OASIS and molecular-replacement model completion

Yao He, De-Qiang Yao, Yuan-Xin Gu, Zheng-Jiong Lin, Chao-De Zheng and Hai-Fu Fan

Copyright © International Union of Crystallography
Author(s) of this paper may load this reprint on their own web site provided that this cover page is retained. Republication of this article or its storage in electronic databases or the like is not permitted without prior permission in writing from the IUCr.
A method of dual-space molecular-replacement model completion has been proposed which involves the programs ARP/wARP, REFMAC, OASIS and DM. OASIS is used in reciprocal space for phase refinement based on models built by ARP/wARP. For this purpose, the direct-method probability formula of breaking SAD/SIR phase ambiguities has been redefined. During the phase refinement, $\varphi_h$ in the expression $\varphi_h = \varphi_h^0 \pm |\Delta \varphi_h|$ is redefined as a reference phase calculated from a randomly selected 5% of the atoms in the current structure model, while $|\Delta \varphi_h|$ is defined as the absolute difference between the phase of the current model and $\varphi_h^0$. The probability formula

$$P_h(\Delta \varphi_h) = \frac{1}{2} + \frac{1}{2} \tanh \left[ \frac{1}{2} \sin \left| \Delta \varphi_h \right| \sin \left( \varphi_h^0 + \Delta \varphi_h^{\text{best}} + \Delta \varphi_h - \varphi_h^{\text{best}} \right) \sin \delta_h \right]$$

is then used to derive the sign of $\Delta \varphi_h$. In this way the ‘0–2\pi’ phase problem is reduced to a ‘plus or minus’ sign problem. The redefinition implies that during the refinement phases close to the true values will probably be kept unchanged, while those distant from the true values will probably undergo a large shift. This is the desired property of phase refinement. The procedure has been tested using protein diffraction data without SAD/SIR signals. The results show that dual-space MR-model completion making use of OASIS is much more efficient than that without.

1. Introduction

Dual-space fragment extension with SAD/SIR information (Wang, Chen, Gu, Zheng & Fan, 2004; Yao et al., 2006) has proved to be very efficient for model completion in solving protein crystal structures using SAD/SIR data. Can this technique also be applied in the absence of SAD/SIR information? This has still to be explored. In a different context, the hybrid direct method proposed by Fan & Gu (1985) has been applied to improve SAD/SIR phasing. Is this method also useful for improving the molecular-replacement procedure? This is worth trying, as about 70% of protein structures submitted to the Protein Data Bank (Berman et al., 2000) in recent years have been solved using the molecular-replacement method. The present work is aimed at answering these questions.

2. Method

2.1. Dual-space model completion

Dual-space model completion is an iterative procedure consisting of three parts: (i) real-space model building and refinement, (ii) reciprocal-space phase refinement based on a partial model and (iii) phase improvement by density modification. The first part is implemented by the programs ARP/
wARP (Perrakis et al., 1999) and REFMAC (Murshudov et al., 1997). These two programs are integrated in the ARP/wARP program suite. The second part is implemented by the program OASIS06 (Zhang et al., 2007). The third part is implemented by the program DM (Collaborative Computational Project, Number 4, 1994; Cowtan, 1994). Fig. 1 shows a flowchart of the procedure. Further descriptions of the second part are given below. The program OASIS06 is freely available for academic purposes at the website http://cryst.iphy.ac.cn.

2.2. Redefining the \( P_i \) formula

The \( P_i \) formula

\[
P_i(\Delta \varphi_h) = \frac{1}{2} + \tanh \left( \sin |\Delta \varphi_h| \left[ \sum_k m_k m_{h-k} \kappa_{h,h} \times \sin(\Phi'_j + \Delta \varphi_{h,\text{best}} + \Delta \varphi_{h,j} + \chi \sin \delta_h) \right] \right)
\]

(1)

was originally proposed by Fan & Gu (1985) to break the phase ambiguity intrinsic in SAD/SIR data, i.e. to estimate the probability of \( \Delta \varphi_h \) being positive. In the SAD case, the symbols in (1) are defined as follows.

\[
\Delta \varphi_h = \varphi_h - \varphi_h^c \text{ or } \varphi_h = \varphi_h^c \pm |\Delta \varphi_h|,
\]

(2)

where \( \varphi_h \) is the phase of a reflection associated with the reciprocal vector \( h \). \( \varphi_h^c \) is the phase contributed from imaginary-part scattering of anomalous scatterers, i.e. the phase of

\[
F'_h = i \sum_{j=1}^N \Delta F'_j \exp(i2\pi h \cdot r_j),
\]

(3)

where \( \Delta F'_j \) is the imaginary-part correction to the atomic scattering factor of the \( j \)th atom. Both \( \varphi_h^c \) and \( |\Delta \varphi_h| \) are known quantities provided that the anomalous-scattering substructure is known.

\[
m_k = \exp(-\sigma_k^2/2) \left| \left[ 2(P_+ - 2) + 1 \right] (1 - \cos 2\Delta \varphi_h) + \cos 2\Delta \varphi_h \right|^{1/2},
\]

(4)

with

\[
\sigma_k^2 = \left( \frac{n \sigma_{\Delta \varphi_h}}{2F_k^0} \right)^2,
\]

(5)

where \( n \) is a scaling factor and \( \sigma_{\Delta \varphi_h} \) is the standard deviation of Bijvoet difference \( \Delta F_h \) (see Wang, Chen, Gu, Zheng, Jiang et al., 2004).

The phase of a reflection associated with the reciprocal vector \( h \) is the phase contributed from real-part scattering of anomalous scatterers, i.e.

\[
\varphi_h = \varphi_h^r + \varphi_h^c
\]

(6)

where \( E_h \) is the normalized structure-factor magnitude associated with the reciprocal vector \( h \) and \( Z_j \) is the atomic number of the \( j \)th atom in the unit cell.

\[
\Phi'_j = -\varphi_h^r + \varphi_h^c + \varphi_h^c
\]

(7)

is the three-phase structure invariant contributed from the imaginary-part scattering of anomalous scatterers.

\[
\tan(\Delta \varphi_{h,\text{best}}) = 2(P_+ - \frac{1}{2}) \sin |\Delta \varphi_h| / \cos \Delta \varphi_h,
\]

(8)

\[
\varphi_{h,\text{best}} = \varphi_h^c + \Delta \varphi_{h,\text{best}},
\]

(9)

\[
\chi = 2E_h E_{h,\text{known}} \left( \sum_{j=1}^{\text{unknown}} Z_j^2 / \sum_{j=1}^{\text{total}} Z_j^2 \right)
\]

(10)

where ‘known’ refers to the known substructure, ‘unknown’ refers to the unknown part of the unit cell and ‘total’ refers to the whole unit cell.

\[
\delta_h = \varphi_h - \varphi_h^c
\]

(11)

where \( \varphi_h^c \) is the phase contributed from real-part scattering of the known substructure.

The main points implied by the \( P_i \) formula are as follows.

(i) Bimodal SAD/SIR phase distributions lead the ‘0–2\( \pi \)’ phase problem to a much simpler ‘plus or minus’ sign problem. This is of great benefit to the reliability of phase derivation.

(ii) The concept of best phases and figures of merit in protein crystallography (Blow & Crick, 1959) is introduced into direct-methods phasing.

(iii) Partial structure information is incorporated into the phasing process.

In the absence of SAD/SIR information, (1) and (2) have to be redefined in order to retain the above advantages. The phase \( \varphi_h^c \) in (2) is now defined as a reference phase, which is calculated from a randomly selected 5% of the total atoms in the current structure model. \( |\Delta \varphi_h| \) is then redefined as the absolute difference between the phase of the current model and \( \varphi_h^c \).

In order to use (1) to derive the sign of \( \Delta \varphi_h \), we can either simulate the SAD case or the SIR case. The SAD case is chosen in the present work. This ensures that the algorithm is equally efficient for noncentric reflections and centric reflections and there is no need to consider whether or not the ‘heavy-atom substructure’ is centrosymmetric. The randomly selected 5% of the atoms in the current structure model are now taken as ‘anomalous scatterers’. An artificial imaginary-

---

**Figure 1**

Flowchart of dual-space MR model completion.
part correction $\Delta f'' = 2.0$ to the atomic scattering factor is assumed for these atoms. The reference phase $\psi_h^*$ can then be calculated using (3). Note that the value of $\Delta f''$ does not affect the resultant $\psi_h^*$. For MR model completion in the absence of SAD/SIR signals there is no lack-of-closure error. Hence, in principle, (4) should reduce to

$$m_b = \left[ \left( 2(P_s - \frac{1}{2}) \right)^2 + \frac{1}{2} \right] \left( 1 - \cos 2\Delta \psi_h^* + \cos 2\Delta \psi_h \right)^{1/2}. \quad (12)$$

However, an alternative treatment is to keep (4) unchanged by replacing $\sigma_{\Delta \psi_h}$ with $\sigma_b$ in (5) so that

$$\sigma_b^2 = \frac{(m \sigma_b)^2}{2|F_h|^2}. \quad (13)$$

Originally, $\sigma_b^2$ was the variance of two Gaussian functions used to approximate the bimodal SAD/SIR distribution (see Fan et al., 1984); the use of (13) thus implies the introduction of an artificial bimodal SAD distribution with the 'sharpness' depending on the errors of experimentally measured structure-factor magnitudes. (4) and (13) are actually used in the program OASIS 06 to calculate figures of merit. No further modification is necessary for formulae (6)–(11), keeping in mind that the term 'known' in (10) now means the current structure model and $\psi_h^*$ in (11) is now the phase from the current structure model. Finally, (4), (8) and (9) are used to update figures of merit and 'best phases' during iteration.

**2.3. Implications of the redefinition**

Phase derivation in OASIS is associated with both 'best phases' and 'figures of merit'. For reasons of simplicity, the discussion in this section will focus only on 'phases' and will disregard 'figures of merit'. The redefinition of (1) implies the following phasing properties. At the beginning of model completion, according to (2), the initial phases will be those calculated from the starting model. By applying (1) and (2), a new phase will be given to each reflection. There are only two choices for each new phase: (i) if the sign of $\Delta \psi_h$ is positive then $\psi_h$ will be kept unchanged and (ii) if the sign of $\Delta \psi_h$ is negative then $\psi_h$ will be changed by $-2|\Delta \psi_h|$. Since $|\Delta \psi_h| \leq 180^\circ$, ranges from 0° to 180° with a mean of ~90°, $2|\Delta \psi_h|$ will mostly be a large angle. In principle, (1) tends to move $\psi_h$ towards its true value. Hence, if the initial $\phi_h$ is close to the true value, (1) will probably give a plus sign to keep $\phi_h$ unchanged. On the other hand, if the initial $\phi_h$ is distant from the true value, (1) will probably give a minus sign and let $\phi_h$ have a large shift. This property is what is expected for a good phase-refinement procedure.

**2.4. Comparison of the redefined $P_s$ formula with the tangent formula in Karle recycling**

The direct method used in the present study is in fact a phase refinement incorporated with partial structure information. A similar type of method is Karle recycling (Karle, 1968), which has long been used successfully for model completion in small-molecule crystallography. However, Karle recycling is not as efficient for proteins as for small molecules. In comparison with Karle recycling, the present method has the following advantages.

(i) The '0–2$\pi$' phase problem reduces to a much simpler 'plus or minus' sign problem. Obviously, the latter will provide much more reliable results.

(ii) Experimental errors in measured structure-factor magnitudes have been taken into account.

**Table 1**

Summary of the test data.

<table>
<thead>
<tr>
<th></th>
<th>Acids phospholipase A&lt;sub&gt;2&lt;/sub&gt;</th>
<th>E7_C–Im7_C complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB code</td>
<td>1m8t</td>
<td>1ujz</td>
</tr>
<tr>
<td>Space group</td>
<td>P6&lt;sub&gt;1&lt;/sub&gt;</td>
<td>I222</td>
</tr>
<tr>
<td>Unit-cell parameters (Å·&lt;sup&gt;2&lt;/sup&gt;)</td>
<td>$a = b = 83.27$, $c = 123.20$, $\gamma = 120$</td>
<td>$a = 62.88$, $b = 74.55$, $c = 120.44$</td>
</tr>
<tr>
<td>Resolution limit (Å)</td>
<td>1.9</td>
<td>2.1</td>
</tr>
<tr>
<td>X-rays</td>
<td>Cu Kα</td>
<td>Synchrotron</td>
</tr>
<tr>
<td>No. of reflections</td>
<td>8697</td>
<td>17107</td>
</tr>
<tr>
<td>No. of residues in the ASU</td>
<td>124</td>
<td>215 (molecule A, 87; molecule B, 128)</td>
</tr>
<tr>
<td>No. of residues in the starting model</td>
<td>Model 1, 63; model 2, 51; model 3, 49</td>
<td>46; 13 with side chains, all in molecule A</td>
</tr>
</tbody>
</table>

**Table 2**

Summary of model completion of acidic phospholipase A<sub>2</sub>.

<table>
<thead>
<tr>
<th></th>
<th>63-residue model</th>
<th>51-residue model</th>
<th>40-residue model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>Protocol A</td>
<td>D-A</td>
<td>O-D-A</td>
</tr>
<tr>
<td>Cycle 0</td>
<td>119 (119)</td>
<td>37 (37)</td>
<td>32 (0)</td>
</tr>
<tr>
<td>Cycle 1</td>
<td>33 (0)</td>
<td>41 (0)</td>
<td>14 (0)</td>
</tr>
<tr>
<td>Cycle 2</td>
<td>38 (27)</td>
<td>121 (121)</td>
<td>5 (0)</td>
</tr>
<tr>
<td>Cycle 3</td>
<td>72 (59)</td>
<td>0 (0)</td>
<td>57 (0)</td>
</tr>
<tr>
<td>Cycle 4</td>
<td>121 (121)</td>
<td>66 (0)</td>
<td></td>
</tr>
<tr>
<td>Cycle 5</td>
<td>121 (121)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2

Model completion of acidic phospholipase A<sub>2</sub> with the 63-residue starting model. Left, starting model; middle, model built after a default run of ARPhwARP; right, refined (final) model.
(iii) In Karle recycling, partial structure information is implicitly included in the tangent formula by associating partial structure phases with normalized structure-factor magnitudes. This may cause strong model bias in three-phase structure invariants. On the other hand, in the $P_+$ formula (1), partial structure information is explicitly included in the term $\chi \sin \sigma_6$, which is separated from the term involving three-phase structure invariants. Model bias in phase relationships is thus weakened.

3. Test data

Two known proteins were selected for testing. They are summarized in Table 1. Structure-factor amplitudes without SAD/SIR signals were taken from the Protein Data Bank (Berman et al., 2000).

4. Results and discussion

4.1. Model completion of the protein acidic phospholipase A$_2$

Acidic phospholipase A$_2$ from the venom of Agkistrodon halys Pallas is a 124-residue protein; its structure was originally solved by the molecular-replacement method (Wang et al., 1996; Xu et al., 2003). Three starting models were used in this test. They were obtained from a partial model truncated to 63, 51 and 40 residues, which correspond to 51, 41 and 32% of the whole structure, respectively. The 40-residue model contains two highly conserved $\alpha$-helices that form the core of the molecule. The results of the test are summarized in Table 2. The models automatically built by ARP/wARP and plotted using PyMOL (DeLano, 2002) at various stages of model completion based on the three starting models are shown in Figs. 2, 3 and 4, respectively. For the 63-residue starting model, a default run of ARP/wARP automatically led to a model of 119 residues (96% of the whole structure) all docked into the sequence. In this case neither DM nor OASIS was necessary. For the 51-residue starting model, a default run of ARP/wARP did not improve the model. However, either iteration of DM-ARP/wARP (bypassing OASIS in the flowchart in Fig. 1) or iteration of OASIS-DM-ARP/wARP led to a model of 121 residues (98% of the whole structure) all docked into the sequence. DM-ARP/wARP needed...
Table 4
Summary of model completion of E7_C–Im7_C.

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Cycle 0</th>
<th>Cycle 1</th>
<th>Cycle 2</th>
<th>Cycle 3</th>
<th>Cycle 4</th>
<th>Cycle 5</th>
<th>Cycle 6</th>
<th>Cycle 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>46 (13)</td>
<td>14 (0)</td>
<td>66 (0)</td>
<td>94 (50)</td>
<td>81 (61)</td>
<td>151 (111)</td>
<td>191 (161)</td>
<td>201 (201)</td>
</tr>
<tr>
<td>D-A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O-D-A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5
Correlation between input phase errors and output signs of $\Delta \phi_h$ from OASIS during model completion of E7_C–Im7_C.

<table>
<thead>
<tr>
<th>Range of phase errors</th>
<th>Cycle 1</th>
<th>Cycle 3</th>
<th>Cycle 5</th>
<th>Cycle 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–30</td>
<td>3337</td>
<td>3761</td>
<td>60.0</td>
<td>4197</td>
</tr>
<tr>
<td>30–60</td>
<td>2856</td>
<td>2874</td>
<td>49.1</td>
<td>3049</td>
</tr>
<tr>
<td>60–90</td>
<td>2635</td>
<td>2359</td>
<td>38.0</td>
<td>2279</td>
</tr>
<tr>
<td>90–120</td>
<td>2291</td>
<td>2189</td>
<td>27.3</td>
<td>1975</td>
</tr>
<tr>
<td>120–150</td>
<td>2175</td>
<td>2097</td>
<td>21.1</td>
<td>1858</td>
</tr>
<tr>
<td>150–180</td>
<td>2064</td>
<td>2071</td>
<td>17.3</td>
<td>1991</td>
</tr>
</tbody>
</table>

Figure 5
Variation of overall average phase errors during model completion of acidic phospholipase A$_2$ with the 40-residue starting model.

4.2. Model completion of the protein–protein complex E7_C–Im7_C

This test was designed to simulate a difficult situation. A protein complex consisting of two different molecules A and B was selected as the test sample. The MR model was derived from a search model that only bears structural similarity to molecule A. This MR model was then extended by dual-space iteration to solve the whole structure. The complex of colicin E7 DNAse and the IM7 immunity protein (E7_C–Im7_C; PDB code 1ujj) is a 215-residue protein containing two molecules: A of 87 residues and B of 128 residues. Molecule A of a similar complex (PDB code 1bxi) bears 60% sequence identity to molecule A of E7_C–Im7_C. The root-mean-square deviation between the two A molecules is 1.38 Å. Molecule A of 1bxi was pruned using the program CHAINSAW (Schwarzenbacher et al., 2004) to provide a search model for Phaser (Read, 2001) to implement molecular replacement. Based on the MR model obtained from Phaser, ARP/wARP automatically built a model of 46 residues, none of which were within the region of molecule B of E7_C–Im7_C and of which only 13 have side chains. This 46-residue model, amounting to ~20% of the structure of E7_C–Im7_C, was used as the starting point for OASIS–DM–ARP/wARP iteration in the test. The results are summarized in Table 4. Models at different stages automatically built by ARP/wARP and plotted using PyMOL are shown in Fig. 6. It was found in this case that iteration of DM–ARP/wARP was unable to obtain a solution and was corrupted completely in cycle 2, listed in the first column. It is seen in each cycle that smaller input phase errors lead to larger percentages of reflections receiving a plus sign for $\Delta \phi_h$. This means that input phases with smaller errors have a higher possibility of being kept unchanged by OASIS, while those with larger errors have a higher possibility of receiving a large phase shift. Comparing the statistics of cycle 1 to those of cycle 5, it is found that the number of reflections with the smallest input phase errors increases and the percentage of these reflections that receive a plus sign for $\Delta \phi_h$ also increases after phase refinement. The opposite is found for the reflections with the largest input phase errors. This signifies that the model completion is progressing satisfactorily.
leading to a set of dummy atoms that contained no secondary-structure information. However, seven cycles of OASIS–DM–ARP/wARP iteration managed to produce a model of 201 residues (≈94% of the whole structure), all docked into the sequence. The variations of overall average phase errors during model completion are plotted in Fig. 7. The black curve shows DM–ARP/wARP iteration, while the cyan curve shows OASIS–DM–ARP/wARP iteration. The phasing property of OASIS described in §2.3 is validated in Table 5. Here, it is again seen in each cycle that smaller input phase errors lead to a larger percentage of reflections receiving a plus sign for |Δφ|= . On the other hand, from cycle 1 to cycle 7 the number of reflections having smallest input phase errors increases and their percentage of receiving a plus sign for |Δφ|= also increases. The opposite is found for the reflections with the largest input phase errors. This implies a successful phase refinement.

5. Concluding remarks

The diffraction data used in the present study were at about 2 Å resolution. For data at lower resolution, the program RESOLVE (build only; Terwilliger, 2003a,b) can be used to replace ARP/wARP and manual adjustment to model building may sometimes be inevitable during the iteration. Dual-space procedures involving direct methods have proved to be very efficient for ab initio solution of protein structures using diffraction data at atomic resolution (Weeks et al., 1993; Sheldrick & Gould, 1995; Foadi et al., 2000), for locating heavy-atom sites in proteins (Weeks et al., 1993; Sheldrick & Gould, 1995), for SAD/SIR phasing (Gu, Chang et al., 2002; Gu, Jiang et al., 2002) and for model completion in the presence of SAD/SIR signals (Yao et al., 2006). It is shown in the present study that they are also very efficient in MR model completion in the absence of SAD/SIR information. Direct methods thus have an important influence on nearly all aspects of single-crystal diffraction analysis of proteins.

H-FF would like to thank Professor D. C. Liang for helpful discussions. This work was supported by the Innovation Project of the Chinese Academy of Sciences and by the 973 Project (grant No. 2002CB713801) of the Ministry of Science and Technology of China.

References


