Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

A case study on the treatment of protein SIRAS data

Deqiang Yao,[a,b]‡ Tao Zhang,[c]‡ Yao He,[c]‡ Pu Han,[c] Maia Cherney,[b] Yuanxin Gu,[c] Miroslaw Cygler[b]* and Haifu Fan[c]*

[a]Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China, [b]Department of Biochemistry, University of Saskatchewan, Saskatoon, Saskatchewan S7N 5E5, Canada, and [c]Beijing National Laboratory for Condensed Matter Physics and Key Laboratory of Soft Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, People's Republic of China

‡ The first three authors contributed equally.

Correspondence e-mail: miroslaw.cygler@usask.ca, fanhf@cryst.iphy.ac.cn

Received 2 April 2014
Accepted 26 July 2014

A case study has been made on the treatment of the SIRAS (single isomorphous replacement with anomalous scattering) data of the originally unknown protein LegC3N. An alternative treatment has been proposed which led to improved results in this particular test case. The treatment involves iterative direct-method SAD (single-wavelength anomalous diffraction) phasing and direct-method-aided model completion, both of which are implanted in the *IPCAS* (*Iterative Protein Crystal-structure Automatic Solution*) pipeline. Apart from the experimental data, a simulated SIRAS data set for LegC3N with the derivative data truncated to 5.0 Å resolution has also been tested. SAD phasing and phase/model extension in *PHENIX* without direct methods failed to solve the structure using these simulated SIRAS data. However, the procedure proposed here involving direct methods in both SAD phasing and phase/model extension led to a nearly complete structure model. This shows the potential ability of treating SIRAS data with a derivative diffracting to lower resolution.

1. Introduction

Although the SAD (single-wavelength anomalous diffraction) and MAD (multi-wavelength anomalous diffraction) methods remain major choices for solving protein crystal structures *de novo*, preparing heavy-atom derivatives is still inevitable when there are no or insufficient anomalous diffracting atoms in the protein molecule and SeMet (selenomethionyl) proteins are difficult to produce. Crystals of derivatives could have a significant anomalous scattering effect if proper heavy atoms and X-ray wavelengths were selected. This provides the basis of the SIRAS (single isomorphous replacement with anomalous scattering) method, which is one of the important techniques in solving large protein structures. However, SIRAS has its own problems: the crystallographic isomorphism between the native and the derivative may not be perfect and the derivative may often diffract to a much lower resolution than the native. In order to eliminate the effects of these problems, an alternative treatment has been proposed and tested in a case study using the SIRAS data of the originally unknown protein LegC3N. The efficiency of the proposed treatment was found to be satisfactory.

2. Data

LegC3 is an effector protein (de Felipe *et al.*, 2008) from the intracellular bacterial pathogen *Legionella pneumophila* (the causative agent of Legionnaires' disease) and the N-terminal part [LegC3N; residues 2–367 (Yao *et al.*, 2014)] was crystallized with the aid of limited proteolysis. Both native data and

methylmercuric acetate derivative (hereafter referred to as Hg-derivative) data were collected at the Canadian Light Source. Crystallographic data of the native and the derivative are summarized in Table 1. Figs. 1(*a*) and 1(*b*) show the SIR and SAD signals, respectively, as a function of data resolution. The figures were produced by the program *HKL2MAP* (Pape & Schneider, 2004) after it finished running *SHELXC* (Usón & Sheldrick, 1999; Sheldrick *et al.*, 2001).

3. Programs and methods

The program names and versions and their usage in the present test are listed in Table 2. Two direct-method procedures were involved in the test. They are briefly described in the following.

3.1. Iterative direct-method SAD phasing

Iterative direct-method SAD phasing has the following features.

(i) The 0–2π phase problem is reduced to a sign problem by the expression

$$\varphi_{\mathbf{h}} + \varphi''_{\mathbf{h}} \pm |\Delta\varphi_{\mathbf{h}}|, \qquad (1)$$

where $\varphi_{\mathbf{h}}$ is the phase of the reflection with its reciprocal vector equal to $\mathbf{h}$, $\varphi''_{\mathbf{h}}$ is the phase contributed by the anomalous scattering part of the heavy atoms and $\Delta\varphi_{\mathbf{h}} = \varphi_{\mathbf{h}} - \varphi''_{\mathbf{h}}$; its absolute value can be derived from the SAD experiment. Now, instead of finding the phase $\varphi_{\mathbf{h}}$, we need only to make a choice between plus and minus for the sign of $\Delta\varphi_{\mathbf{h}}$.

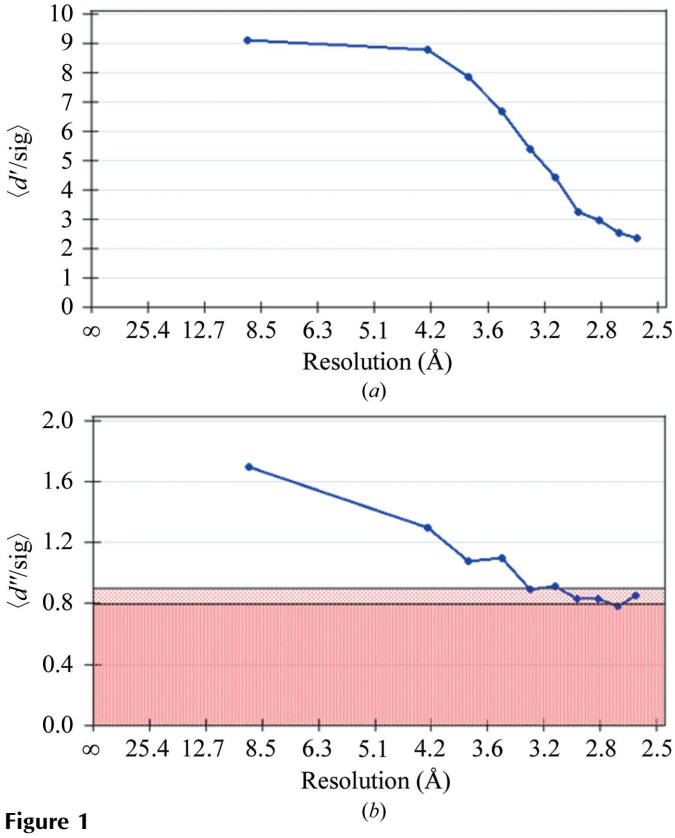

Figure 1
(*a*) SIR signals as a function of data resolution calculated by *SHELXC* as implemented in *HKL2MAP*. (*b*) SAD signals as a function of data resolution calculated by *SHELXC* as implemented in *HKL2MAP*.

Table 1
Summarized crystallographic data of LegC3N.

	Native	Hg-derivative
Space group	$P2_12_12$	$P2_12_12$
Unit-cell parameters		
a (Å)	108.874	107.039
b (Å)	150.246	149.507
c (Å)	24.240	24.207
X-ray wavelength (Å)	0.97887	0.99186
Resolution range (Å)	45.50–2.08 (2.18–2.08)	45.50–2.60 (2.64–2.60)
No. of observed reflections	266588	86709
No. of unique reflections	23598	12829
Completeness (%)	94.9 (84.9)	99.7 (96.3)
Multiplicity	11.3 (10.3)	6.8 (6.0)
$R_{\rm merge}$	0.144 (0.932)	0.203 (0.905)
$\langle I/\sigma(I)\rangle$	16.6 (2.3)	9.5 (1.7)

Table 2
Programs used in the test.

Program	Version	Usage
AutoBuild	*PHENIX* 1.8.2 or 1.8.4†	Model building/refinement
AutoSol	*PHENIX* 1.8.2 or 1.8.4†	SAD phasing and model building/refinement
DM	Version 6.3 in *CCP*4‡ 6.3.0	Density modification
OASIS	Version 4.2 in *IPCAS* 1.1	Direct-method SAD phasing and direct-method-aided model completion
SIGMAA	Version 6.3 in *CCP*4 6.3.0	Estimating weights from the partial model
Buccaneer	Version 1.5.2 in *CCP*4 6.3.0	Model building
REFMAC	Version 5.7.0032 in *CCP4* 6.3.0	Structure refinement
SHELXC/*D*	Version 2006 in *CCP*4 6.3.0 and *HKL2MAP*	Locating heavy atoms

† *AutoSol* and *AutoBuild* used in the calculations for §4 are from *PHENIX* 1.8.2, while those for §5 are from *PHENIX* 1.8.4. ‡ Winn *et al.* (2011).

(ii) The sign of $\Delta\varphi_{\mathbf{h}}$ is determined based on the product of the experimental bimodal SAD phase distribution, the Cochran distribution (Cochran, 1955) in direct methods, and the phase distribution of the heavy atoms and/or the known part of the protein structure.

(iii) Each cycle of the iterative phasing process consists of three parts: direct-method phasing, density modification and model building/refinement. From the second cycle onwards, the model built from the previous cycle will feed back to the direct-method phasing.

For further details, the reader is referred to Wang *et al.* (2004) and Yao *et al.* (2006).

3.2. Direct-method-aided model completion

This procedure was originally proposed for MR (molecular replacement) model completion (He *et al.*, 2007). However, it is in fact a general-purpose model-completion tool. While the procedure does not make use of either SAD or SIR information, the algorithm is similar to that described in the previous section with the following modifications.

(i) In a model-completion process SAD/SIR information may not necessarily be available or we may not want to make use of this information to avoid large experimental errors (see,

for example, Zhang *et al.*, 2010). In order to keep using (1) so that the 0–2π phase problem can be reduced to a sign problem, $\varphi''_{\mathbf{h}}$ is now redefined as the phase calculated from a randomly selected 5% (the value is adjustable) of the atoms from the current model.

(ii) In each iterating cycle, a number ($\geq$1) of trials with different randomly selected atoms from the current model are run in parallel. The result from the trial that leads to the smallest R factor will be passed on to the next cycle. Increasing the number of trials in each cycle would lead to better results at the cost of more complicated calculations.

By the above redefinition of $\varphi''_{\mathbf{h}}$ in (1), the direct-method phasing is actually a kind of phase-flipping process, *i.e.* for reflections having an absolute contribution from the current model smaller than that from the Cochran distribution with their signs opposite to each other a large phase change (in practice the average is $\sim$50°) will be obtained, while in other cases the phase change will be small (in practice the average is <10°). This feature is good for eliminating model bias during phase refinement/extension. For further details, the reader is referred to He *et al.* (2007).

All of the direct methods for protein crystallography developed in the Institute of Physics in Beijing up to the year 2013 have been integrated into the *IPCAS* (*Iterative Protein Crystal-structure Automatic Solution*) pipeline written by Tao Zhang. Typical applications to difficult SAD phasing, model completion and low-resolution phase/model extension are given in Fan *et al.* (2014).

Table 3
Resultant structure models from *AutoSol* in *PHENIX* running with SIRAS data and different heavy-atom substructures.

	SIRAS models from *AutoSol* in *PHENIX*					
	No. of residues					
Heavy-atom substructure	Built	Placed	$\Delta C^{\alpha} < 1$ Å†	R	$R_{\rm free}$	Model–map CC
Auto‡	134	0	0	0.53	0.56	0.31
SHELXC/*D* with 2.9 Å SIRAS data	145	11	1	0.49	0.53	0.47
SHELXC/*D* with 4.0 Å SIRAS data	119	0	2	0.50	0.52	0.40
SHELXC/*D* with 2.9 Å SIR data	81	52	3	0.52	0.53	0.35
SHELXC/*D* with 4.0 Å SAD data	163	108	12	0.46	0.51	0.49

† ΔC^{α} is the positional deviation of C^{α} atoms in the built model from those of the final structure. ‡ The substructure was created and refined within *AutoSol* automatically.

Table 4
Completion of the best SIRAS model from *AutoSol* in *PHENIX*.

	No. of residues				
	Built	Placed (sequenced)	$\Delta C^{\alpha} < 1$ Å†	R	$R_{\rm free}$
Starting model (last row of Table 3)	163	108	12	0.46	0.51
Completion by *AutoBuild* in *PHENIX*	260	260	256	0.25	0.28
Completion by *IPCAS*: *OASIS–DM–Buccaneer*/*REFMAC*	276	269	271	0.26	0.31
Final model	274	274	274	0.218	0.278

† ΔC^{α} is the positional deviation of C^{α} atoms in the built model from those of the final structure.

4. SIRAS data treated by *AutoSol* in *PHENIX*

It is convenient to treat SIRAS data with *AutoSol* in *PHENIX* (Adams *et al.*, 2010). The determination of the heavy-atom substructure is important for obtaining a good result. Table 3 lists results from *AutoSol* based on five different substructures. The first substructure was created and refined automatically within *AutoSol*, while the other four were created by *SHELXC*/*D* (Usón & Sheldrick, 1999; Sheldrick *et al.*, 2001) *via HKL2MAP* (Pape & Schneider, 2004) and refined within *AutoSol*. In dealing with SIRAS data using *SHELXC*/*D*, there are three different ways to derive the heavy-atom substructure. In theory, the best way is to use the whole set of SIRAS data to derive amplitudes $|F_{\rm A}|$ of the heavy-atom substructure. However, in practice this will include all kinds of experimental errors, *i.e.* errors from measurement of the weak signals $|\Delta F_{\rm iso}|$ and $|\Delta F_{\rm ano}|$ and errors from the imperfect isomorphism. Alternatively, instead of finding $|F_{\rm A}|$ we can use SIR data to derive $|F_{\rm A}|\cos\alpha$ or use SAD data to derive $|F_{\rm A}|\sin\alpha$, where α is the phase difference between the protein and its heavy-atom substructure. Either $|F_{\rm A}|\cos\alpha$ or $|F_{\rm A}|\sin\alpha$ can be used as an approximation to $|F_{\rm A}|$. In Table 3, the last four substructures cover all reasonable ways of using *SHELXC*/*D* to derive amplitudes of the substructure. The high-resolution cutoff of 2.9 Å for the LegC3N SIR data was suggested by *SHELXC*, while that of 4.0 Å for the LegC3N SAD data was suggested by both *SHELXC* and *phenix.xtriage*. As can be seen in Table 3, the best SIRAS model based on the heavy-atom substructure found by *SHELXC*/*D* with 4.0 Å resolution SAD data is by far the best model. This indicates that the SAD signals in LegC3N SIRAS data are much more reliable than the SIR signals, implying a considerable deviation from perfect isomorphism. Starting from the SIRAS model in the last row of Table 3, either *AutoBuild* in *PHENIX* or the direct-method-aided model completion involving *OASIS*, *DM* (Cowtan, 1994) and *Buccaneer*/*REFMAC* (Cowtan, 2006; Murshudov *et al.*, 2011) led to a nearly complete structure (see Table 4).

5. An alternative treatment for SIRAS data: SAD phasing of the derivative followed by phase/model extension to the native and final model completion with the native

As seen in the previous section, the success of SIRAS solution of LegC3N was actually based on the heavy-atom substructure derived from a set of 4.0 Å resolution SAD data. This implies the possibility of an alternative treatment of the LegC3N SIRAS data. Let us first clarify what a conventional treatment of SIRAS data is. A set of SIRAS data is in fact a set of SIR

data with sufficiently useful anomalous scattering signals in the derivative data or, in other words, a set of coexistent SIR data and SAD data. The conventional treatment would be combining the SIR and SAD information to derive the heavy-atom contribution $|F_A|$ and then to solve the heavy-atom substructure. Based on this, the protein phases can then be uniquely derived. In the previous section, it can be seen that this may not be the best way to perform this in practice owing to imperfect isomorphism. An alternative treatment may involve a three-stage process. In the first stage, only the derivative is used to obtain low-resolution phases and an initial structure model *via* SAD phasing; in the second stage, the results from the first stage are used with the native data to perform phase/model extension; finally, in the third stage a simple model completion is performed using the native data and the results from the second stage. The advantage of such a process is that it can avoid the measurement of SIR signals, *i.e.* $|\Delta F_{iso}| = |F_{derivative} - F_{native}|$. The disadvantage is that we should overcome the difficulty of SAD phase ambiguity. Hence, the success of the alternative treatment relies on the results of SAD phasing with only the derivative data.

Table 5
Structure models from SAD phasing of 5.0 Å resolution derivative data by different methods.

Method	No. of residues: Built	Placed (sequenced)	ΔC^α < 1 Å†	R	R_{free}	Model–map CC
IPCAS step 1: *OASIS* (SAD) plus *AutoBuild* (quick, helices and strands only)	198	15	12	0.35	0.46	0.65
IPCAS step 2: *OASIS* (SAD based on step 1) plus *AutoBuild* (quick, helices and strands only)	202	16	39	—	—	0.64
AutoSol (SAD) (thorough, helices and strands only)	138	72	1	0.38	0.48	0.65
AutoSol (SAD, with *SHELXD–SOLVE* heavy-atom sites‡) (thorough, helices and strands only)	149	149	0	0.35	0.46	0.76

† ΔC^α is the positional deviation of C^α atoms in the built model from those of the final structure. ‡ The same heavy-atom sites were used in the *IPCAS* treatment.

Table 6
Phase/model extension from the result of 5.0 Å resolution SAD phasing (*IPCAS* step 2) to the 2.1 Å resolution native data.

Method	No. of residues: Built	Placed (sequenced)	ΔC^α < 1 Å†	R	R_{free}
IPCAS: *OASIS* (direct-method phase/model extension) plus *AutoBuild* (quick, helices and strands only)	195	151	109	0.37	0.42
AutoBuild (phase/model extension) (quick, helices and strands only)	193	12	11	0.46	0.56
AutoBuild (phase/model extension) (thorough, helices and strands only)	207	0	16	0.48	0.55
AutoBuild (phase/model extension) (thorough)	122	12	0	0.51	0.56

† ΔC^α is the positional deviation of C^α atoms in the built model from those of the final structure.

Table 7
Model completion based on the result of *IPCAS* (see Table 6) by different methods.

Method	No. of residues: Built	Placed (sequenced)	ΔC^α < 1 Å†	R	R_{free}	CPU time‡ (h:min:s)
IPCAS (MR iteration): *DM* + *Buccaneer*	277	271	271	0.25	0.31	2:58:00 for ten cycles
IPCAS (MR iteration): *OASIS* + *DM* + *Buccaneer*	277	267	268	0.25	0.30	8:26:00 for ten cycles, each includes five trials
AutoBuild (model completion) (thorough)	267	253	262	0.24	0.31	4:39:00

† ΔC^α is the positional deviation of C^α atoms in the built model from those of the final structure. ‡ Running on a MacBookPro4.1 under OS X 10.9.3.

5.1. SAD phasing of the derivative data at 5 Å resolution

As mentioned previously, the Hg-derivative crystal diffracted to 2.6 Å resolution. However, *phenix.xtriage* reported that 'the anomalous signal seems to extend to about 5.7 Å (or to 3.9 Å from a more optimistic point of view)'. In the following test, the 5.0 Å resolution truncated Hg-derivative data were used. Recently, it has been reported (Fan *et al.*, 2014) that SAD phasing by *OASIS* as implemented in *IPCAS* resulted in a reasonable 5.0 Å resolution electron-density map. However, the partial model from the map failed to extend the 2.1 Å resolution native data. In the present work, improvement has been made to SAD phasing in *OASIS*. By default, the original unresolved SAD phase distribution will be created in *OASIS* and passed on to density modification and model building/refinement until the number of sequenced residues is greater than 30% of the whole structure (see Wu *et al.*, 2009). However, during the 30 cycles of SAD iteration of the 5.0 Å resolution LegC3N Hg-derivative data, the number of sequenced residues is always far below 30% of those in the whole structure. Consequently, the original unresolved SAD phase distribution was kept in density modification and model building/refinement throughout the 30 cycles of iteration. This could affect the rate and quality of convergence. Hence, here we used a two-step procedure for the SAD phasing. In step 1, a 30-cycle default run was performed. In step 2, the best result from step 1, cycle 16 in the present case, was then used as the starting model to continue the SAD iteration, while the use of the original unresolved SAD phase distribution was disabled. After iterating for six cycles, an improved result was obtained. Details are listed in Table 5 in comparison with the results

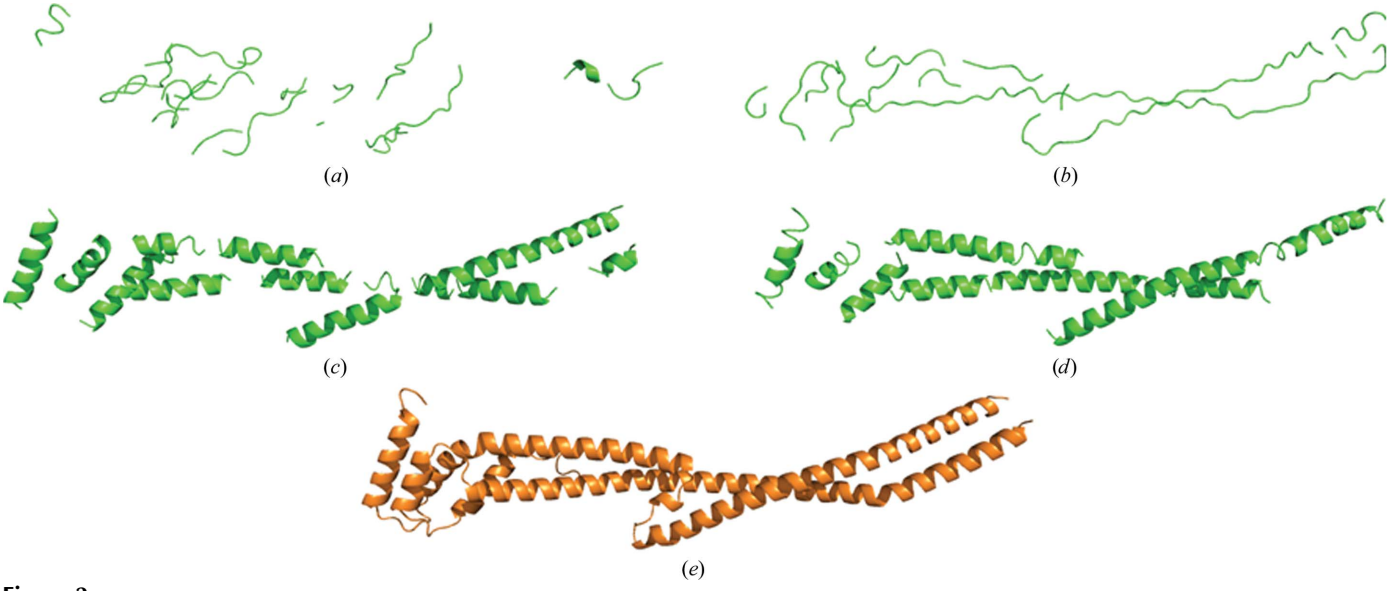

Figure 2
SAD phasing of the 5.0 Å resolution LegC3N Hg-derivative data by different methods: (*a*) *AutoSol* default SAD phasing; (*b*) *AutoSol* SAD phasing with *SHELXD–SOLVE* heavy-atom sites; (*c*) *OASIS* SAD phasing step 1; (*d*) *OASIS* SAD phasing step 2; (*e*) final structure.

from *AutoSol* SAD phasing. Ribbon models plotted by *PyMOL* (DeLano, 2002) are shown in Fig. 2.

5.2. Phase/model extension from the 5.0 Å resolution SAD phasing result to the 2.1 Å resolution native data

The direct-method-aided phase/model extension described by Fan *et al.* (2014) is in fact a special application of MR iteration (He *et al.*, 2007). Fig. 3 shows the flowchart, which is a duplicate of Fig. 6 in Fan *et al.* (2014). In the present work, an improved procedure is used; the flowchart is shown in Fig. 4. Comparing Figs. 3 and 4, it is clear that the main difference between the two procedures is where to input the starting data. For the former procedure the starting data are first input to *AutoBuild* in *PHENIX*, while for the improved procedure the starting data are input to *OASIS*. As can be seen in Table 6, while *AutoBuild* working in different modes failed to extend the SAD phasing result of *IPCAS* step 2, the combination of *OASIS* and *AutoBuild* (working in quick, helices and strands only mode) led to a much better result based on the same starting data.

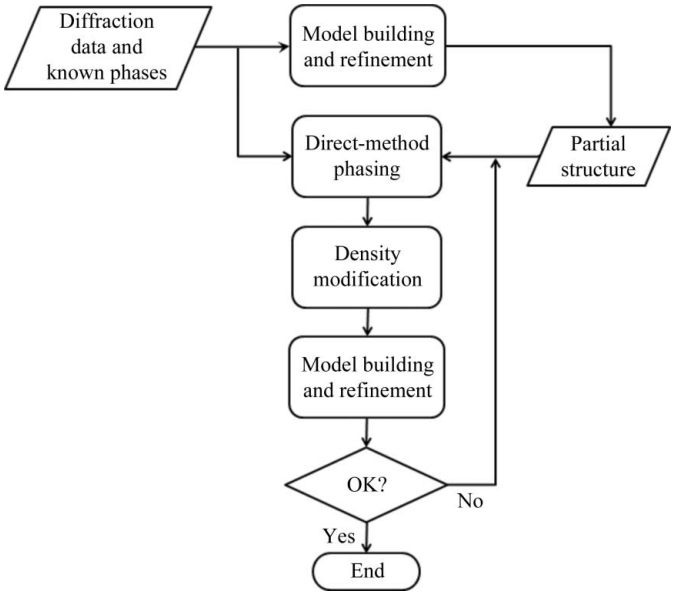

Figure 3
Flowchart of the original direct-method-aided phase/model extension as described by Fan *et al.* (2014).

5.3. Model completion

Since the result from the improved direct-method-aided phase/model extension is so good, further model completion in this case becomes trivial. Table 7 lists results from three

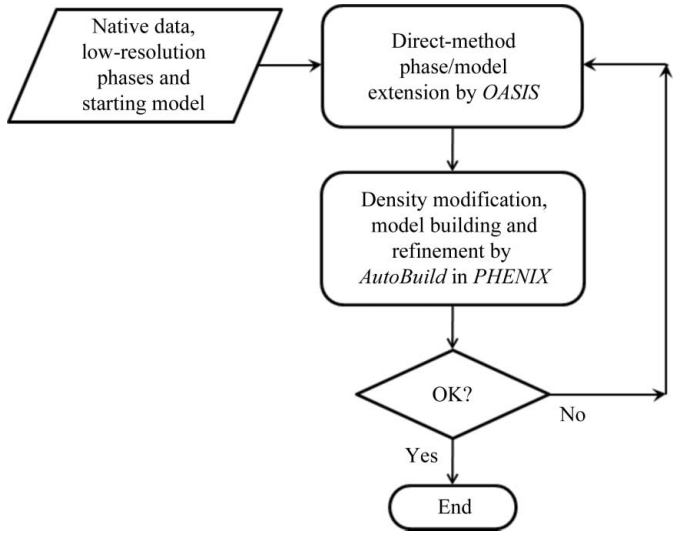

Figure 4
Flowchart of the improved direct-method-aided phase/model extension.

different methods. All led to a nearly complete structure model.

6. Concluding remarks

The iterative direct-method SAD phasing implemented by *OASIS* plus *AutoBuild* is capable of phasing low-resolution SAD data down to 5.0 Å resolution.

The improved direct-method-aided model completion implemented by *OASIS–AutoBuild* is capable of extending a 5.0 Å resolution secondary-structure model to a nearly complete structure using the 2.1 Å resolution native data. This technique is useful in dealing with SIRAS data with derivative crystals that diffract to low resolution. It also enables the combination of low-resolution phases/models from other sources, *e.g.* cryo-electron microscopy, with X-ray data.

This work was supported by the 973 Project (Grant No. 2011CB911101) of the Ministry of Science and Technology of China. We thank the Canadian Macromolecular Crystallography Facility (CMCF) at the Canadian Light Source (CLS) for access to beam time. M. Cygler would like to acknowledge the support from the CIHR grant GSP-48370. The authors would like to thank the referees, whose comments/suggestions substantially improved the manuscript.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Cochran, W. (1955). *Acta Cryst.* **8**, 473–478.
Cowtan, K. (1994). *Jnt CCP4 ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.
Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.
DeLano, W. L. (2002). *PyMOL*. http://www.pymol.org.
Fan, H., Gu, Y., He, Y., Lin, Z., Wang, J., Yao, D. & Zhang, T. (2014). *Acta Cryst.* A**70**, 239–247.
Felipe, K. S. de, Glover, R. T., Charpentier, X., Anderson, O. R., Reyes, M., Pericone, C. D. & Shuman, H. A. (2008). *PLoS Pathog.* **4**, e1000117.
He, Y., Yao, D.-Q., Gu, Y.-X., Lin, Z.-J., Zheng, C.-D. & Fan, H.-F. (2007). *Acta Cryst.* D**63**, 793–799.
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.* D**67**, 355–367.
Pape, T. & Schneider, T. R. (2004). *J. Appl. Cryst.* **37**, 843–844.
Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, M. & Usón, I. (2001). *International Tables for Macromolecular Crystallography*, Vol. *F*, edited by E. Arnold & M. Rossmann, pp. 333–345. Dordrecht: Kluwer Academic Publishers.
Usón, I. & Sheldrick, G. M. (1999). *Curr. Opin. Struct. Biol.* **9**, 643–648.
Wang, J. W., Chen, J. R., Gu, Y. X., Zheng, C. D. & Fan, H. F. (2004). *Acta Cryst.* D**60**, 1991–1996.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Wu, L.-J., Zhang, T., Gu, Y.-X., Zheng, C.-D. & Fan, H.-F. (2009). *Acta Cryst.* D**65**, 1213–1216.
Yao, D., Cherney, M. & Cygler, M. (2014). *Acta Cryst.* D**70**, 436–441.
Yao, D., Huang, S., Wang, J., Gu, Y., Zheng, C., Fan, H., Watanabe, N. & Tanaka, I. (2006). *Acta Cryst.* D**62**, 883–890.
Zhang, T., Wu, L.-J., Gu, Y.-X., Zheng, C.-D. & Fan, H.-F. (2010). *Chin. Phys. B* **19**, 096101.