Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Direct-method SAD phasing of proteins enhanced by the use of intrinsic bimodal phase distributions in the subsequent phase-improvement process

**Li-Jie Wu,[a] Tao Zhang,[a,b] Yuan-Xin Gu,[a] Chao-De Zheng[a] and Hai-Fu Fan[a]***

[a]Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, People's Republic of China, and [b]School of Physical Sciences and Technology, Lanzhou University, Gansu, Lanzhou 730000, People's Republic of China

Correspondence e-mail: fanhf@cryst.iphy.ac.cn

Received 11 April 2009
Accepted 11 September 2009

A modified SAD (single-wavelength anomalous diffraction) phasing algorithm has been introduced in the latest version of the program *OASIS*. In addition to direct-method phases and figures of merit, Hendrickson–Lattman coefficients that correspond to the original unresolved bimodal phase distributions are also output and used in subsequent phase-improvement procedures in combination with the improved phases. This provides the possibility of rebreaking the SAD phase ambiguity using the ever-improving phases resulting from the phase-improvement process. Tests using experimental SAD data from six known proteins showed that in all cases the new treatment produced significant improved results.

## 1. Introduction

SAD phasing is nowadays the method of choice in solving *de novo* protein structures. However, the problem of phase ambiguity is intrinsic to the SAD method, *i.e.* SAD experiments do not lead to unique phase estimations for individual reflections but instead just to a phase doublet. Direct methods have proved to be very efficient in breaking the SAD phase ambiguity (Watanabe *et al.*, 2005; Yao *et al.*, 2006). The relevant theory and practice has been summarized and discussed by Yao *et al.* (2008). On the other hand, since anomalous diffraction signals are rather weak, initial phases from SAD methods are accompanied by large errors. It is essential to improve the initial phases (Sha *et al.*, 1995) by using some kind of phase-improvement procedure, in particular the solvent-flattening technique (Wang, 1985). A typical implementation of an *OASIS*-aided SAD phasing and phase-improvement process is the iteration of *OASIS*, *DM* (Cowtan & Main, 1993; Collaborative Computational Project, Number 4, 1994) and a model-building program such as *ARP*/*wARP* (Perrakis *et al.*, 1999), *RESOLVE* (Terwilliger, 2000) or *AutoBuild* (Terwilliger *et al.*, 2008) in *PHENIX* (Adams *et al.*, 2002). In one of our previous papers (Wang *et al.*, 2004), we stated

> the Hendrickson–Lattman coefficients are also output and can be used by the subsequent density modification.

At the time, what we meant by 'the Hendrickson–Lattman coefficients' were those corresponding to the direct-method phases and figures of merit. We subsequently found that this is unnecessary, since either *DM* or *RESOLVE* will calculate such Hendrickson–Lattman (HL) coefficients (Hendrickson & Lattman, 1970) from the input direct-method phases and figures of merit in the absence of input HL coefficients. Hence, up to *OASIS*-2006 (Zhang *et al.*, 2007) no HL coefficients are

**Table 1**
Test samples.

| Protein | No. of residues per AU | Space group | X-ray wavelength (Å) | Anomalous scatterers per AU | $\langle\lvert\Delta F\rvert\rangle/\langle F\rangle$ (%) | Data multiplicity | Solvent content (%) | Resolution limit (Å) | Reference |
|---|---|---|---|---|---|---|---|---|---|
| Azurin | 129 | $P4_122$ | 0.97 | 1 × Cu | 1.45 | 10.0 | 0.45 | 1.9 | Dodd *et al.* (1995) |
| Set7/9 | 586 | $P2_12_12_1$ | 0.9794 | 12 × Se | 7.03 | 3.8 | 0.50 | 2.8 | Wilson *et al.* (2002) |
| Tom70p | 1086 | $P2_1$ | 0.9789 | 24 × Se | 4.35 | 3.3 | 0.45 | 3.3 | Wu & Sha (2006) |
| TT0570 | 1206 | $P2_12_12$ | 1.5418 (Cu $K\alpha$) | 22 × S | 0.57 | 29.2 | 0.47 | 2.1 | Watanabe *et al.* (2005) |
| TTHA1012 | 213 | $P2_12_12_1$ | 2.291 (Cr $K\alpha$) | 2 × S | 0.83 | 13.5 | 0.47 | 2.2 | PDB code 2yzy† |
| Xylanase | 303 | $P2_1$ | 1.49 | 5 × S | 0.56 | 15.9 | 0.37 | 1.8 | Ramagopal *et al.* (2003) |

† A. Ebihara, N. Watanabe, S. Yokoyama & S. Kuramitsu (unpublished work).

**Table 2**
Phase errors (°) in different phasing stages.

| | | *DM* | | *AutoBuild* | |
|---|---|---|---|---|---|
| Protein | *OASIS* | (i) | (ii) | (i) | (ii) |
| Azurin | 60.9 | 47.1 | 48.2 | 34.9 | 47.1 |
| Set7/9 | 48.7 | 27.5 | 31.5 | 22.3 | 28.6 |
| Tom70p | 62.3 | 45.0 | 50.4 | 39.9 | 43.9 |
| TT0570 | 61.1 | 48.8 | 50.6 | 21.4 | 41.2 |
| TTHA1012 | 62.2 | 52.7 | 53.7 | 34.8 | 52.8 |
| Xylanase | 62.5 | 56.9 | 57.4 | 31.8 | 56.3 |

actually output in addition to the direct-method phases and figures of merit. The program *OASIS* has now been modified to enable the calculation and output of HL coefficients that correspond to unresolved bimodal phase distributions from a SAD experiment. For this purpose, the subroutine *HENDFT.F* of the program *SOLVE* (Terwilliger & Berendzen, 1999) has been incorporated into *OASIS*. The HL coefficients are calculated according to the formula

$$P_{\rm anom}(\varphi_{\mathbf{h}}) = N\exp\{-[\Delta F - 2F''\sin(\varphi_{\mathbf{h}} - \varphi'_h)]^2/2E^2\} \quad (1)$$

This is the same formula as used in the solution of the structure of crambin (Hendrickson & Teeter, 1981), where $\varphi_{\mathbf{h}}$ is the phase of $\langle\mathbf{F}_{\mathbf{h}}\rangle$, which is equal to $(\mathbf{F}^+_{\mathbf{h}} + \mathbf{F}^-_{\mathbf{h}})/2 = [(|F^+ + F^-|)/2]\exp(i\varphi_{\mathbf{h}})$, $N$ is the normalizing factor, $\Delta F = F^+ - F^-$, $F''$ is the structure-factor magnitude contributed from the imaginary-part scattering of the anomalous scatterers, $\varphi'_{\mathbf{h}}$ is the phase contributed from the real-part scattering of the anomalous scatterers and $E$ is the standard error. These HL coefficients will be output directly to the subsequent phase-improvement programs. Hence, in the latest version of *OASIS* two different kinds of phase information are output simultaneously. One is the result of SAD phasing expressed as the direct-method phases and figures of merit, while the other is the original unresolved bimodal SAD-phase distribution expressed as HL coefficients. During the subsequent phase-improvement process, the former is used to calculate an electron-density map and is to be improved, while the latter is used for combination with the improved result. Doing this implies that experimental SAD-phase ambiguities are rebroken in each stage of improvement by a set of ever-improving phases. In the following, it will be seen how the power of SAD phasing is enhanced by this treatment.

## 2. Test data

SAD data from the six proteins listed in Table 1 were used in the test. They cover selenium, copper and sulfur SAD data collected using synchrotron radiation, Cr $K\alpha$ and Cu $K\alpha$ X-rays. The size of the proteins ranges from 129 to 1206 amino acids per AU (asymmetric unit). The high-resolution limit of the data is far lower than 'atomic resolution', ranging from 1.8 to 3.3 Å. The sample data represent situations of varying difficulty arising from various unfavourable features. The azurin data have an overall completeness of only ∼60%. Set7/9 has a high-resolution limit of 2.8 Å and a multiplicity of 3.8. Tom70p has a high-resolution limit of 3.3 Å with a multiplicity as low as 3.3 and the crystal has suffered serious radiation damage during data collection. The other three data sets are all difficult sulfur SAD data with very low Bijvoet ratios ($\langle|\Delta F|\rangle/\langle F\rangle$). Of the data sets, Tom70p, TTHA1012 and xylanase are three extremely difficult cases. In particular, xylanase has a solvent content of only 37%. The data set was concluded by Ramagopal *et al.* (2003) as lacking 'sufficient phasing power to produce interpretable electron-density maps'. Structures of all the sample proteins are already known. Calculations of phase errors were made against the known structure model given by the original authors.

## 3. Test and results

Test calculations for each set of data passed through three steps: direct-method SAD phasing using *OASIS*, density modification using *DM* (Cowtan & Main, 1993; Collaborative Computational Project, Number 4, 1994) and model building and refinement using *AutoBuild* (Terwilliger *et al.*, 2008) in *PHENIX* (Adams *et al.*, 2002). From the second step onward, the calculations were split into two parallel paths: (i) both HL coefficients and direct-method phases and figures of merit obtained from *OASIS* were input to *DM* and *AutoBuild*, and (ii) only direct-method phases and figures of merit from *OASIS* were input to *DM* and *AutoBuild*. Results from the two paths are listed in Tables 2 and 3 under the column headings (i) and (ii), respectively. Table 2 shows overall averaged phase errors (weighted by $mF_{\rm o}$, where $m$ is the figure of merit and $F_{\rm o}$ is the observed structure-factor magnitude) in different phasing stages for the six test proteins. In all cases, the phase errors that result from *DM* and *AutoBuild* through path (i) are obviously lower than those through path (ii). Table 3 shows

**Table 3**
Results of model building by *AutoBuild* in *PHENIX*.

| | $R$† (%) | | $R_{free}$‡ (%) | | No. of built residues§ | | No. of assigned residues¶ | | Built percentage†† (%) | |
|---|---|---|---|---|---|---|---|---|---|---|
| Protein | (i) | (ii) | (i) | (ii) | (i) | (ii) | (i) | (ii) | (i) | (ii) |
| Azurin | 38 | 44 | 42 | 44 | 87 | 86 | 17 | 0 | 67 | 67 |
| Set7/9 | 24 | 26 | 27 | 31 | 548 | 478 | 530 | 410 | 94 | 82 |
| Tom70p | 31 | 36 | 37 | 40 | 859 | 748 | 66 | 50 | 79 | 69 |
| TT0570 | 19 | 30 | 23 | 33 | 1182 | 1065 | 1156 | 881 | 98 | 88 |
| TTHA1012 | 33 | 49 | 38 | 53 | 128 | 122 | 78 | 11 | 60 | 57 |
| Xylanase | 16 | 46 | 20 | 48 | 309‡‡ | 175 | 301 | 0 | 100 | 58 |

† $R$ is the residual factor, $R = \sum_h w||F_o| - |F_c||/\sum_h w|F_o|$. ‡ $R_{free}$ is the residual factor calculated from a randomly selected subset of reflections which are not involved in refinement of the structural model. § The number of residues built in the structure model. ¶ The number of residues in the structure model which have been assigned into the amino-acid sequence. †† The ratio between the number of built residues and the total number of residues in the AU expressed as a percentage. ‡‡ The *AutoBuild* output for xylanase consists of the total of 303 residues in the AU plus six residues in two separate short chains. The latter may come from ghost densities in the map input to model building.

the results of model building using *AutoBuild* in *PHENIX*. Again, the use of HL coefficients output from *OASIS* led to significantly better results. Note that while the *OASIS*-aided model completion (involving *OASIS*, *DM* and *AutoBuild*) was running in iterative mode, only results from the first cycle of iteration are shown in Table 3. It can be seen in Table 3 that for the data sets azurin, Tom70p and TTHA1012 a common feature of the results is the low built percentage (<80%) and the large difference between the number of built residues and the number of assigned residues. This may be a consequence of certain unfavourable features of the data sets. For azurin, the low overall completeness (∼60%) is probably the reason. Nevertheless, one more cycle of path (i) iteration led to a model that consisted of 121 residues all assigned into the sequence and the built percentage increased from 67 to 94%. For TTHA1012, one more cycle of path (i) iteration resulted in a model consisting of 160 residues with 141 assigned into the sequence and the built percentage increased from 60 to 75%. Finally, for Tom70p, as the high-resolution limit is rather low (3.3 Å) *AutoBuild* was run in the 'helices_strands_only' mode. This causes the large difference between the number of built residues and the number of assigned residues. Two more cycles of path (i) iteration resulted in a slight decrease in the number of built residues (from 859 to 849). However, the number of assigned residues increased from 66 to 151. While a large difference remains between the number of built residues and the number of assigned residues, the result still provides an adequate starting point for successful manual model completion.

## 4. Concluding remarks

The HL coefficients of unresolved bimodal phase distributions preserve information about the intrinsic SAD-phase ambiguity. The combination of such HL coefficients with improved phases in each stage of phase improvement implies that the original phase ambiguities are rebroken in each stage with the ever-improving phases. This is reasonable in phasing philosophy, while its efficiency in practice has been proved by the significant improvement of results with the representative test samples. On the other hand, when the model becomes sufficiently large the combination of improved phases with the unresolved bimodal phase distributions may be equivalent to applying a kind of damping factor to the phase-improvement process. This may slow the convergence rate. Hence, a criterion has been set in the new version of *OASIS* that only when the ratio of number of assigned residues to the number of total residues in the AU is smaller than 30% (this value can be changed by the user) will the output HL coefficients from *OASIS* be input to subsequent phase-improvement processes. The new version of *OASIS* will be available on the web at http://cryst.iphy.ac.cn in due course.

The authors are grateful to Dr T. C. Terwilliger for his very kind permission for the incorporation of the subroutine *HENDFT.F* in *OASIS*. Thanks are also due to Professor S. Hasnain for making available the SAD data of azurin, Dr B. Xiao for the data for Set7/9, Dr B.-D. Sha for the data for Tom70p, Professor N. Watanabe for the data for TT0570 and TTHA1012 and Dr Z. Dauter for the data for xylanase. 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

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1948–1954.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. D. & Main, P. (1993). *Acta Cryst.* D**49**, 148–157.
Dodd, F. E., Hasnain, S. S., Abraham, Z. H. L., Eady, R. R. & Smith, B. E. (1995). *Acta Cryst.* D**51**, 1052–1064.
Hendrickson, W. A. & Lattman, E. E. (1970). *Acta Cryst.* B**26**, 136–143.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Ramagopal, U. A., Dauter, M. & Dauter, Z. (2003). *Acta Cryst.* D**59**, 1020–1027.
Sha, B.-D., Liu, S. P., Gu, Y. X., Fan, H. F., Ke, H. M., Yao, J. X. & Woolfson, M. M. (1995). *Acta Cryst.* D**51**, 342–346.
Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L.-W., Read, R. J. & Adams, P. D. (2008). *Acta Cryst.* D**64**, 61–69.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
Wang, J. W., Chen, J. R., Gu, Y. X., Zheng, C. D., Jiang, F., Fan, H. F., Terwilliger, T. C. & Hao, Q. (2004). *Acta Cryst.* D**60**, 1244–1253.
Watanabe, N., Kitago, Y., Tanaka, I., Wang, J., Gu, Y., Zheng, C. & Fan, H. (2005). *Acta Cryst.* D**61**, 1533–1540.
Wilson, J. R., Jing, C., Walker, P. A., Martin, J. R., Howell, S. A., Blackburn, G. M., Gamblin, S. J. & Xiao, B. (2002). *Cell*, **111**, 105–115.
Wu, Y. & Sha, B. (2006). *Nature Struct. Mol. Biol.* **13**, 589–593.

Yao, D., Huang, S., Wang, J., Gu, Y., Zheng, C., Fan, H., Watanabe, N. & Tanaka, I. (2006). *Acta Cryst.* D**62**, 883–890.
Yao, D.-Q., Li, H., Chen, Q., Gu, Y.-X., Zheng, C.-D., Lin, Z.-J., Fan, H.-F., Watanabe, N. & Sha, B.-D. (2008). *Chin. Phys. B*, **17**, 1–9.
Zhang, T., He, Y., Gu, Y.-X., Zheng, C.-D. Hao, Q., Wang, J.-W. & Fan, H.-F. (2007). *OASIS-2006: A Direct-methods Program for SAD/SIR Phasing and Reciprocal-space Fragment Extension*. Institute of Physics, Chinese Academy of Sciences, People's Republic of China. http://cryst.iphy.ac.cn/Project/program/oasis.html.