**Chacko Jobichen and
Kunchithapadam Swaminathan***Department of Biological Sciences, National
University of Singapore, Singapore 117543,
Singapore

Correspondence e-mail: dbsks@nus.edu.sg

Received 22 May 2014
Accepted 18 June 2014

Molecular replacement with a large number of molecules in the asymmetric unit

The exponential increase in protein structures deposited in the Protein Data Bank (PDB) has resulted in the elucidation of most, if not all, protein folds, thus making molecular replacement (MR) the most frequently used method for structure determination. A survey of the PDB shows that most of the structures determined by molecular replacement contain less than ten molecules in the asymmetric unit and that it is predominantly virus and ribosome structures that contain more than 20 molecules in the asymmetric unit. While the success of the MR method depends on several factors, such as the homology and the size of an input model, it is also a well known fact that this method can become significantly difficult in cases with a large number of molecules in the asymmetric unit, higher crystallographic symmetry and tight packing. In this paper, five representative structures containing 16–18 homomeric molecules in the asymmetric unit and the strategies that have been used to solve these structures are described. The difficulties faced and the lessons learned from these structure-determination efforts will be useful for selected and similar future situations with a large number of molecules in the asymmetric unit.

1. Introduction

Determination of protein structures by X-ray crystallography is carried out using the anomalous scattering, isomorphous replacement or molecular-replacement methods. The exponential increase in the number of structures deposited in the Protein Data Bank (PDB) has resulted in the elucidation of most, if not all, protein folds, making molecular replacement the most frequently used method. While the success of this method depends on the homology and suitability of an input model, it is also a well known fact that molecular replacement can become significantly difficult in cases with a large number of molecules in the asymmetric unit, higher symmetry and tight packing (McCoy, 2007).

Our survey of the PDB (as of June 2014; Supplementary Table S1¹) shows that the majority of the structures determined by molecular replacement contain less than ten molecules in the asymmetric unit (Berman *et al.*, 2000). Entries with 20 or more molecules in the asymmetric unit are predominantly viral assemblies and ribosomes. Based on the number of molecules in the asymmetric unit, we can group the entries as low ($1 \leq x \leq 6$), medium ($7 \leq x \leq 12$) and high ($13 \leq x \leq 18$) (Fig. 1 and Supplementary Table S1). It is interesting to note that the number of entries with an even number of molecules in the asymmetric unit is greater than the number of adjacent entries with odd numbers of molecules. For the present study, we focus on the highest end of the high group, where 60% of the entries have 16–18 molecules in the asymmetric unit.

The entries with a large number of molecules in the asymmetric unit can be divided into three major categories: (i) inherent or physiological assemblies, such as viruses, (ii) homo-multimers and (iii) hetero-multimers, such as ribosomes. Interestingly, the members of the second category can be further subdivided into three cases: (1) homo-oligomers with point-group symmetry (a defined local symmetry), say a hexameric ring, (2) randomly oriented monomers (no local symmetry) and (3) randomly oriented monomers in the

asymmetric unit but with a symmetrical final assembly (monomers oriented in such a way that crystallographic symmetry produces a homo-oligomer). Even though any molecular-replacement attempt can pose a challenge, the case of a large number of homo-multimers in the asymmetric unit with no local symmetry tends to be particularly difficult.

In this paper, we describe five representative structures containing 16–18 homomeric molecules in the asymmetric unit. The strategies that have been used to solve these structures are discussed in detail. The difficulties faced and the lessons learned from these structures will be useful for selected and similar future situations with large numbers of molecules in the asymmetric unit.

2. Results

2.1. Randomly oriented symmetrical monomers

In the case of chloramphenicol acetyltransferase I (CAT_I; PDB entries 3u9b and 3u9f; Biswas *et al.*, 2012), the apo form was

Figure 1

The distribution of Protein Data Bank entries (as of June 3, 2014) with up to 18 molecules in the asymmetric unit which were solved by the molecular-replacement method. Based on the number of molecules in the asymmetric unit (Supplementary Table S1), we can group the entries as (a) low ($1 \leq x \leq 6$), (b) medium ($7 \leq x \leq 12$) and (c) high ($13 \leq x \leq 18$).

crystallized with nine molecules (three trimers) and the complex with chloramphenicol (CAM) was crystallized with 18 molecules (six

Figure 2

(a) The asymmetric unit of the CAT_I-CAM complex crystal (PDB entry 3u9f). The six trimers in the asymmetric unit do not form any local symmetry and are randomly oriented. The members of each trimer are coloured the same and the local threefold axis for one of the trimers is shown. (b) The asymmetric unit of the *L. mexicana* pyruvate kinase (*LmPYK*; PDB entry 3hqp) crystal contains 16 monomers. The molecules are arranged as four randomly oriented tetramers which have local 'D₂' symmetry. (c) The *LmPYK* molecule (PDB entry 1pkl) was split into four domains, residues 1–86 (yellow), 87–187 (cyan), 188–481 (magenta) and 489–498 (green), which were grouped into two ensembles and used as input in molecular replacement.

trimers) in the asymmetric unit. The biological unit is a trimer and the six trimers in the CAT_I–CAM complex do not form any further local symmetry and are randomly oriented in the asymmetric unit (Fig. 2*a*). In fact, the complex was first crystallized in space group *P*₁ with six trimers in the asymmetric unit, and the difficulty in molecular replacement using this crystal form made further crystallization attempts necessary. Further crystallization screening with CAT_I alone yielded crystals in space group *P*₂₁ that diffracted X-rays to 3.2 Å resolution. The structure was determined by molecular replacement with *MOLREP* using this new data and a CAT_I trimer (PDB entry 1noc; Crane *et al.*, 1997) as a search model, which yielded the apo CAT_I structure with three trimers in the asymmetric unit. This trimeric structure was used as the input model and the CAT_I–CAM

complex structure was subsequently, and successfully, determined at 2.9 Å resolution using *MOLREP*.

In the case of *Leishmania mexicana* pyruvate kinase (*LmPYK*; PDB entries 3hqn, 3hqo, 3hqp and 3hqq; Morgan *et al.*, 2010), there are 16 monomers in the asymmetric unit in PDB entry 3hqp (Fig. 2*b*). The molecules are arranged as four randomly oriented tetramers with each tetrameric assembly following local ‘*D*₂’ symmetry. The biological unit is a monomer and an earlier structure of the same protein (PDB entry 1pkl; Rigden *et al.*, 1999) was used as a search model in molecular replacement. *Phaser* (McCoy, 2007) was used to determine the structure at 2.3 Å resolution. However, instead of using an intact model, the input was split into four fragments organized as two ensembles (ensemble 1, residues 87–187; ensemble 2, residues 1–86,

Figure 3
(*a*) The asymmetric unit of the *E. coli* purine nucleoside phosphorylase crystal (PDB entry 3ooh) contains 18 monomers arranged as three linearly displaced hexameric rings. (*b*) One hexameric ring, shown down the threefold axis, was used as the input model in molecular replacement for structure determination.

Figure 4
Stereoview of the asymmetric unit of the cytochrome *c*₅₅₂ crystal (PDB entry 3zow), showing the 18 randomly distributed molecules.

188–481 and 89–498; Fig. 2c) for successful structure determination. This example shows a clever use of a ‘divide-and-conquer’ strategy.

2.2. Regularly oriented symmetrical monomers

In some crystals, the arrangement of a large number of homomultimers in the asymmetric unit can follow certain local symmetry.

For example, the biological unit of *Escherichia coli* purine nucleoside phosphorylase (PNP; Mikleušević *et al.*, 2011) is a hexamer. This protein and its mutant have been crystallized with H_2PO_4^- and SO_4^{2-} ions (PDB entries 3ooe, 3onv, 3ooh and 3opv) and the structures were determined by molecular replacement. The complex of wild-type PNP with phosphate ions (PDB entry 3ooh) was crystallized in space group $P2_1$ and the crystals diffracted X-rays to 2.9 Å resolution. The

Figure 5 Rv1498A dodecamer (a) and monomer (b). (c) Eight dodecins, represented as black circles, in the unit cell of the *H. salinarum* dodecin crystal (PDB entry 2cc7). The fractional coordinates of the centres of the dodecins are $(0, 0, 0)$, $(0, \frac{1}{2}, \frac{1}{2})$, $(\frac{1}{2}, 0, \frac{1}{2})$, $(\frac{1}{2}, \frac{1}{2}, 0)$, $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $(\frac{3}{4}, \frac{3}{4}, \frac{3}{4})$, $(\frac{3}{4}, \frac{1}{4}, \frac{3}{4})$ and $(\frac{1}{4}, \frac{3}{4}, \frac{3}{4})$. (d) 16 Rv1498A dodecin molecules are packed in the unit cell of the *M. tuberculosis* Rv1498A crystal (PDB entry 3oqt). The fractional coordinates of the centres of the dodecins are $(0, 0, 0)$, $(\frac{1}{2}, \frac{1}{2}, 0)$, $(0, \frac{1}{2}, 0)$, $(0, 0, \frac{1}{2})$, $(\frac{1}{2}, 0, \frac{1}{2})$, $(\frac{1}{2}, \frac{1}{2}, 0)$, $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $(\frac{3}{4}, \frac{3}{4}, \frac{3}{4})$, $(\frac{3}{4}, \frac{1}{4}, \frac{3}{4})$, $(\frac{3}{4}, \frac{3}{4}, \frac{1}{4})$, $(\frac{1}{4}, \frac{3}{4}, \frac{3}{4})$, $(\frac{1}{4}, \frac{1}{4}, \frac{3}{4})$, $(\frac{1}{4}, \frac{3}{4}, \frac{1}{4})$ and $(\frac{3}{4}, \frac{1}{4}, \frac{1}{4})$.

18 subunits in the asymmetric unit (Fig. 3) are arranged as three linearly translated homohexameric rings along the c axis. Phasing was

carried out by molecular replacement with *Phaser* (McCoy, 2007) in *PHENIX* (Adams *et al.*, 2010), using the ternary complex (hexamer) of purine nucleoside phosphorylase (PDB entry 1k9s; Koellner *et al.*, 2002) as a model. This example suggests that we need to prepare a hexamer, which is a subset of a larger symmetrical arrangement, as an input model to obtain a successful solution.

2.3. Randomly oriented monomers

In several crystals, there may be no local symmetrical relationship among the molecules in the asymmetric unit. Also, each molecule in the asymmetric unit may be a biological or functional unit and the asymmetric unit molecules may not form any further assembly. A typical example is the cytochrome c_{552} molecule from *Nitrosomonas europaea* (NeC552n; PDB entry 3zow; Can *et al.*, 2013). The biological unit is a monomer and the 18 molecules in the asymmetric unit are randomly oriented (Fig. 4). The NeC552n structure was determined at 2.35 Å resolution by using NeN64D (PDB entry 3zox; Can *et al.*, 2013) as a search model in *Phaser* (McCoy, 2007).

2.4. Randomly oriented monomers in the asymmetric unit but a symmetry-generated biological unit

Unlike in the above section, in some crystals there may be no local symmetrical relationship among the molecules in the asymmetric unit but the biological or functional unit is a larger and symmetrical assembly. Recently, we have solved the crystal structure of Rv1498A, a dodecimer from *Mycobacterium tuberculosis* (PDB entry 3oqt, space group $P2_13$; Liu *et al.*, 2011). The biological molecule (Fig. 5a) always exists as a dodecamer (an assembly of 12 monomers, each of 70 residues; Fig. 5b), hereafter called a ball. The crystals of Rv1498A and another dodecimer (PDB entry 2cc7; space group $F4_132$; only one molecule in the asymmetric unit; eight dodecamers in the unit cell; Fig. 5c) have the same unit-cell parameters. This unit-cell mimicry convinced us to assume that Rv1498A could contain only eight dodecamers in the unit cell, even though the space group is different ($P2_13$). A Matthews coefficient calculation (Matthews, 1968) suggested the plausible presence of 8–16 monomers in the asymmetric unit. Our molecular-replacement attempts using *MOLREP* (Vagin & Teplyakov, 2010) with the 2cc7 monomer or an Rv1498A monomer (prepared from 2cc7; Grininger *et al.*, 2006), based on sequence alignment using *ClustalW* (Larkin *et al.*, 2007), did not produce any successful result.

The strategy was changed to locate all dodecimer balls in the entire unit cell by using a full ball as the input search model in the triclinic space group $P1$. *MOLREP* now identified six balls in the unit cell at locations 1, 5, 6, 7, 9 and 16 (Fig. 5d). As these locations are also common to the 2cc7 unit cell, we believed that Rv1498A should pack as 2cc7. All eight balls were manually placed at locations 1, 5, 6, 7, 9, 12, 13 and 16. However, the crystallographic R value was stuck around 0.55, suggesting either an error in our positioning of the dodecimer balls or a need for more diffracting matter in the unit cell.

The 12 dodecimer molecules are assembled as a tetrahedron, with a trimer on each face (Fig. 6a). Furthermore, a tetrahedron can be oriented in two possible ways without jeopardizing symmetry (Fig. 6b), with the face normals at quadrants I and III projected above the xy plane when viewed down the z axis (left panel) or below the xy plane (right panel). Each of the two possible orientations of all eight dodecimer balls would still validate the symmetry requirements of the space group. This further complicated the manual placement of the balls and refinement. All combinations were attempted and the R factor did not improve.

Figure 6

(a) Three monomer molecules on each face of the tetrahedral arrangement of the Rv1498A dodecimer. (b) Two equivalent orientations of a tetrahedron (viewed down the z axis) satisfying the $P2_13$ symmetry. The central solid line, one of the tetrahedron edges, is close to the viewer (above the xy plane) and the dashed line edge is below the xy plane. In the left panel, the face normals of quadrants I and II are above the xy plane and those of quadrants III and IV are below the xy plane. This arrangement is reversed in the right panel, where the tetrahedron is rotated by 90° around the z axis. (c) Views of the tetrahedrons down the x axis. The upper and lower edges of the tetrahedrons are parallel to the xy plane. Note that in both (b) and (c), the axes and O just represent the local unit-cell edge directions and the centre of the tetrahedron, respectively, and are not the actual crystallographic axes or the origin of the unit cell.

A close observation of electron density throughout the unit cell clearly indicated that we could place 16 dodecin molecules in the unit cell. We placed 16 dodecins with their centres at fractional coordinates $(0, 0, 0)$, $(\frac{1}{2}, 0, 0)$, $(0, \frac{1}{2}, 0)$, $(0, 0, \frac{1}{2})$, $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, $(\frac{1}{2}, \frac{1}{2}, 0)$, $(\frac{1}{2}, 0, \frac{1}{2})$, $(0, \frac{1}{2}, \frac{1}{2})$, $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $(\frac{1}{4}, \frac{1}{4}, \frac{3}{4})$, $(\frac{3}{4}, \frac{1}{4}, \frac{1}{4})$, $(\frac{3}{4}, \frac{1}{4}, \frac{3}{4})$, $(\frac{1}{4}, \frac{3}{4}, \frac{1}{4})$, $(\frac{1}{4}, \frac{3}{4}, \frac{3}{4})$ and $(\frac{3}{4}, \frac{3}{4}, \frac{3}{4})$ (Fig. 5*d*). Even though the *R* factor was better (40%), the structure could not be refined further. A ‘ring-like assembly’ of four symmetrically independent monomers (Fig. 7*a*) was now derived and the last

(and successful) operation was a run of *Phaser* (McCoy *et al.*, 2007) with a ring as the input and a search for four such rings (or 16 monomers) in the asymmetric unit that would account for 16 balls in the unit cell. The *R* factor was 0.27 and *R*_{free} was 0.37. The missing N- and C-terminal residues (1, 2, 69 and 70), ions and water molecules were added and the structure was refined with *REFMAC* (Murshudov *et al.*, 2011) to final *R* and *R*_{free} values of 0.25 and 0.28, respectively. Thus, each of the 16 chains in the asymmetric unit generates 12 chains in the unit cell through the 2₁3 symmetry and all 16 chains in the asymmetric unit generate 192 chains, assembled as 16 dodecins, in the unit cell.

After the completion of refinement, we analyzed the tetrahedral orientations of the 16 dodecins in the unit cell. Monomers *A*, *B*, *C* and *D* of the asymmetric unit form the dodecins at positions 2, 3, 4 and 8, monomers *E*, *F*, *G* and *H* form those at 10, 11, 14 and 15, monomers *I*, *J*, *K* and *L* form those at 9, 12, 13 and 16 and monomers *M*, *N*, *O* and *P* form the dodecins at positions 1, 5, 6 and 7. When viewed down the *z* axis, the *ABCD* and *IJKL* rings are well aligned with the crystallographic axes (Fig. 7*b*). The pseudo plane of the *ABCD* ring is parallel to the *xy* plane and that of the *IJKL* ring is parallel to the *xz* plane. However, the *EFGH* and *MNOP* rings are misaligned with respect to the crystallographic *x* axis by -30° and 60° , respectively (Fig. 7*c*), which means the corresponding tetrahedrons are also misaligned from ideality (Fig. 6*b*). This misalignment must have led to the resistance to refinement and halted the *R* factor at 0.4 when we manually, and perfectly, placed the 16 dodecamers in the unit cell.

3. Conclusion

Several proteins function only as an assembly of homo- or heteropolypeptide chains. When the number of molecules in the asymmetric unit becomes significantly large, structure determination becomes considerably difficult (Scapin, 2013). It becomes obvious that in such

Figure 7

(*a*) A ring-like structure formed by four monomers. There are four such rings in the asymmetric unit. (*b*) When viewed down the *z* axis, the planes of the pseudo rings formed by monomers *A*, *B*, *C*, *D* and *I*, *J*, *K*, *L* are parallel to the crystallographic *xy* and *xz* planes, whereas (*c*) the *EFGH* and *MNOP* rings are misaligned by -30° and 60° with respect to the crystallographic *xz* plane. The misalignment is measured as the angle between the line *Ox* and the dashed line that represents the planarity of the rings.

cases molecular replacement needs various ‘tricks’ to obtain a correct solution. In §2.1, a proper solution could be obtained when a large model was first divided into fragmented domains and ensembles of such domains were used as input. In §2.2, subsets of a final larger and symmetrical arrangement, such as a dimer, trimer or tetramer, were prepared as inputs and the hexamer model gave a successful solution. In the last case, §2.4, a complete biological assembly, instead of the monomers of the asymmetric unit, was used as a model to search the entire unit cell in space group *P1*.

The usual and most frequent approach in the molecular-replacement method is to take a monomer of a homologous structure and attempt structure determination. In certain cases, truncations of flexible, N- or C-terminal regions were attempted. In the current study, we focus on some challenging examples of molecular replacement and highlight the attempts undertaken to determine the structure. Firstly, we described two strategies: arrangement of quaternary structures into an ensemble and dividing a single molecule into multiple ensembles. In the last example, we witnessed the need for two different approaches in tandem: the use of a full molecular assembly to locate such assemblies in the unit cell and a tear-down strategy to identify the molecules in the asymmetric unit. This approach is quite unusual but was necessary owing to the misorientation of certain molecular assemblies at selected locations in the unit cell.

Two important lessons can be learned. The first is that preliminary planning regarding the selection of an input model is very important. While the ‘divide-and-conquer’ rule wins on several occasions, the ‘unity is strength’ formula also should not be ignored. The approach of going to the unit-cell level with a full dodecin search model, instead of an individual monomer as bait at the asymmetric unit level, was the key to success in §2.4. Another lesson learned is with regard to the misorientation of only some molecules in the unit cell. This warns and prepares us for future surprises in which the misorientation of a large assembly, which cannot be identified at the ‘microscopic’ level, will become obvious at the ‘intermediate’ or ‘macroscopic’ level.

The authors hope that powerful algorithms to handle problems of even greater magnitudes (with even a larger number of molecules in the asymmetric unit) with a small input search monomer will evolve in the future.

The authors thank Michael Rossmann (Purdue University), Jack Johnson (Scripps), Tim Baker (UCSD), Liang Tong (Columbia University), Liang Tang (Kansas University), Jim Pflugrath (Rigaku) and Peter Zwart (Berkeley Center for Structural Biology) for their advice and help. This project was partially supported by grants from the National University of Singapore (NUS)/Singapore Ministry of Education Academic Research Fund and Singapore Biomedical Research Council (BMRC).

References

- Adams, P. D. *et al.* (2010). *Acta Cryst.* **D66**, 213–221.
- 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.
- Biswas, T., Houghton, J. L., Garneau-Tsodikova, S. & Tsodikov, O. V. (2012). *Protein Sci.* **21**, 520–530.
- Can, M., Krucinska, J., Zoppellaro, G., Andersen, N. H., Wedekind, J. E., Hersleth, H. P., Andersson, K. K. & Bren, K. L. (2013). *Chembiochem*, **14**, 1828–1838.
- Crane, B. R., Arvai, A. S., Gachhui, R., Wu, C., Ghosh, D. K., Getzoff, E. D., Stuehr, D. J. & Tainer, J. A. (1997). *Science*, **278**, 425–431.
- Grininger, M., Zeth, K. & Oesterhelt, D. (2006). *J. Mol. Biol.* **357**, 842–857.
- Koellner, G., Bzowska, A., Wielgus-Kutrowska, B., Lubić, M., Steiner, T., Saenger, W. & Stepniński, J. (2002). *J. Mol. Biol.* **315**, 351–371.
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007). *Bioinformatics*, **23**, 2947–2948.
- Liu, F., Xiong, J., Kumar, S., Yang, C., Ge, S., Li, S., Xia, N. & Swaminathan, K. (2011). *J. Struct. Biol.* **175**, 31–38.
- Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
- McCoy, A. J. (2007). *Acta Cryst.* **D63**, 32–41.
- McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
- Miklušević, G., Stefanić, Z., Narczyk, M., Wielgus-Kutrowska, B., Bzowska, A. & Lubić, M. (2011). *Biochimie*, **93**, 1610–1622.
- Morgan, H. P., McNae, I. W., Nowicki, M. W., Hannaert, V., Michels, P. A., Fothergill-Gilmore, L. A. & Walkinshaw, M. D. (2010). *J. Biol. Chem.* **285**, 12892–12898.
- Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* **D67**, 355–367.
- Rigden, D. J., Phillips, S. E., Michels, P. A. & Fothergill-Gilmore, L. A. (1999). *J. Mol. Biol.* **291**, 615–635.
- Scapin, G. (2013). *Acta Cryst.* **D69**, 2266–2275.
- Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* **D66**, 22–25.