Optimizing the error term in direct-method SAD phasing

Optimizing the error term in direct-method SAD phasing

The probability formula of the direct-method SAD (single-wavelength anomalous diffraction) phasing proposed by Fan & Gu (1985, Acta Cryst. A41, 280–284) contains an error term which is related to the lack-of-closure error. This error term is used as a weighting function in the phase derivation and in the subsequent calculation of electron-density maps. Previously, there has been a constant in the error term that has had to be determined empirically for each particular case. It has been found that improper choice of the constant often leads to failure of the direct-method SAD phasing. The problem is resolved by introducing a modified error term and a method of automatically tuning the associated scaling factor.

1. Introduction

In the direct-method SAD (single-wavelength anomalous diffraction) phasing proposed by Fan and coworkers (Fan, Han & Qian, 1984; Fan, Han, Qian et al., 1984; Fan & Gu, 1985), the phase of reflections is expressed as

$$\varphi = \varphi'' \pm |\Delta \varphi|,$$

where $\varphi''$ is the phase of imaginary-part scattering from the heavy-atom (anomalous-scatterer) substructure, i.e.

$$F_\varphi'' = i \sum_{j=1}^{N} f_j'' \exp(i2\pi h \cdot r).$$

Given the known heavy-atom substructure, the absolute value of $\Delta \varphi$ can be calculated as (see Blundell & Johnson, 1976)

$$|\Delta \varphi| \sim \cos^{-1} \left( |F^{+} h| / |F^{0} h| / 2|F'' h| \right) \equiv \cos^{-1} \left( \frac{\Delta F}{2|F'' h|} \right).$$

The sign of $\Delta \varphi$ is then estimated by the probability of $\Delta \varphi$ being positive

$$P_+ (\Delta \varphi_h) = \frac{1}{2} + \frac{1}{2} \tanh \left\{ \sin |\Delta \varphi_h| \left[ \sum_{h'} m_{h'h} m_{h} \right] \times \kappa \sin \left( \Phi_h' + \Delta \varphi_{h'} + \Delta \varphi_{h' \text{best}} + \chi \sin \delta_h \right) \right\},$$

where

$$\kappa = 2 \omega_3 \sigma_2^{-3/2} |E_{h'} E_{h} E_{h'}|,$$

$$\Phi_h' = \varphi''_h + \varphi''_h + \varphi''_{h' \text{best}},$$

$\chi \sin \delta_h$ is from the Sim distribution (Sim, 1959) related to the heavy-atom substructure, $\Delta \varphi_{h \text{best}}$ is defined as

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

and $m_h$ is the figure of merit for $E_h$ and is expressed as...
The value of $P_s(\Delta \varphi)$ in (7) and (8) is initially set to 1/2 and will be updated in each cycle during the phase iteration. The term $\exp(-\sigma_h^2/2)$ in (8) is related to the lack-of-closure error, in which

$$m_h = \exp(-\sigma_h^2/2)(2|P_s(\Delta \varphi)| - \frac{1}{2} + \frac{1}{2}) 	imes (1 - \cos 2\Delta \varphi_h + \cos 2\Delta \varphi_h)^2.$$  

(8)

and $D$ is the lack-of-closure error defined by Blow & Crick (1959). Actually, $\exp(-\sigma_h^2/2)$ is nothing other than a weighting function associated with individual reflections. For details of (4)–(9) the reader is referred to Woolfson & Fan (1995a,b). In the program OASIS (Hao et al., 2000), the term $\sin^2 \Delta \varphi$ in (9) is replaced by its averaged value $\langle \sin^2 \Delta \varphi \rangle = 1/2$ and the lack-of-closure error $D$ is replaced by a constant to be set empirically for each particular case. Thus,

$$\sigma_h^2 = \frac{K^2}{2|F_h'|^2},$$  

(10)

It has been found that failures of OASIS are often a consequence of an improper choice of the constant $K$. In this paper a modified expression is given, i.e.

$$\sigma_h^2 = \frac{(n\sigma_{\Delta F_h})^2}{2|F_h'|^2},$$  

(11)

where $\sigma_{\Delta F_h}$ is the measured standard deviation of the Bijvoet difference $\Delta F_h$ and $n$ is a scaling factor. A method for automatically tuning the scaling factor $n$ is also proposed. The above modification avoids manual intervention and leads to good phasing results for a wide variety of sample proteins.

2. Test samples

Experimental SAD data from eight known proteins (summarized in Table 1) were used in the test calculation. The size of proteins ranges from 129 to 560 residues per asymmetric unit. Anomalous scattering atoms in the sample proteins are S, Cu, Se, Br and Xe. The Bijvoet ratio changes from 0.68 to 8.78%.

3. Influence of $\langle \exp(-\sigma_h^2/2) \rangle$ on the phasing error

From (8) and (11) it is evident that the overall average value of $\exp(-\sigma_h^2/2)$ would affect the resultant overall average phase error. In the following test, for each sample protein, by tuning the value $n$ in (11) we set $\langle \exp(-\sigma_h^2/2) \rangle$ equal to 0.25, 0.50, 0.75 and 1.00 in turn. Phases were derived for each condition by OASIS and then improved using the program DM from the CCP4 suite (Collaborative Computational Project, Number 4, 1994). Results are summarized in Table 2. We observe the following.

(i) For all sample proteins, the condition corresponding to the lowest overall average phase error is not...
h exp \ldots \beta = 2 \hat{\beta} i
to obtain more accurate phases.

(ii) The value \( n \) corresponding to the lowest overall average phase error changes from one sample to the other. This means that the value \( n \) should be fine-tuned in each case. In practice, this would be better performed automatically.

(iii) On average the condition \( h \exp \ldots \beta = 2 \hat{\beta} i \) led to the best phasing result. This can be seen more clearly in Fig. 1, where the overall average phase error is plotted against \( h \exp \ldots \beta = 2 \hat{\beta} i \) for each sample protein.

A simple algorithm has been added to the program OASIS that ensures that the condition \( h \exp \ldots \beta = 2 \hat{\beta} i = 0.5 \) is satisfied for various types of input SAD data by automatically tuning the scaling factor \( n \). A keyword is also provided so that the user can set \( h \exp \ldots \beta = 2 \hat{\beta} i \) equal to any value in the range 0–1. SAD data for all the eight sample proteins were successfully phased by a default run of the modified OASIS followed by a default run of DM. As pointed out by Ramagopal et al. (2003), SAD phasing with a default run of the program OASIS (Hao et al., 2000) failed to solve the structure of xylanase. However in the present test, an easily traceable electron-density map can be obtained based on the SAD phasing with the modified OASIS (see Fig. 2).

4. Modification to the program OASIS

A simple algorithm has been added to the program OASIS that ensures that the condition \( \langle \exp(\epsilon_\beta^2/2) \rangle = 0.5 \) is satisfied for various types of input SAD data by automatically tuning the scaling factor \( n \). A keyword is also provided so that the user can set \( \langle \exp(\epsilon_\beta^2/2) \rangle \) equal to any value in the range 0–1. SAD data for all the eight sample proteins were successfully phased by a default run of the modified OASIS followed by a default run of DM. As pointed out by Ramagopal et al. (2003), SAD phasing with a default run of the program OASIS (Hao et al., 2000) failed to solve the structure of xylanase. However in the present test, an easily traceable electron-density map can be obtained based on the SAD phasing with the modified OASIS (see Fig. 2).

Table 1
Summary of the test data.

| Protein                  | Residues in AU | Anomalous scatterer | \( \lambda (\text{Å}) \) | \( f^0 (\text{e}) \) | \( \langle |\Delta F|/F \rangle (\%) \) | Multiplicity | Reference         |
|--------------------------|----------------|---------------------|-------------------------|----------------------|-----------------------------|--------------|------------------|
| Glucose isomerase        | 388            | S (9)               | 1.54                    | 0.56                 | 0.68                        | 17.4         | Ramagopal et al. (2003) |
| Xylanase                 | 303            | S (5)               | 1.74                    | 0.70                 | 0.69                        | 12.0         | Ramagopal et al. (2003) |
| Azurin                   | 129            | Cu (1)              | 0.97                    | 2.21                 | 1.44                        | 10.0         | Dodd et al. (1995)    |
| Rusticyanin              | 155            | Cu (1)              | 1.376                   | 3.88                 | 2.36                        | 10.2         | Harvey et al. (1998)  |
| Histone methyltransferase SET7/9 | 560          | Se (12)             | 0.9794                  | 5.60                 | 7.03                        | 3.8          | Wilson et al. (2002)  |
| KD93                     | 188            | Se (5)              | 0.9712                  | 3.79                 | 6.60                        | 23.0         | Chen et al. (2004)    |
| Human acyl protein thioesterase | 464        | Br (22)             | 0.9167                  | 5.0                  | 8.78                        | 3.7          | Devedjiev et al. (2000) |
| Porcine pancreatic elastase | 240          | Xe (1)              | 2.1                     | 11.8                 | 5.76                        | 4.0          | Mueller-Dieckmann et al. (2004) |

Table 2
Summary of test results.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Anomalous scatterer</th>
<th>Bijvoet ratio (%)</th>
<th>( n )</th>
<th>Error (%)</th>
<th>( n )</th>
<th>Error (%)</th>
<th>( n )</th>
<th>Error (%)</th>
<th>( n )</th>
<th>Error (%)</th>
<th>( n )</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucose isomerase</td>
<td>S (9)</td>
<td>0.68</td>
<td>1.05</td>
<td>53.07</td>
<td>0.55</td>
<td>41.21</td>
<td>0.27</td>
<td>42.80</td>
<td>0.00</td>
<td>57.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xylanase</td>
<td>S (5)</td>
<td>0.69</td>
<td>1.09</td>
<td>56.62</td>
<td>0.57</td>
<td>52.66</td>
<td>0.28</td>
<td>56.11</td>
<td>0.00</td>
<td>65.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Azurin</td>
<td>Cu (1)</td>
<td>1.44</td>
<td>0.61</td>
<td>62.18</td>
<td>0.30</td>
<td>59.79</td>
<td>0.13</td>
<td>63.15</td>
<td>0.00</td>
<td>70.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rusticyanin</td>
<td>Cu (1)</td>
<td>2.36</td>
<td>1.42</td>
<td>52.70</td>
<td>0.77</td>
<td>49.99</td>
<td>0.35</td>
<td>49.42</td>
<td>0.00</td>
<td>57.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histone methyltransferase SET7/9</td>
<td>Sc (12)</td>
<td>7.03</td>
<td>3.47</td>
<td>44.33</td>
<td>1.83</td>
<td>44.66</td>
<td>0.88</td>
<td>41.31</td>
<td>0.00</td>
<td>44.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KD93</td>
<td>Sc (5)</td>
<td>6.60</td>
<td>1.46</td>
<td>39.80</td>
<td>0.59</td>
<td>42.09</td>
<td>0.21</td>
<td>43.58</td>
<td>0.00</td>
<td>49.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human acyl protein thioesterase</td>
<td>Br (22)</td>
<td>8.78</td>
<td>1.77</td>
<td>53.22</td>
<td>0.90</td>
<td>52.08</td>
<td>0.43</td>
<td>53.77</td>
<td>0.00</td>
<td>62.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Porcine pancreatic elastase</td>
<td>Xe (1)</td>
<td>5.76</td>
<td>3.23</td>
<td>48.01</td>
<td>1.70</td>
<td>45.67</td>
<td>0.83</td>
<td>44.31</td>
<td>0.00</td>
<td>52.34</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

† Bijvoet ratio = \( \langle |F^+ + F^-|/2 \rangle \).
5. Discussion

The weighting function $\exp(-\sigma_h^2/2)$ was introduced by expressing the bimodal SAD phase distribution with the sum of two Gaussian functions taking into account the lack-of-closure error (see Fan, Han & Qian 1984). Introducing the lack-of-closure error into direct methods is important both in theory and in practice of the direct-method SAD phasing. As is seen from the test cases listed in Table 2, $\exp(-\sigma_h^2/2)$ is essential to a successful direct-method SAD phasing. For the calculation of $h$ we need the experimentally measured standard deviation of Bijvoet differences. However, when we wrote the previous version of OASIS, protein diffraction data sets were not always provided with such measurements or they were not accurate enough. Hence, we used (10) instead of (9) in the program. Now with the technical advances in data collection and treatment we can replace (10) with (11), where the experimental standard deviation $\sigma_{\Delta h}$ is included. The constant $n$ preceding $\sigma_{\Delta h}$ is for the compensation of some uncertain systematic errors. In Table 2, it is seen that in order to obtain the best phases, $n$ should be set to different value for different sample. Tuning the value of $n$ is thus inevitable in each particular phasing process. It is found from Fig. 1 that on average the condition $\exp(-\sigma_h^2/2) \approx 0.5$ leads to the best phasing result. The phasing procedure has been made more efficient by automatically tuning the scaling factor $n$ to satisfy the condition. The method proposed in this paper is also useful in dealing with single-isomorphous replacement (SIR) data.

FHF should like to thank Dr Z. Dauter, Dr M. S. Weiss, Dr S. J. Gamblin, Dr B. Xiao and Dr S. S. Hasnain for kindly providing experimental SAD data. This work is supported by the Innovation Project of the Chinese Academy of Sciences and the 973 Project (Grant Nos. G1999075604 and 2002CB713801) of the Ministry of Science and Technology of China.

References