Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Direct-method SAD phasing with partial-structure iteration: towards automation

J. W. Wang, J. R. Chen, Y. X. Gu, C. D. Zheng and H. F. Fan*

Institute of Physics, Chinese Academy of Sciences, Beijing 100080, People's Republic of China

Correspondence e-mail: fan@mail.iphy.ac.cn

Received 8 September 2004
Accepted 23 September 2004

The probability formula of direct-method SAD (single-wavelength anomalous diffraction) phasing proposed by Fan & Gu (1985, *Acta Cryst.* A**41**, 280–284) contains partial-structure information in the form of a Sim-weighting term. Previously, only the substructure of anomalous scatterers has been included in this term. In the case that the subsequent density modification and model building yields only structure fragments, which do not straightforwardly lead to the complete solution, the partial structure can be fed back into the Sim-weighting term of the probability formula in order to strengthen its phasing power and to benefit the subsequent automatic model building. The procedure has been tested with experimental SAD data from two known proteins with copper and sulfur as the anomalous scatterers.

1. Introduction

Fan and coworkers (Fan, Han & Qian, 1984; Fan, Han, Qian *et al.*, 1984; Fan & Gu, 1985) proposed a direct method of breaking the phase ambiguity intrinsic in single isomorphous replacement (SIR) or single-wavelength anomalous diffraction (SAD). The method was first successfully applied to the experimental SAD data from the known protein aPP (Fan *et al.*, 1990). The program *OASIS* (Hao *et al.*, 2000) is based on the principle of Fan and coworkers (Fan, Han & Qian, 1984; Fan, Han, Qian *et al.*, 1984; Fan & Gu, 1985) and the practical implementation of Fan *et al.* (1990). *OASIS* has been tested with a series of known proteins (see Wang *et al.*, 2004 and references therein) and applied to solve a number of originally unknown proteins (Harvey *et al.*, 1998; Huang *et al.*, 2004; Chen *et al.*, 2004). The program is now being revised and a new version, *OASIS*-2004, will be released in due course. The major improvements in the new version will be discussed in a series of forthcoming papers. Here, the incorporation of partial-structure iteration into the direct-method phasing procedure will be discussed in detail.

2. Method

The intrinsic phase ambiguity in SAD data is expressed as

$$\varphi = \langle\varphi\rangle \pm |\Delta\varphi|, \qquad (1)$$

where φ is the phase associated with the average magnitude

$$\langle F\rangle = (F^{+} + F^{-})/2. \qquad (2)$$

$\langle\varphi\rangle$ is the mean value of the phase doublet, which in the SAD case equals φ'', *i.e.* the phase of

DOI: 10.1107/S0907444904023893

$$F''(\mathbf{h}) = i \sum_{j=1}^{N} f''_j \exp(i2\pi\mathbf{h} \cdot \mathbf{r}_j), \qquad (3)$$

and $|\Delta\varphi|$ is the absolute difference between $\langle\varphi\rangle$ and φ.

The value of $|\Delta\varphi|$ in the SAD case can be calculated as (see Blundell & Johnson, 1976)

$$|\Delta\varphi| \simeq \cos^{-1}[(F^+ - F^-)/2|F''|]. \qquad (4)$$

The direct-method SAD phasing procedure is based on the probability of $\Delta\varphi$ being positive, which is expressed as

$$P_+(\Delta\varphi_\mathbf{h}) = \frac{1}{2} + \frac{1}{2}\tanh\left\{\sin|\Delta\varphi_\mathbf{h}|\left[\sum_{\mathbf{h}'} m_{\mathbf{h}'} m_{\mathbf{h}-\mathbf{h}'} \kappa_{\mathbf{h},\mathbf{h}'} \times \sin(\Phi'_3 + \Delta\varphi_{\mathbf{h}'\mathrm{best}} + \Delta\varphi_{\mathbf{h}-\mathbf{h}'\mathrm{best}}) + \chi\sin\delta_\mathbf{h}\right]\right\}. \qquad (5)$$

For details of this formula the reader is referred to Fan & Gu (1985). In (5), the term $\sum_{\mathbf{h}'} m_{\mathbf{h}'} m_{\mathbf{h}-\mathbf{h}'} \kappa_{\mathbf{h},\mathbf{h}'} \sin(\Phi'_3 + \Delta\varphi_{\mathbf{h}'\mathrm{best}} + \Delta\varphi_{\mathbf{h}-\mathbf{h}'\mathrm{best}})$ comes from the Cochran distribution (Cochran, 1955), while $\chi\sin\delta_\mathbf{h}$ comes from the Sim distribution (Sim, 1959) with

$$\chi = 2E_\mathbf{h}E_{\mathbf{h},\mathrm{partial}} \bigg/ \left(\sum_i^{N_\mathrm{unknown}} Z_i^2 \bigg/ \sum_j^{N_\mathrm{total}} Z_j^2\right) \qquad (6)$$

and

$$\delta = \varphi_{\mathbf{h},\mathrm{partial}} - \langle\varphi\rangle. \qquad (7)$$

The result of (5) depends on the balance between the Cochran term and the Sim term. At the beginning of direct-method SAD phasing, $E_{\mathbf{h},\mathrm{partial}}$ and $\varphi_{\mathbf{h},\mathrm{partial}}$ in (6) and (7), respectively, are only contributed from the anomalous scatterers. Hence, the Sim term is weak in comparison with that of Cochran term, in particular when the Bijvoet ratio $\langle|F^+ - F^-|\rangle/\langle(F^+ + F^-)/2\rangle$ is small. In practice, (5) has led to successful SAD phasing for proteins with various kinds of anomalous scatterers (see Wang *et al.*, 2004). However, there have been examples in which the direct-method SAD phases did not lead to an easily interpretable electron-density map. However, even in these cases structure fragments can often be found by automatic model building. While such a partial structure is not sufficient to approach the complete solution using conventional techniques, the contribution of this partial structure, $E_{\mathbf{h},\mathrm{partial}}$ and $\varphi_{\mathbf{h},\mathrm{partial}}$, can considerably enhance the Sim term by feeding it back to (6) and (7) and then to (5). With this feedback information, the phasing power of (5) can be dramatically

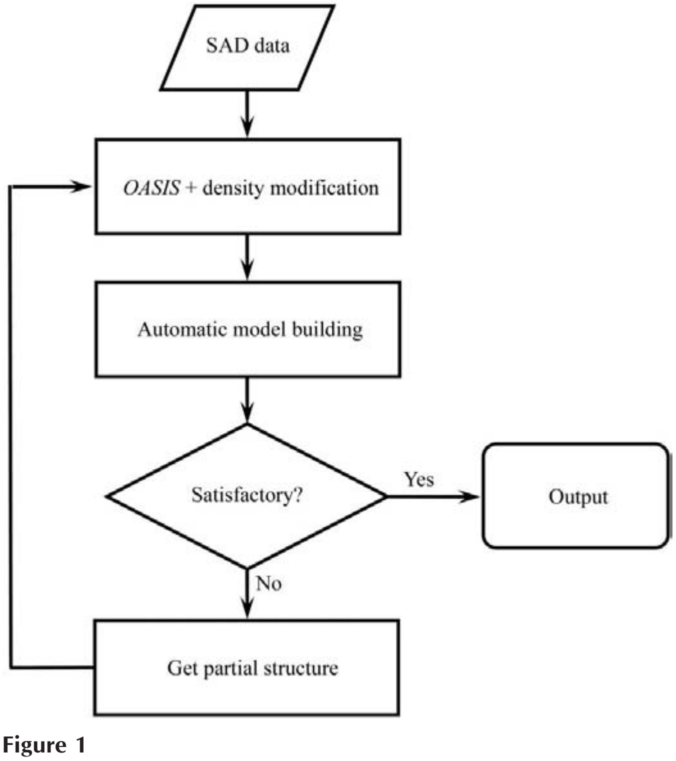

Figure 1
Flowchart of the partial-structure iterative direct-method SAD phasing.

Table 1
Summary of test samples.

	Rusticyanin	Xylanase
Atoms in ASU	1161	2300
Space group	$P2_1$	$P2_1$
Unit-cell parameters (Å, °)	a = 32.43, b = 60.68, c = 38.01, β = 107.82	a = 41.07, b = 67.14, c = 50.81, β = 113.5
Wavelength (Å)	1.376	1.743
Resolution range (Å)	8.0–2.1	25–1.63
Multiplicity	10.2	12.0
Anomalous scatterer	Cu (1) (in centric arrangement)	S (5)
$\Delta f''$	3.88	0.70
$\langle\|\Delta F\|\rangle/\langle F\rangle$ (%)	2.36	0.69

Table 2
Rusticyanin: cumulative phase errors (°) in descending order of F_obs.

	Before iteration		After one cycle of iteration	
No. reflections	Unweighted	F_obs-weighted	Unweighted	F_obs-weighted
500	34.10	33.15	26.64	25.99
1000	34.59	33.93	29.06	28.25
2000	38.22	36.90	33.06	31.58
3000	40.45	38.62	35.33	33.35
4000	42.52	40.09	37.05	34.65
5000	44.49	41.34	39.37	36.10
6000	46.20	42.33	41.92	37.48
7000	48.90	43.43	44.99	38.75
7763	50.94	43.98	47.81	39.47

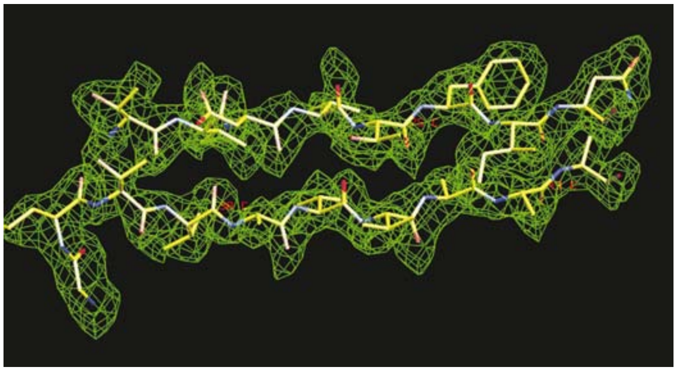

Figure 2
A portion of the electron-density map of rusticyanin after a single-cycle iteration of *OASIS* + *DM*.

strengthened, leading to much improved results. The procedure can be made iterative. The flowchart is shown in Fig. 1.

3. Samples

SAD data from two known proteins, rusticyanin and xylanase, were used as test samples (see Table 1). Rusticyanin was first solved by Walter *et al.* (1996) using MAD data. It was also solved independently by Harvey *et al.* (1998) using a combination of direct-method SAD phasing and the P_s-function method (Hao & Woolfson, 1989). The direct method used by Harvey *et al.* (1998) was the pre-release version of *OASIS*. The resultant electron-density map after density modification can be interpreted with the help of the P_s-function method. Attempts have been made in several laboratories to try phasing the same set of SAD data without using direct methods, but no successful results have been reported so far. The main difficulties in phasing the SAD data of rusticyanin are the low Bijvoet ratio, $\langle|\Delta F|\rangle/\langle F\rangle = 2.36\%$, and in particular the centric arrangement of anomalous scatterers. The latter leads to enantiomorphous ambiguities in SAD phasing if direct methods are not used (see Wang *et al.*, 2004). In the next section it is shown that partial-structure iterative direct-method SAD phasing followed by a default run of *DM* (Cowtan, 1994) from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994) could solve the structure of rusticyanin in a straightforward way. Xylanase was solved by Natesh *et al.* (1999) using the molecular-replacement method. Ramagopal *et al.* (2003) used the 1.63 Å SAD data collected at the wavelength λ = 1.743 Å to test the limit of sulfur-SAD phasing. The SAD data has an extremely low Bijvoet ratio, $\langle|\Delta F|\rangle/\langle F\rangle = 0.69\%$ and the crystal has a low solvent content (37%). Ramagopal *et al.* (2003) succeeded in solving the structure with the program *SHELXE* (Sheldrick, 2002) by tuning some of the input parameters, *i.e.* reducing the solvent fraction to 33% and changing the perturbation value to 2.0 from the default of 1.0. Again, it is seen in the next section that the partial-structure iterative direct-method SAD phasing followed by a default run of *DM* led straightforwardly to the solution.

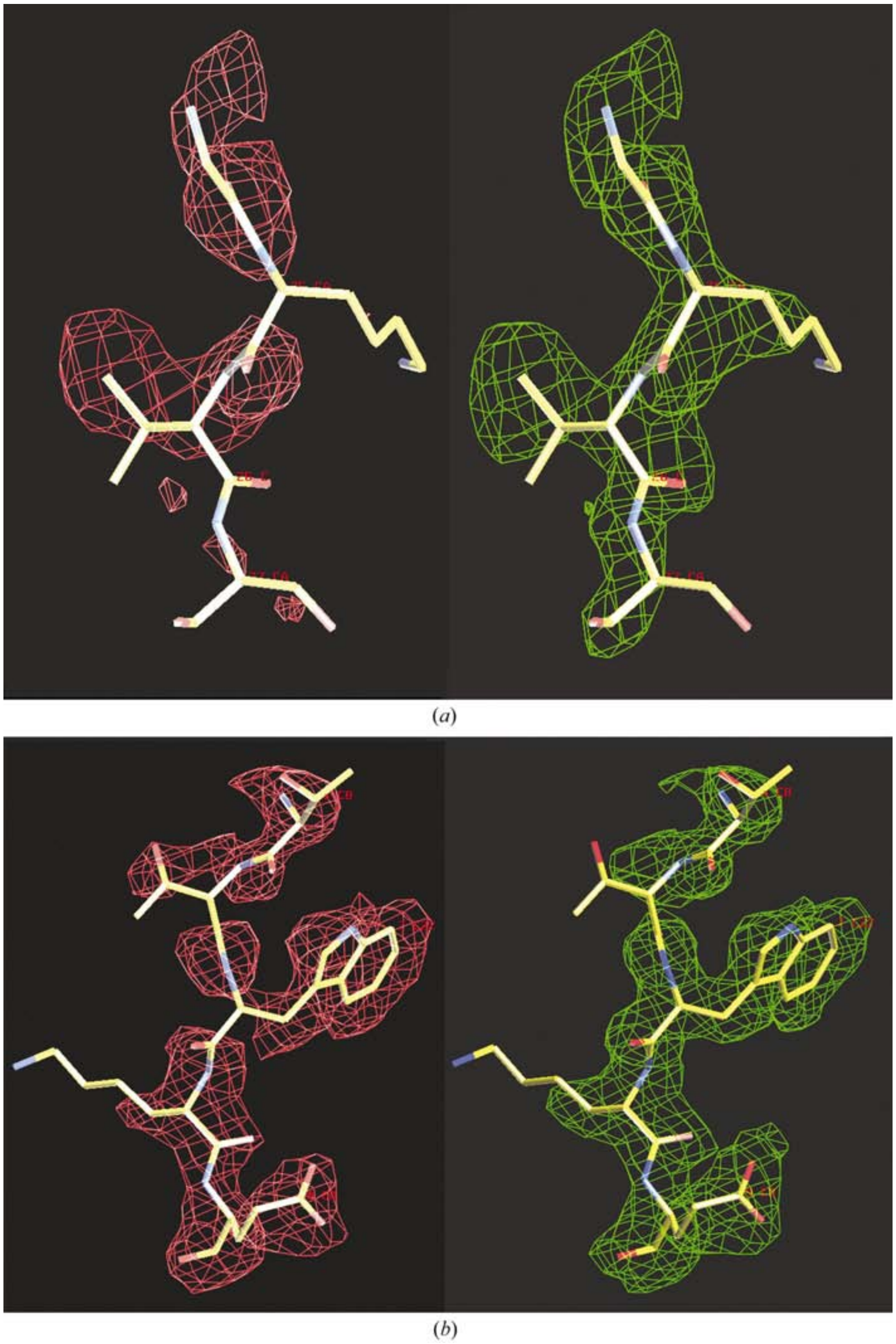

Figure 3
Two portions, (*a*) and (*b*), of the electron-density maps of rusticyanin before (red) and after (green) a single-cycle iteration of *OASIS* + *DM*.

4. Test and results

In the following test *OASIS*-2004 was used to obtain initial SAD phases. *DM* (Cowtan, 1994) was used for subsequent density modification. *RESOLVE* (Terwilliger, 2003*a*,*b*) was only used for model building when the electron-density map contains relatively large errors. *ARP*/*wARP* (Perrakis *et al.*, 1999) was used for model building when the electron-density map was sufficiently accurate. All programs were run with their default controlling parameters.

Table 3
Number of residues found by automatic model building using *RESOLVE BUILD* and *ARP*/*wARP* before and after partial-structure direct-method iteration.

	Before iteration		After iteration
Sample protein	*ARP*/*wARP*	*RESOLVE*	*ARP*/*wARP*
Rusticyanin	8	84	140 (of 155)
Xylanase	12	164	302 (of 303)

4.1. Rusticyanin

Using the SAD data from rusticyanin, a single run of *OASIS*-2004 + *DM* was first performed. The cumulative phase errors of the resultant phases are listed in columns 2 and 3 of Table 2. An electron-density map was then calculated. The map is manually traceable. However, both *ARP*/*wARP* and *RESOLVE* failed in automatic model building. *ARP*/*wARP* could find only eight residues of the total of 155, while *RESOLVE* managed to give 84 residues without side chains. The partial structure from *RESOLVE* is difficult to expand further. However, by feeding back this partial structure to (5), a single-cycle iteration of *OASIS*-2004 + *DM* led to dramatically more accurate phases (see columns 4 and 5 of Table 2). A portion of the corresponding electron-density map is shown in Fig. 2. A comparison of the electron-density maps before and after iteration is shown in Fig. 3. Based on the improved electron-density map, *ARP*/*wARP* automatically built a model containing 140 residues (including side chains) of the total of 155 (see Table 3). Models built before and after iteration are compared in Fig. 4.

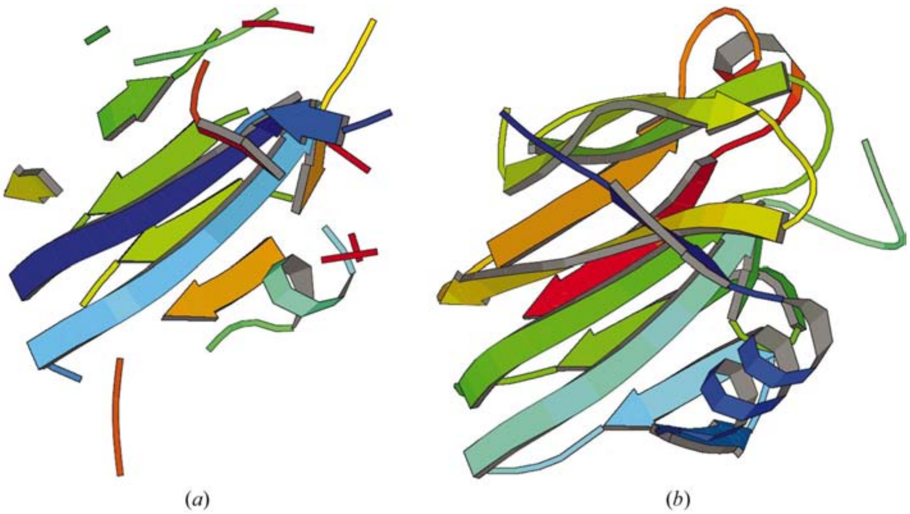

Figure 4
(*a*) Partial model of rusticyanin (without side chains) obtained by *RESOLVE BUILD* based on the resultant phases from a single run of *OASIS* + *DM*. The model contains 88 of the total of 155 residues. (*b*) Structure model of rusticyanin (with side chains) built by *ARP*/*wARP* after a single-cycle iteration of *OASIS* + *DM*. The model contains 140 of the total of 155 residues.

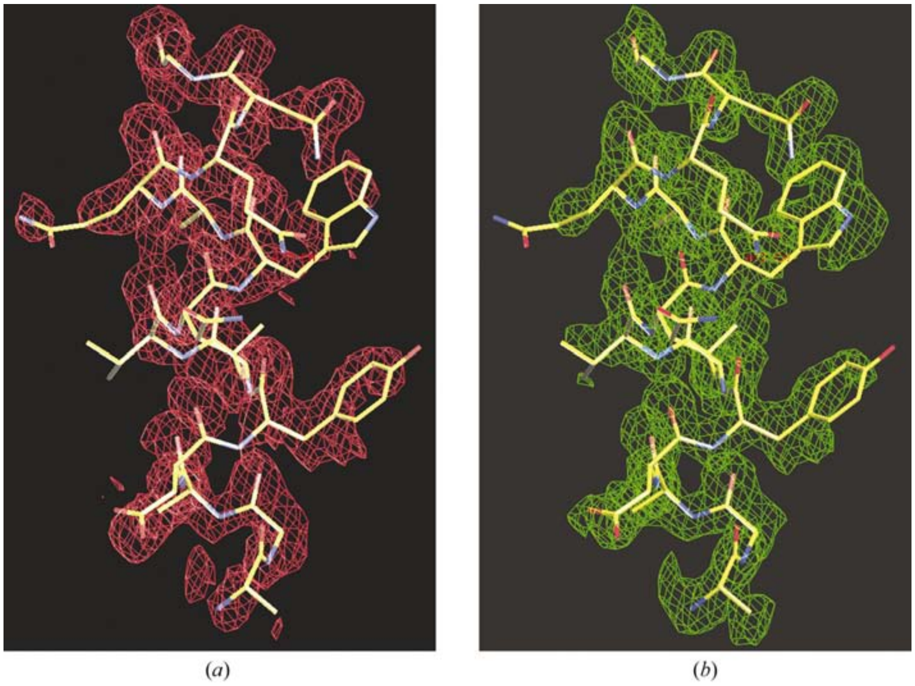

Figure 5
A portion of the electron-density map of xylanase (*a*) before and (*b*) after a single-cycle iteration of *OASIS* + *DM*.

4.2. Xylanase

A single run of *OASIS*-2004 + *DM* resulted in a manually traceable electron-density map, a portion of which is shown in Fig. 5(*a*). However, on automatic model building, *ARP*/*wARP* gave only 12 of the total of 303 residues, while *RESOLVE BUILD* found 164 residues without side chains (see Table 3 and Fig. 6*a*). A single-cycle iteration of *OASIS*-2004 + *DM* based on the *RESOLVE*-built partial model improved the resultant phases dramatically (see Table 4). A portion of the improved electron-density map is shown in Fig. 5(*b*). From the improved electron-density map *ARP*/*wARP* automatically built a model containing 302 of the total of 303 residues including side chains (see Table 3 and Fig. 6*b*). The above test was based on the refined sulfur substructure. An additional test was performed with the same xylanase SAD data based on the unrefined sulfur substructure. The electron-density map resulting from the initial run of *OASIS*-2004 + *DM* contains larger errors than that based

Table 4
Xylanase: cumulative phase errors (°) in descending order of $F_{\rm obs}$.

	Before iteration		After one cycle of iteration	
No. reflections	Unweighted	$F_{\rm obs}$-weighted	Unweighted	$F_{\rm obs}$-weighted
500	36.59	35.75	24.53	23.91
5000	41.41	40.68	32.45	31.47
10000	43.97	42.78	36.11	34.48
15000	46.06	44.33	38.60	36.38
20000	47.73	45.46	40.84	37.87
25000	49.56	46.44	43.24	39.16
30000	52.41	47.34	46.74	40.28
30570	52.86	47.40	47.29	40.35

on the refined sulfur substructure. *RESOLVE BUILD* yielded a partial model (see Fig. 7*a*) that contained only 110 instead of the previous 138 residues without side chains. A single-cycle iteration of *OASIS*-2004 + *DM* led to an *ARP*/*wARP* model containing 172 (including side chains) of the total of 303 residues (see Fig. 7*b*). This is still far from the complete structure. However, a second cycle of iteration of *OASIS*-2004 + *DM* yielded an *ARP*/*wARP* model consisting of 299 residues including side chains, just four residues less than the complete structure (see Fig. 7*c*). This test demonstrates that further cycles of *OASIS*-2004 + *DM* iteration may compensate for larger errors from the anomalous-scatterer substructure.

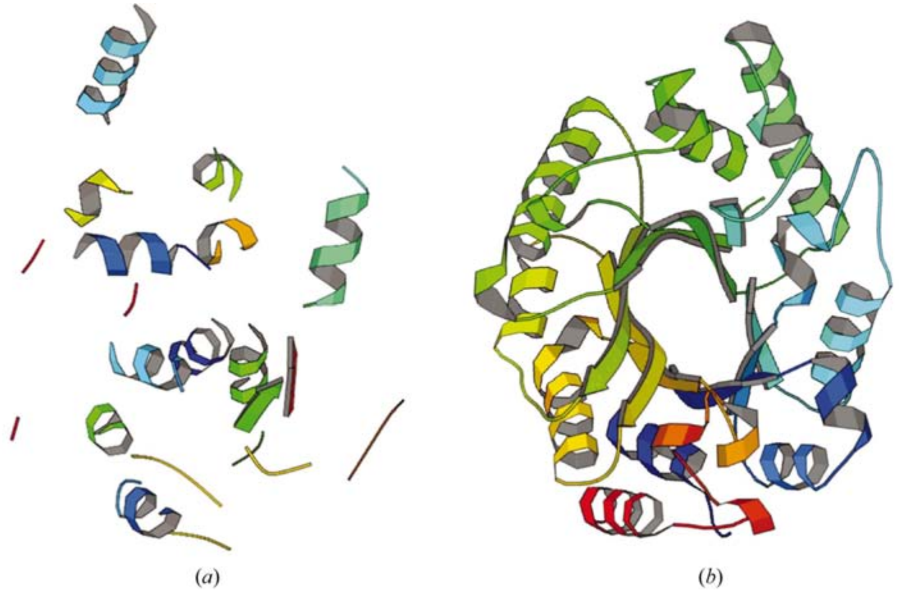

Figure 6
(*a*) Partial model of xylanase (without side chains) obtained by *RESOLVE BUILD* based on the resultant phases from a single run of *OASIS* + *DM*. The model contains 138 of the total of 303 residues. (*b*) Structure model of xylanase (with side chains) built by *ARP*/*wARP* after a single-cycle iteration of *OASIS* + *DM*. The model contains 302 (including side chains) of the total of 303 residues.

5. Discussion

Partial-structure iterative direct-method SAD phasing is much more powerful than single-run direct-method SAD phasing (see Wang *et al.*, 2004 and references therein). It yields much more accurate phases and thus will be beneficial to the automation of model building and thus to the high-throughput structure determination of proteins. In a different context, partial-structure iterative direct-method SAD phasing is a powerful tool for partial-structure expansion if SAD data is available. Unlike the case in the solution of small molecular structures, a fragment less than ~70% of the complete structure will be difficult to expand *via* Fourier recycling. However,

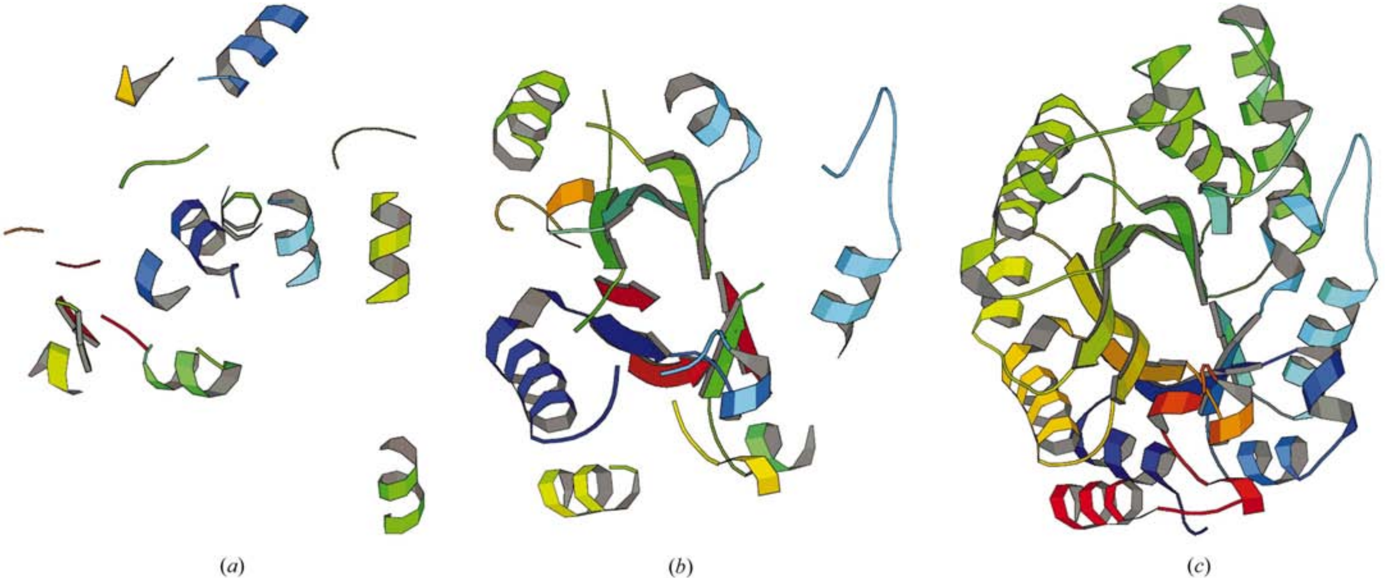

Figure 7
(*a*) Partial model of xylanase (without side chains) obtained by *RESOLVE BUILD* based on the resultant phases from the unrefined sulfur substructure and a single run of *OASIS* + *DM*. The model contains 110 of the total of 303 residues. (*b*) Structure model of xylanase (with side chains) built by *ARP*/*wARP* after a single-cycle iteration of *OASIS* + *DM*. The model contains 172 of the total of 303 residues. (*c*) Structure model of xylanase (with side chains) built by *ARP*/*wARP* after the second cycle of iteration of *OASIS* + *DM*. The model contains 299 of the total of 303 residues.

the partial-structure iterative direct-method SAD phasing can lead to the complete structure with a much smaller starting fragment. The principle proposed in this paper will also be applicable to the SIR case.

HFF would like to thank Drs S. S. Hasnain and Z. Dauter for kindly making available the SAD data of rusticyanin and xylanase, respectively. Thanks are also due to Professor D. C. Liang for helpful discussion. 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

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*, p. 177. London: Academic Press.

Chen, J. R., Gu, Y. X., Zheng, C. D., Jiang, F., Jiang, T., Liang, D. C. & Fan, H. F. (2004). In the press.

Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cowtan, K. (1994). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.

Fan, H. F. & Gu, Y. X. (1985). *Acta Cryst.* A**41**, 280–284.

Fan, H. F., Han, F. S. & Qian, J. Z. (1984). *Acta Cryst.* A**40**, 495–498.

Fan, H. F., Han, F. S., Qian, J. Z. & Yao, J. X. (1984). *Acta Cryst.* A**40**, 489–495.

Fan, H. F., Hao, Q., Gu, Y. X., Qian, J. Z., Zheng, C. D. & Ke, H. (1990). *Acta Cryst.* A**46**, 935–939.

Hao, Q., Gu, Y. X., Zheng, C. D. & Fan, H. F. (2000). *J. Appl. Cryst.* **33**, 980–981.

Hao, Q. & Woolfson, M. M. (1989). *Acta Cryst.* A**45**, 794–797.

Harvey, I., Hao, Q., Duke, E. M. H., Ingledew, W. J. & Hasnain, S. S. (1998). *Acta Cryst.* D**54**, 629–635.

Huang, Q. Q., Liu, Q. & Hao, Q. (2004). In the press.

Natesh, R., Bhanumoorthy, P., Vithayathil, P. J., Sekar, K, Ramakumar, S. & Viswamitra, M. A. (1999). *J. Mol. Biol.* **288**, 999–1012.

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.

Sheldrick, G. M. (2002). *Z. Kristallogr.* **217**, 644–650.

Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.

Terwilliger, T. C. (2003*a*). *Acta Cryst.* D**59**, 38–44.

Terwilliger, T. C. (2003*b*). *Acta Cryst.* D**59**, 45–49.

Walter, R. L., Ealick, S. E., Friedman, A. M., Blake, R. C. II, Proctor, P. & Shoham, M. (1996). *J. Mol. Biol.* **263**, 730–751.

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.