Acta Crystallographica Section A
Foundations and Advances
ISSN 2053-2733

Applications of direct methods in protein crystallography for dealing with diffraction data down to 5 Å resolution

Haifu Fan, Yuanxin Gu, Yao He, Zhengjiong Lin, Jiawei Wang, Deqiang Yao and Tao Zhang

Acta Cryst. (2014). A**70**, 239–247

Acta Crystallographica Section A: Foundations and Advances publishes articles reporting fundamental advances in all areas of crystallography in the broadest sense. This includes metacrystals such as photonic or phononic crystals, *i.e.* structures on the meso- or macroscale that can be studied with crystallographic methods. The central themes are, on the one hand, experimental and theoretical studies of the properties and arrangements of atoms, ions and molecules in condensed matter, periodic, quasiperiodic or amorphous, ideal or real, and, on the other, the theoretical and experimental aspects of the various methods to determine these properties and arrangements. In the case of metacrystals, the focus is on the methods for their creation and on the structure–property relationships for their interaction with classical waves.

Acta Crystallographica Section A
Foundations and Advances
ISSN 2053-2733

Received 11 September 2013
Accepted 30 December 2013

Applications of direct methods in protein crystallography for dealing with diffraction data down to 5 Å resolution

Haifu Fan,[a]* Yuanxin Gu,[a]* Yao He,[a] Zhengjiong Lin,[b] Jiawei Wang,[c] Deqiang Yao[d] and Tao Zhang[a]

[a]Institute of Physics, Chinese Academy of Sciences, Beijing, 100190, People's Republic of China, [b]Institute of Biophysics, Chinese Academy of Sciences, Beijing, 100101, People's Republic of China, [c]School of Life Sciences, Tsinghua University, Beijing, 100084, People's Republic of China, and [d]Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, People's Republic of China. Correspondence e-mail: fanhf@cryst.iphy.ac.cn, gu@cryst.iphy.ac.cn

Apart from solving the heavy-atom substructure in proteins and *ab initio* phasing of protein diffraction data at atomic resolution, direct methods have also been successfully combined with other protein crystallographic methods in dealing with diffraction data far below atomic resolution, leading to significantly improved results. In this respect, direct methods provide phase constraints in reciprocal space within a dual-space iterative framework rather than solve the phase problem independently. Applications of this type of direct methods to difficult SAD phasing, model completion and low-resolution phase extension will be described in detail.

1. Introduction

Applications of direct methods in protein crystallography can be divided into three categories: (i) solving the heavy-atom substructure in proteins; (ii) *ab initio* solution of protein structures (atomic resolution data at ∼1.2 Å or better is required); and (iii) combining direct methods with existing protein crystallographic methods.

Direct-methods solution of the heavy-atom substructure in proteins was initiated by Steitz (1968) using centrosymmetric projections and by Neidle (1973) with acentric three-dimensional data. An early attempt at *ab initio* direct-methods solution of protein structures was made by Woolfson & Yao (1990). They found that it was possible to solve a small protein using the direct-methods program *SAYTAN*. Further developments include improved algorithms encoded in the programs *Shake-and-Bake* (Miller *et al.*, 1993; Weeks *et al.*, 1993; Xu *et al.*, 2000), *SHELXD* (Sheldrick & Gould, 1995; Sheldrick, 1997), *ACORN* (Foadi *et al.*, 2000; Yao, Dodson *et al.*, 2006) and the *SIR* suite (Burla *et al.*, 2003, 2005, 2012). All these programs require atomic resolution (∼1.2 Å or better) data. Only about 5% of the entries in the Protein Data Bank can fulfil the atomic resolution requirement. On the other hand, the above programs are widely used in solving heavy-atom substructures of proteins. For this purpose, since distances between heavy atoms in protein crystals are mostly longer than 3–4 Å, data at resolution down to ∼4 Å or even lower are still 'atomic resolution' data for solving heavy-atom substructures.

The earliest proposals on combining direct methods with existing protein crystallographic methods were that of breaking the SAD (single-wavelength anomalous diffraction) or SIR (single isomorphous replacement) phase ambiguity (Coulter, 1965; Fan, 1965*a*,*b*; Karle, 1966). Similar studies were reported later by Hazell (1970), Hendrickson (1971), Sikka (1973) and Heinerman *et al.* (1978). From the early 1980s to the early 2000s, the combination of direct methods with SAD/SIR data had been a hot topic in direct-methods research worldwide (Hauptman, 1982, 1996; Hauptman *et al.*, 1982; Fortier *et al.*, 1985; Xu & Hauptman, 2003; Giacovazzo, 1983; Giacovazzo *et al.*, 1988, 1995; Giacovazzo & Siliqi, 2004; Fan *et al.*, 1984, 1990; Fan & Gu, 1985; Bing-Dong *et al.*, 1995; Zheng *et al.*, 1997; Liu *et al.*, 1999; Klop *et al.*, 1987; Verwer *et al.*, 1991; Kyriakidis *et al.*, 1993; Woolfson *et al.*, 1997). The first application of direct-methods SAD phasing in solving originally unknown protein structures was reported by Yu-dong *et al.* (1999) with the data of rusticyanin at 2.1 Å resolution. Based on the method used in solving rusticyanin, the program *OASIS* (One-wavelength Anomalous Scattering and Single Isomorphous Substitution) was released (Hao *et al.*, 2000). This is the first direct-methods program to be incorporated in *CCP4* (Collaborative Computational Project, Number 4, 1994) for SAD/SIR phasing of protein diffraction data below atomic resolution.

Two major developments in the application of direct methods have been achieved since the year 2000. The first is the iterative direct-methods SAD/SIR phasing and structure model extension (Wang, Chen, Gu, Zheng & Fan, 2004); the

second is the direct-methods-aided model completion without using additional experimental phase information such as SAD, SIR *etc*. (He *et al.*, 2007). Both procedures are dual-space iterative phasing and model-building procedures. Recently, the *IPCAS* (iterative protein crystal structure automatic solution) pipeline has been released (Zhang *et al.*, 2012*b*), which was used for most of the calculations in the present paper. In the following, direct methods belonging to the third category and their applications to difficult SAD phasing, model completion and low-resolution phase extension will be discussed in detail.

2. Dual-space iterative algorithms and diffraction analysis

The direct methods to be discussed in the next part of this paper work within a dual-space iterative framework. They provide phase constraints in reciprocal space rather than solve the phase problem independently. The dual-space phase-retrieval algorithm in crystallography was proposed by Gerchberg & Saxton (1971, 1972) for phase retrieval in electron microscopy. In protein crystallography, a dual-space phasing and density-modifying algorithm was proposed by Wang (1985) for resolving the phase ambiguity in the SAD or SIR methods and for phase improvement *via* the solvent-flattening technique. The programs *Shake-and-Bake*, *SHELXD* and *ACORN* mentioned above are all dual-space iterative procedures. Modern protein structure solution pipelines like *PHENIX* (Adams *et al.*, 2010) and *SHELXC/D/E* (Sheldrick, 2010) contain various kinds of dual-space iterative algorithms. The oversampling phasing algorithm dealing with phase retrieval in single-molecule coherent diffraction imaging of proteins (Miao *et al.*, 2001) is also a typical dual-space iterative phasing procedure.

2.1. The dual-space iterative phasing/model-building framework of solving crystal structures

Fig. 1 shows schematically the dual-space iterative framework of crystal structure analysis. Any particular process of solving crystal structures can be fitted in it. This can be explained by taking the trial-and-error method as an example. A trial-and-error procedure builds a structure model at the beginning by guessing. This implies constraints in the real space to modifying the electron-density function $\rho(r)$. The Fourier transform of the structure model or the modified electron density $\rho'(r)$ is then passed onto the reciprocal space. Here the basic constraint of 'replacing $|F_c|$ with $|F_o|$' is applied. Since the trial-and-error process makes no additional constraints on phases, the inverse Fourier transform of $|F_o| \exp(i\varphi)$ is then used to produce a new electron-density function $\rho(r)$ and then go back to the real space. By comparing $\rho(r)$ and $\rho'(r)$, a new model or modified electron-density function $\rho'(r)$ can be created and a new cycle of iteration starts.

2.2. The efficiency of a particular structure-solving process depends on what constraints it sets in the real and/or the reciprocal space

Trial-and-error methods are capable of solving crystal structures containing about ten symmetry-independent atoms. The Patterson method (Patterson, 1934) sets additional mathematical constraints to the atomic arrangement in real space. Hence a partial or even complete structure model can be derived without relying on guessing. The iteration performed in Patterson procedures is similar to that of trial-and-error methods. Up to the 1980s Patterson methods were capable of solving crystal structures containing about 100 independent atoms if there were some heavy atoms in the structure, otherwise Patterson methods were capable of solving structures containing about 20 light independent atoms. In contrast to Patterson methods, direct methods set mathematical constraints to phases in reciprocal space. Up to the 1980s direct methods were capable of solving structures containing about 100 independent atoms no matter whether the structure contained heavy atoms or not. In the context of a dual-space iterative framework, it is easy to understand that Patterson methods and direct methods can be synergic in crystal structure analysis. This can be seen clearly from numerous successful applications of the program *SHELXD* (Sheldrick & Gould, 1995; Sheldrick, 1997).

In protein crystallography, for solving structures containing thousands of independent atoms with diffraction data far below atomic resolution, additional experimental data are needed so as to set more constraints to phases and/or the structure model. One kind of additional experimental data can be the known structure of a homologous protein. This kind of data can be used to constrain the structure model of the unknown target protein structure. The MR (molecular replacement) procedure (Rossmann & Blow, 1962; Rossmann &

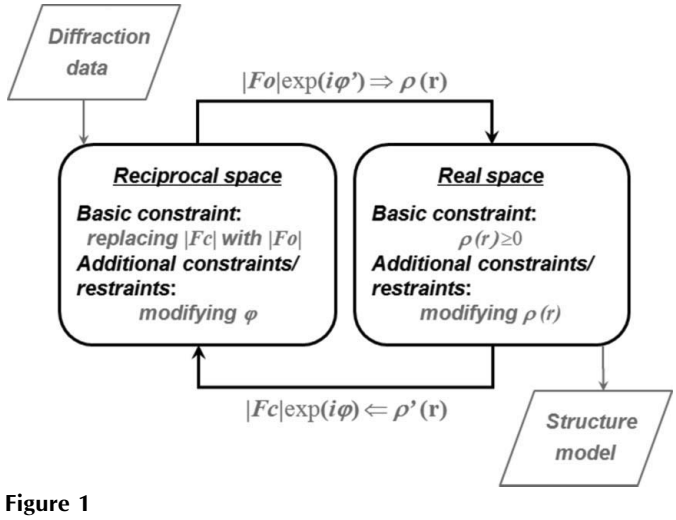

Figure 1
Schematic representation of the dual-space iterative framework of crystal structure analysis. $|F_o|$, experimental observed structure-factor amplitudes; $|F_c|$, calculated structure-factor amplitudes; φ, structure-factor phases calculated from the real-space modified electron-density function $\rho'(r)$ or from the structure model; φ', reciprocal-space modified structure-factor phases.

Table 1
Summary of samples for testing iterative direct-methods SAD phasing and model building.

AU: asymmetric unit.

Protein	Anomalous scatterers in the AU	X-ray λ (Å)	Bijvoet ratio $\langle\lvert\Delta F\rvert\rangle/\langle F\rangle$ (%)	Data redundancy	High-resolution limit (Å)	No. of residues in the AU	Reference
Xylanase	S (6)†	1.49	0.56	15.9	1.80	303	Ramagopal *et al.* (2003)
TT0570	S (20)†	1.542 (Cu $K\alpha$)	0.55	29.2	1.86	1206	Watanabe *et al.* (2005)
TTHA	S (2)	2.291 (Cr $K\alpha$)	1.14	13.5	2.27	215	Private communication
Tom70p	Se (22)	0.9798	4.3	3.3	3.3	1086	Wu & Sha (2006)
LegC3N	Hg (3)	0.9919	8.1	6.8	5.0	367	Yao, Huang *et al.* (2006)

† All anomalous scatterers found are treated as sulfur atoms.

Arnold, 2001) in protein crystallography can be understood as the principles of the Patterson method combined with data of the homologous pair. Like the Patterson procedure in small molecular crystallography, the MR procedure sets strong constraints to the structure model in real space, but has nothing to do directly with phases in reciprocal space. Hence there is room for improvement, in particular by adding constraints to phases in reciprocal space. When the known structure of a homologous protein is unavailable, other kinds of additional experimental data should be used. MIR (multiple isomorphous replacement) and MAD (multi-wavelength anomalous diffraction) are typical procedures for this purpose. They set strong constraints to phases in reciprocal space. However, sometimes it is not easy to prepare MIR samples and there are difficulties in collecting MAD data. In these cases, SIR or SAD procedures may be used instead. Unlike MIR and MAD procedures, either SIR or SAD can only provide weak constraints to phases in reciprocal space owing to the phase ambiguity intrinsic in these techniques. Direct methods can help in resolving the phase ambiguity leading to the iterative direct-methods SAD/SIR phasing. Direct methods are actually universal techniques for restraining phases in reciprocal space. Hence direct methods can be combined with most kinds of existing structure-solving procedures. This will be described in the following section.

3. Combining direct methods with existing methods in protein crystallography

It is well known in the crystallographic community that direct methods are *ab initio* phasing methods. However, if we look at things in a different way, we may see something different. According to the two basic formulae of direct methods, the Sayre equation (1) (Sayre, 1952) and the tangent formula (2) (Karle & Hauptman, 1956), it is clear that, if no initial phases are put into their right-hand side, we can get nothing about the phase of reflections **h** on the left. Hence direct methods are rather a kind of phase-extension/refinement procedure than an *ab initio* phasing one. In this context, direct methods may work better in combination with other methods than on their own.

$$F_{\mathbf{h}} = \frac{\theta}{V}\sum_{\mathbf{h}'} F_{\mathbf{h}'}F_{\mathbf{h}-\mathbf{h}'}, \tag{1}$$

$$\tan\alpha_{\mathbf{h}} = \frac{\sum_{\mathbf{h}'}|E_{\mathbf{h}'}E_{\mathbf{h}-\mathbf{h}'}|\sin(\alpha_{\mathbf{h}'}+\alpha_{\mathbf{h}-\mathbf{h}'})}{\sum_{\mathbf{h}'}|E_{\mathbf{h}'}E_{\mathbf{h}-\mathbf{h}'}|\cos(\alpha_{\mathbf{h}'}+\alpha_{\mathbf{h}-\mathbf{h}'})}. \tag{2}$$

3.1. The P_+ formula

The P_+ formula was proposed (Fan & Gu, 1985) originally to combine direct methods with SAD/SIR data for breaking the enantiomorphic phase ambiguity. The main points are as follows:

(i) The phase $\varphi_{\mathbf{h}}$ is expressed as

$$\varphi_{\mathbf{h}} = \varphi''_{\mathbf{h}} \pm |\Delta\varphi_{\mathbf{h}}|. \tag{3}$$

In SAD cases, $\varphi''_{\mathbf{h}}$ is the phase contribution of the imaginary-part scattering of anomalous scatterers, while in SIR cases $\varphi''_{\mathbf{h}}$ is that of the real-part scattering of isomorphous replacing atoms. $\Delta\varphi_{\mathbf{h}}$ is the phase difference between $\varphi_{\mathbf{h}}$ and $\varphi''_{\mathbf{h}}$. Both $\varphi''_{\mathbf{h}}$ and $|\Delta\varphi_{\mathbf{h}}|$ in equation (3) can be calculated once the heavy atoms are located. Hence, equation (3) turns the 0–2π phase problem into a sign problem of making a choice between plus and minus. As is well known, direct methods are much more reliable in solving sign problems than in solving phase problems.

(ii) The sign of $\Delta\varphi_{\mathbf{h}}$ in equation (3) is estimated by the P_+ formula, which gives the probability of $\Delta\varphi_{\mathbf{h}}$ being positive:

$$P_+(\Delta\varphi_{\mathbf{h}}) = \tfrac{1}{2} + \tfrac{1}{2}\tanh\Big\{\sin|\Delta\varphi_{\mathbf{h}}|\Big[\sum_{\mathbf{h}'} m_{\mathbf{h}'}m_{\mathbf{h}-\mathbf{h}'}\kappa_{\mathbf{h},\mathbf{h}'}\sin(\Phi'_3 + \Delta\varphi_{\mathbf{h}'\text{best}} + \Delta\varphi_{\mathbf{h}-\mathbf{h}'\text{best}}) + \chi\sin\delta_{\mathbf{h}}\Big]\Big\}. \tag{4}$$

For details of this formula, the reader is referred to the original publication (Fan & Gu, 1985). What should be emphasized here is that the P_+ formula provides a platform for combining information from various sources. Three species of information are 'merged' inside the hyperbolic tangent function 'tanh' on the right of equation (4). The first is $|\Delta\varphi_{\mathbf{h}}|$, which comes from the phase doublet or bimodal distribution of a SAD/SIR experiment; the second is $\sum_{\mathbf{h}'} m_{\mathbf{h}'}m_{\mathbf{h}-\mathbf{h}'}\kappa_{\mathbf{h},\mathbf{h}'}\sin(\Phi'_3 + \Delta\varphi_{\mathbf{h}'\text{best}} + \Delta\varphi_{\mathbf{h}-\mathbf{h}'\text{best}})$, which is the phase constraint set by the Cochran distribution in direct methods (Cochran, 1955); while the third is $\chi\sin\delta_{\mathbf{h}}$, which

Table 2
Summary of SAD-phasing and model-building results.

Key to methods used: O, *OASIS* for phasing; D, *DM* for density modification; R, *RESOLVE* for model building; A, *ARP/wARP* and *REFMAC* for model building/refinement; B, *Buccaneer* and *REFMAC* for model building/refinement; P(qhs), *PHENIX.AutoBuild* running in 'quick' and 'helices_strands_only' mode for density modification and model building/refinement.

Protein	Methods	*R*	*R*-free	No. of residues built	No. of residues sequenced	No. of iteration cycles	Further model completion†
Xylanase	ODRA	0.168	0.204	298	298	5	No need
TT0570	ODRA	0.189	0.241	1174	1151	11	No need
TTHA	ODRA	0.269	0.333	169	160	11	Yes
Tom70p	ODRB	0.319	0.412	960	777	13	Yes
LegC3N	OP(qhs)	0.34	0.45	216	0	22	No

† Here, 'model completion' means 'direct-methods-aided model completion'; 'no need' means that the resultant model is good enough and there is no need to pass through further automatic extension; 'Yes' means that the resultant model did successfully extend by 'direct-methods-aided model completion' to more than 90% of the final structure; 'No' means that the resultant model failed to extend with 'direct-methods-aided model completion'.

comes from the contribution of the heavy-atom substructure expressed as the Sim distribution (Sim, 1959). The P_+ formula is suitable for restraining phases in reciprocal space. It is also capable of accepting known phases and/or partial structures from different sources. The first successful applications of the P_+ formula in breaking the SAD and the SIR phase ambiguity from experimental data were, respectively, reported by Fan *et al.* (1990) and Liu *et al.* (1999). The latter provides a special SIR example, which possesses a centric substructure of replacing heavy atoms. Hence the SIR phase ambiguity is difficult to break by other methods.

3.2. Dual-space iterative direct-methods SAD/SIR phasing and model building

The procedure was proposed by Wang, Chen, Gu, Zheng & Fan (2004). The flowchart is shown in Fig. 2. A systematic test of the procedure was reported by Yao, Huang *et al.* (2006). Here, some difficult SAD-phasing cases are given to demonstrate the general ability of the procedure. Table 1 summarizes the samples. The structure of xylanase has been solved by other methods; the SAD data quoted here are given by Ramagopal *et al.* (2003), who stated that SAD phasing with existing programs failed. Yao, Huang *et al.* (2006) reported that such xylanase data could be successfully phased by *OASIS*-2004 (Wang, Chen, Gu, Zheng & Fan, 2004; Wang, Chen, Gu, Zheng, Jiang & Fan, 2004).

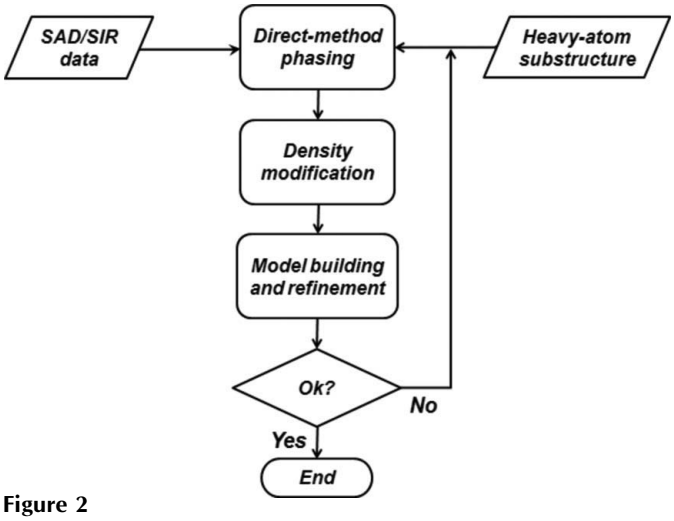

Figure 2
Flowchart of dual-space iterative direct-methods SAD/SIR phasing and model building. Programs involved: direct-methods phasing, *OASIS* (Zhang *et al.*, 2012*a*); density modification, *DM* (Cowtan, 1994) or *RESOLVE* (Terwilliger, 2000); model building/refinement, *ARP/wARP* (Langer *et al.*, 2008) and *REFMAC* (Murshudov *et al.*, 2011), *Buccaneer* (Cowtan, 2006) and *REFMAC*, or *PHENIX.AutoBuild* (Terwilliger *et al.*, 2008).

TT0570 is one of the largest proteins with a very low Bijvoet ratio so far solved with sulfur-SAD data. TTHA is a very difficult sulfur-SAD-phasing case; it consists of more than 200 residues with only two sulfur atoms. Tom70p is also a very difficult SAD-phasing case owing to rather low data resolution, low redundancy and big molecular size. LegC3N is a test case performed recently in order to probe the limit of direct methods in dealing with low-resolution SAD data. The mercury derivative of LegC3N diffracts to 2.6 Å resolution. However *phenix.xtriage* reported: 'the anomalous signal seems to extend to about 5.7 Å (or to 3.9 Å, from a more optimistic point of view)'. After failing to phase the 5.7 Å truncated data, we tried the 5.0 Å data. The results of all the samples are listed in Table 2. As can be seen, LegC3N is the only sample where the resultant model from SAD phasing failed to be extended with the direct-methods-aided model completion. However, the 5.0 Å resolution electron-density map matches reasonably the final structure and the secondary-structure model built automatically from the map covers about 60% of the total residues, revealing the essential feature of the final structure (see Fig. 3).

The success of phasing the 5.0 Å resolution SAD data of LegC3N relied on the combination of *OASIS* for phasing and *PHENIX.AutoBuild* for density modification and for model

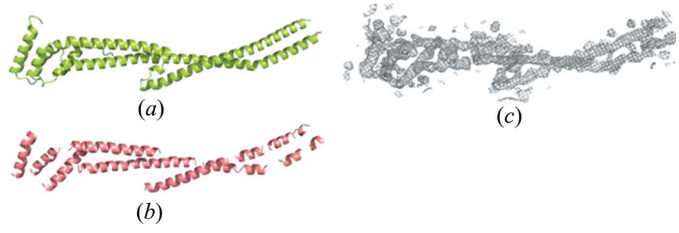

Figure 3
Comparison of LegC3N final structure with the 5.0 Å electron-density map and the structure model built from it. (*a*) Ribbon model of the final structure of LegC3N; (*b*) secondary-structure model of LegC3N from 22 cycles of iterative direct-methods SAD phasing and model building with 5.0 Å derivative data; (*c*) LegC3N 5.0 Å electron-density map at the 1σ level phased by iterative direct-methods SAD phasing and model building. All ribbon structure models in the present paper and the electron-density map in Fig. 3 were plotted by *PyMOL* (DeLano, 2002).

Table 3
$|F_o|$-weighted phase errors (°) from the initial cycle of *OASIS* SAD phasing.

Protein	All reflections	Centric reflections	Reflections with $\Delta\varphi_{\mathbf{h}} > 0$	
			Error calculated with $-\lvert\Delta\varphi_{\mathbf{h}}\rvert$	Error calculated with $+\lvert\Delta\varphi_{\mathbf{h}}\rvert$
Xylanase	67.17	74.45	81.96	72.29
TT0570	71.45	83.53	83.81	76.40
TTHA	72.00	77.63	85.91	69.93
Tom70p	71.41	74.28	89.67	73.47
LegC3N	70.32	77.41	84.97	83.58

building (in 'quick' and 'helices_and_strands_only' mode). We have tried other combinations, but they failed to give results comparable to those from *OASIS* plus *AutoBuild*. Two features of *OASIS* in SAD phasing are explained with data in Table 3, in which averaged phase errors of the initial cycle calculated for the five test samples are listed. The first feature relates to centric reflections. Most SAD-phasing procedures reject centric reflections in calculating the initial electron-density map, since phases of centric reflections are not available at that stage. However, *OASIS* can derive phases of centric reflections from those of acentric reflections just after resolving the SAD-phase ambiguity. The overall average phase errors of centric reflections thus obtained are listed in the third column of Table 3. They are all well below the averaged error (∼90°) for random phases. Hence the quality of the initial electron-density map can be improved by including these centric reflections. The second feature relates to $\Delta\varphi_{\mathbf{h}}$. According to equation (3), the phase of a given reflection **h** is either $\varphi''_{\mathbf{h}} + |\Delta\varphi_{\mathbf{h}}|$ or $\varphi''_{\mathbf{h}} - |\Delta\varphi_{\mathbf{h}}|$ depending on whether $\Delta\varphi_{\mathbf{h}}$ is positive or negative. In practice, phases resulting from techniques that break the SAD-phase ambiguity based on only the contribution of anomalous substructures will always correspond to a negative $\Delta\varphi_{\mathbf{h}}$. On the other hand, *OASIS* SAD phasing can give positive as well as negative $\Delta\varphi_{\mathbf{h}}$'s (for a detailed explanation the reader is referred to §3.2 in Wang, Chen, Gu, Zheng, Jiang, Fan, Terwilliger & Hao, 2004). The last column of Table 3 lists phase errors for reflections having a positive $\Delta\varphi_{\mathbf{h}}$ resulting from *OASIS*. For comparison, phase errors of the same portion of reflections calculated by changing the sign of $\Delta\varphi_{\mathbf{h}}$ to negative are listed in the fourth column. As can be seen, errors in the last column are obviously smaller than those in the fourth column. This means on average *OASIS* predicts properly a portion of reflections having a positive $\Delta\varphi_{\mathbf{h}}$.

3.3. Direct-methods-aided model completion

The procedure was originally proposed for MR model completion (He *et al.*, 2007). However, it has been successfully applied to partial models from a wide variety sources. The flowchart is shown in Fig. 4. Here the partial structure or the feedback model from model building/refinement set constraints to the structure model in real space, while direct methods restrain phases in reciprocal space. The procedure

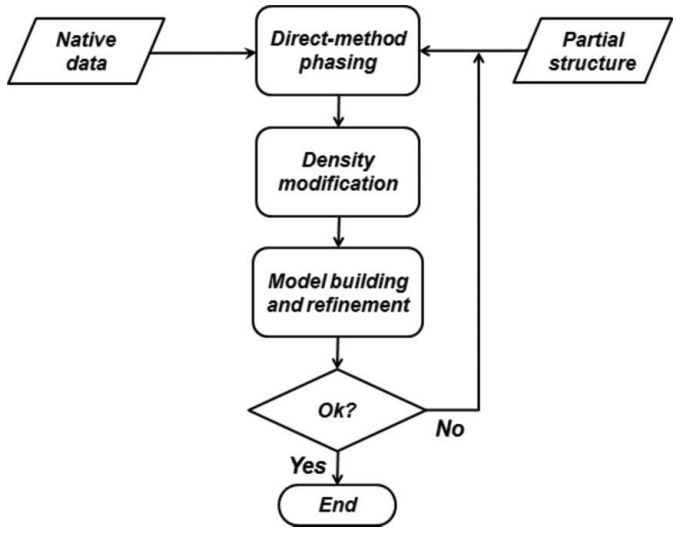

Figure 4
Flowchart of direct-methods-aided model completion. For programs involved, see Fig. 2.

does not make use of SAD/SIR information so as to extend the application field and to avoid large systematic errors from SAD/SIR signals (for the latter, see Zhang *et al.*, 2010). The flowchart and algorithm are similar to those of 'iterative direct-method SAD/SIR phasing and model building/extension'.

In order to use equation (3) so that the 0–2π phase problem can be reduced to a sign problem, now $\varphi''_{\mathbf{h}}$ will no longer be related to the heavy-atom substructure. It is redefined as the reference phase calculated from a randomly selected 5% (the value is adjustable) of the atoms from the current model. During each iterative cycle, a number (from 1 onwards) of trials run in parallel with different randomly selected atoms from the current model. The result from the trial that leads to the smallest R factor will be passed onto 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 equation (3), the direct-method phasing is actually a kind of phase flipping, *i.e.* for reflections having the absolute contribution from the current model smaller than that from the Cochran distribution and signs that are opposite to each other, a large phase change (in practice the average is ∼50°) will be obtained. However, 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 the original publications (He *et al.*, 2007).

In the following we will show a recent test with a known membrane protein UraA (PDB entry 3qe7) consisting of 429 residues in the asymmetric unit (Lu *et al.*, 2011). The structure was solved using a set of SIRAS (single isomorphous replacement with anomalous scattering) data with the mercury-derivative crystal diffracting to 3.6 Å and the native crystal diffracting to 2.8 Å resolution. The starting model to be completed was prepared by preliminary calculations not involving direct methods. During the preparation *PHENIX*.-

Table 4
Model completion of UraA (3qe7).

Model	*R*	*R*-free	No. of residues built	No. of residues sequenced
Start	0.32	0.37	327	273
Result	0.26	0.35	495	427
Final	0.25	0.30	409	409

AutoSol was used to derive an initial model from the SAD data of the Hg derivative. *PHENIX.AutoBuild* was then used to extend the model with derivative data at 3.6 Å. Finally *PHENIX.AutoBuild* was used to extend the model with 2.8 Å native data. The resultant model was then used as the starting point of direct-methods-aided model completion. The results are shown in Table 4 and Fig. 5. It is clear that the test is successful.

3.4. Dual-space collaborative phase/model extension

Phase extension is important in solving complicated protein structures. When using MIR, SIR or SIRAS techniques, it is often the case that the high-resolution cutoff of the native data is much higher than that of the derivatives. Similarly, when using the MAD technique, data from the 'high-remote' wavelength may be at much higher resolution than those from the 'peak' one. Sometimes, we have only a set of native data at moderate high resolution, but there is no SAD/SIR information and no homologous structures. Meanwhile, some low-resolution NMR data or cryo-EM (electron microscopy) data are available. In all these cases, we could probably first solve the phase problem at low resolution, leaving the difficult task of phase extension to obtain the high-resolution structure. On the other hand, as already mentioned above, the P_+ formula is a good platform for merging together information from different sources. There are no difficulties in including known phases in the 'direct-methods-aided model completion'. The flowchart will be changed from Fig. 4 to Fig. 6. The input data consist of high-resolution diffraction data and low-resolution known phases. They are first passed through a model-building and refinement process implemented by *PHENIX.AutoBuild* to produce an initial model. This can be recognized as a real-space phase extension. Then the original input data together with the initial model are input to *OASIS* for direct-method phasing. Here a reciprocal-space phase extension is performed. The low-resolution known phases will be kept fixed during the beginning cycles of iteration until the resultant model has grown to bigger than say 80% (adjustable) of the whole structure. The remaining parts in the flowchart are identical to those of 'direct-methods-aided model completion'. Programs in charge of these parts are *DM* (Cowtan, 1994) for density modification, *Buccaneer* (Cowtan, 2006) and *REFMAC* (Murshudov *et al.*, 2011) for model building and refinement. Two test calculations are provided below.

3.4.1. Dual-space collaborative phase/model extension for 1h3i. The structure of 1h3i was solved at 2.8 Å by the MAD method (Wilson *et al.*, 2002). For the present test, we take only diffraction data from the 'high remote' wavelength and a set of known MAD phases extended to 5.3 Å resolution. Phase extension was performed by 30 cycles of iteration. The results of cycles 0, 8, 20 and 30 are listed in Table 5 and Fig. 7. As can be seen, the procedure is very successful. Within 20 cycles of iteration the phase error decreased to below 30° while the model grew to more than 95% of the whole structure. On the

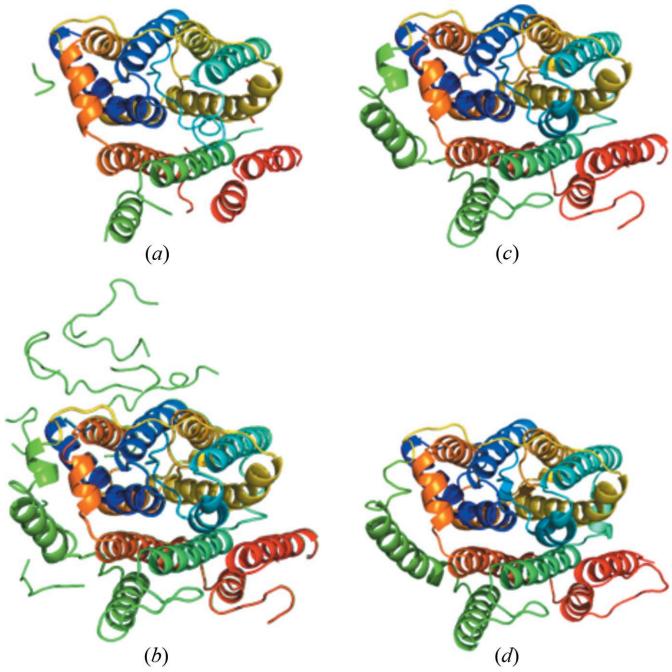

Figure 5
Model completion of UraA (3qe7). (*a*) The starting model from *PHENIX.AutoBuild* with 2.8 Å native data based on the output of *AutoSol* with 3.6 Å Hg-derivative data. (*b*) Model resulting from direct-methods-aided model completion in seven cycles of iteration based on model (*a*). (*c*) Model (*b*) after manually removing the isolated coils. (*d*) The final model from the PDB entry 3qe7.

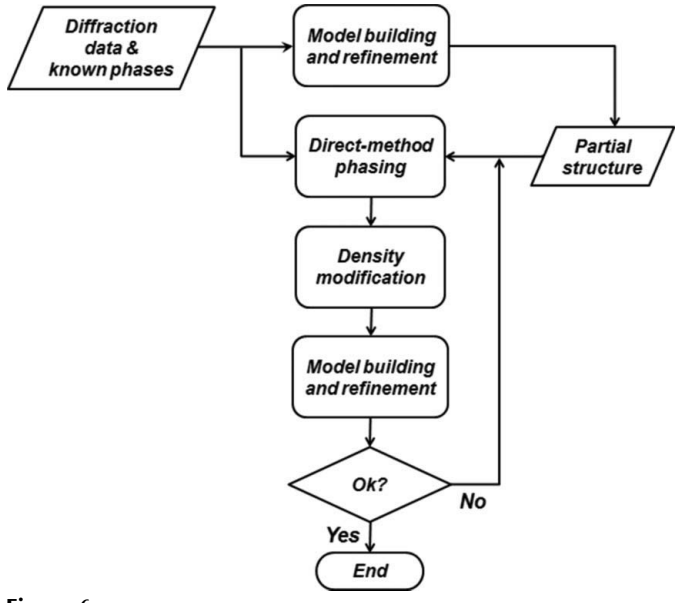

Figure 6
Flowchart of dual-space collaborative phase/model extension.

Table 5
Results of dual-space collaborative phase/model extension for 1h3i (from 5.3 Å MAD phases to 2.8 Å structure).

Cycle	$\|F_o\|$-weighted phase error (°)	*R*	*R*-free	No. of residues built	No. of residues placed (sequenced)
0†	47.2	0.39	0.48	410	22
8	46.6	0.407	0.495	527	418
20	25.0	0.259	0.340	580	564
30	22.9	0.238	0.317	570	564
Reference‡	56.1	0.435	0.537	552	287

† Cycle 0 involves only the initial model building by *PHENIX.AutoBuild*. ‡ The reference is from a 20-cycle iteration of 'direct-methods-aided model completion' with the same starting model from cycle 0 but without using the known phase information. Results of the 'best model' among the 20-cycle iteration are listed here.

other hand, the process in fact applies additional phase restraints to the 'direct-methods-aided model completion' procedure. It would be interesting to see how the phase restraint affects the result. 20 cycles of direct-methods-aided model completion without using the MAD phases were performed, starting with the partial structure from cycle 0. The 'best result' is listed in Table 5 as the 'Reference' and the 'best model' is shown in Fig. 7(*f*). Hence, for this example, structure model extension from cycle 0 will simply fail without restraints from the 5.3 Å MAD phases. Finally, not only phases but also the structure model are to be extended in the process and all programs involved are indispensable for success. This example shows also that changing constraints/restraints in the dual-space iterative framework of crystal structure analysis may lead to greatly different results.

3.4.2. Dual-space collaborative phase/model extension for 1lia. The structure of 1lia was solved at 2.8 Å by the MIR method (Chang *et al.*, 1996). For the present test, we take only

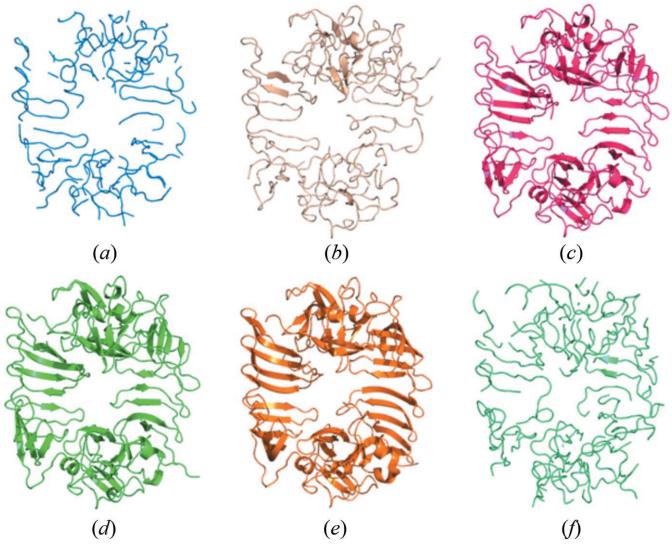

Figure 7
Results of dual-space collaborative phase/model extension for 1h3i (from 5.3 Å MAD phases to 2.8 Å structure). (*a*) Model from the initial cycle; (*b*) model from cycle 8; (*c*) model from cycle 20; (*d*) model from cycle 30; (*e*) final model from the PDB entry 1h3i; (*f*) model obtained without using known phases.

Table 6
Results of dual-space collaborative phase/model extension for 1lia (from 5.8 Å MIR phases to 2.8 Å structure).

Cycle	*R*	*R*-free	No. of residues built	No. of residues placed (sequenced)
0†	0.39	0.46	468	32
3	0.361	0.455	714	626
11	0.283	0.376	736	656

† Cycle 0 involves only the initial model building by *PHENIX.AutoBuild*.

diffraction data from the native crystal and a set of known MIR phases extended to 5.8 Å resolution. Phase extension was performed by 15 cycles of iteration. The results of cycles 0, 3 and 11 are listed in Table 6 and Fig. 8. Again, the extension is very successful.

4. Concluding remarks

Looking at the dual-space iterative framework, different methods work together for a common target – the structure. They are collaborators rather than competitors. The function of direct methods in the dual-space framework is to provide phase restraints. The combination of direct methods with other methods has produced fruitful results. There are unlimited ways to modify constraints/restraints within the dual-space iterative framework.

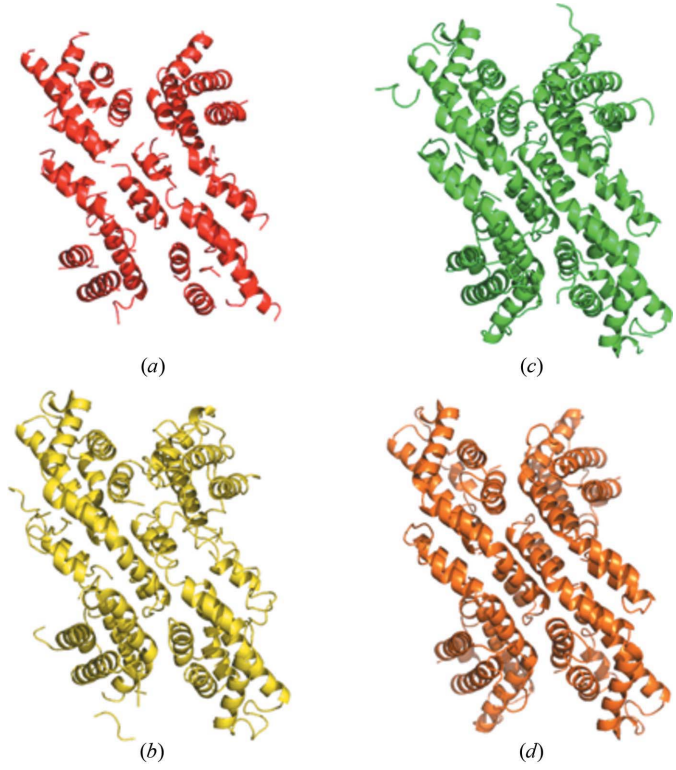

Figure 8
Results of dual-space collaborative phase/model extension for 1lia (from 5.8 Å MIR phases to 2.8 Å structure). (*a*) Model from the initial cycle; (*b*) model from cycle 3; (*c*) model from cycle 11; (*d*) final model from the PDB entry 1lia.

We acknowledge support from the Innovation Project of the Chinese Academy of Sciences and the 973 Project (grant No. 2002CB713801) of the Ministry of Science and Technology of China. HF and YG would like to thank Professor M. Cygler for the data for LegC3N, Dr Z. Dauter for the data for xylanase, Drs S. J. Gamblin and B. Xiao for the data for 1h3i, Professor D. C. Liang for the data for 1lia, Professor B. D. Sha for the data for Tom70p, Professor N. Watanabe for the data for TT0570 and TTHA, and Dr N. Yan for the data for UraA. The authors also thank the referees for their comments, which improved the manuscript. In the author list (arranged in ascending order of the family names), H. Fan and Y. Gu are joint principal investigators of the long-term project, and have been since the early 1990s. Z. Lin is the co-supervisor during Y. He's PhD work. Y. He is the main worker in developing algorithms for direct-methods-aided model completion. He also took part in preparing the latest version of the program *OASIS* and performed test calculations for the present work. J. Wang is the main worker in developing algorithms for iterative direct-methods SAD phasing and took part in the test with the membrane protein UraA. D. Yao took part in the development of iterative direct-methods SAD phasing and in the test with LegC3N. T. Zhang wrote the pipeline *IPCAS*. He also took part in preparing the latest version of the program *OASIS* and performed test calculations for the present work.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.

Bing-Dong, S., Shen-Ping, L., Yuan-Xin, G., Hai-Fu, F., Ke, H., Jia-Xing, Y. & Woolfson, M. M. (1995). *Acta Cryst.* D**51**, 342–346.

Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Spagna, R. (2005). *J. Appl. Cryst.* **38**, 381–388.

Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Mallamo, M., Mazzone, A., Polidori, G. & Spagna, R. (2012). *J. Appl. Cryst.* **45**, 357–361.

Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Spagna, R. (2003). *J. Appl. Cryst.* **36**, 1103.

Chang, W. R., Jiang, T., Wan, Z. L., Zhang, J. P., Yang, Z. X. & Liang, D. C. (1996). *J. Mol. Biol.* **262**, 721–731.

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

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

Coulter, C. L. (1965). *J. Mol. Biol.* **12**, 292–295.

Cowtan, K. (1994). *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography*, **31**, 34–38.

Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.

DeLano, W. L. (2002). The *PyMOL* Molecular Graphics System. DeLano Scientific, San Carlos, USA.

Fan, H.-F. (1965*a*). *Acta Phys. Sin.* **21**, 1114–1118.

Fan, H.-F. (1965*b*). *Chin. Phys.* pp. 1429–1435.

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

Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Jia-xing, Y. & Chao-de, Z. (2000). *Acta Cryst.* D**56**, 1137–1147.

Fortier, S., Moore, N. J. & Fraser, M. E. (1985). *Acta Cryst.* A**41**, 571–577.

Gerchberg, R. W. & Saxton, W. O. (1971). *Optik*, **34**, 275–284.

Gerchberg, R. W. & Saxton, W. O. (1972). *Optik*, **35**, 237–246.

Giacovazzo, C. (1983). *Acta Cryst.* A**39**, 585–592.

Giacovazzo, C., Cascarano, G. & Zheng Chao-de (1988). *Acta Cryst.* A**44**, 45–51.

Giacovazzo, C. & Siliqi, D. (2004). *Acta Cryst.* D**60**, 73–82.

Giacovazzo, C., Siliqi, D. & Platas, J. G. (1995). *Acta Cryst.* A**51**, 811–820.

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

Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.

Hauptman, H. A. (1996). *Acta Cryst.* A**52**, 490–496.

Hauptman, H., Potter, S. & Weeks, C. M. (1982). *Acta Cryst.* A**38**, 294–300.

Hazell, A. C. (1970). *Nature (London)*, **227**, 269.

He, Y., Yao, D.-Q., Gu, Y.-X., Lin, Z.-J., Zheng, C.-D. & Fan, H.-F. (2007). *Acta Cryst.* D**63**, 793–799.

Heinerman, J. J. L., Krabbendam, H., Kroon, J. & Spek, A. L. (1978). *Acta Cryst.* A**34**, 447–450.

Hendrickson, W. A. (1971). *Acta Cryst.* B**27**, 1474–1475.

Karle, J. (1966). *Acta Cryst.* **21**, 273–276.

Karle, J. & Hauptman, H. (1956). *Acta Cryst.* **9**, 635–651.

Klop, E. A., Krabbendam, H. & Kroon, J. (1987). *Acta Cryst.* A**43**, 810–820.

Kyriakidis, C. E., Peschar, R. & Schenk, H. (1993). *Acta Cryst.* A**49**, 557–569.

Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nat. Protoc.* **3**, 1171–1179.

Liu, Y.-D., Gu, Y.-X., Zheng, C.-D., Hao, Q. & Fan, H.-F. (1999). *Acta Cryst.* D**55**, 846–848.

Lu, F., Li, S., Jiang, Y., Jiang, J., Fan, H., Lu, G., Deng, D., Dang, S., Zhang, X., Wang, J. & Yan, N. (2011). *Nature (London)*, **472**, 243–246.

Miao, J., Hodgson, K. O. & Sayre, D. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 6641–6645.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.

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.

Neidle, S. (1973). *Acta Cryst.* B**29**, 2645–2647.

Patterson, A. (1934). *Phys. Rev.* **46**, 372–376.

Ramagopal, U. A., Dauter, M. & Dauter, Z. (2003). *Acta Cryst.* D**59**, 1020–1027.

Rossmann, M. G. & Arnold, E. (2001). *Patterson and molecular replacement techniques.* In *International Tables for Crystallography*, Vol. B, *Reciprocal space*, edited by U. Shmueli, pp. 235–263. Dordrecht: Kluwer Academic Publishers.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.

Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.

Sheldrick, G. M. (1997). *Proceedings of the CCP4 Study Weekend: Recent Advances in Phasing*, edited by K. S. Wilson, G. Davies, A. Ashton & S. Bailey, pp. 147–158. Warrington: Daresbury Laboratory.

Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.

Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423–431.

Sikka, S. K. (1973). *Acta Cryst.* A**29**, 211–212.

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

Steitz, T. A. (1968). *Acta Cryst.* B**24**, 504–507.

Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.

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.

Verwer, P., Krabbendam, H. & Kroon, J. (1991). *Acta Cryst.* A**47**, 143–144.

Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.

Wang, J. W., Chen, J. R., Gu, Y. X., Zheng, C. D. & Fan, H. F. (2004). *Acta Cryst.* D**60**, 1991–1996.

Wang, J. W., Chen, J. R., Gu, Y. X., Zheng, C. D., Jiang, F. & Fan, H. F. (2004). *Acta Cryst.* D**60**, 1987–1990.

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.
Weeks, C. M., DeTitta, G. T., Miller, R. & Hauptman, H. A. (1993). *Acta Cryst.* D**49**, 179–181.
Wilson, J. R., Jing, C., Walker, P. A., Martin, S. R., Howell, S. A., Blackburn, G. M., Gamblin, S. J. & Xiao, B. (2002). *Cell*, **111**, 105–115.
Woolfson, M. M., Jia-Xing, Y. & Hai-Fu, F. (1997). *Acta Cryst.* D**53**, 673–681.
Woolfson, M. M. & Yao, J.-X. (1990). *Acta Cryst.* A**46**, 409–413.
Wu, Y. & Sha, B. (2006). *Nat. Struct. Mol. Biol.* **13**, 589–593.
Xu, H. & Hauptman, H. A. (2003). *Acta Cryst.* A**59**, 60–65.
Xu, H., Hauptman, H. A., Weeks, C. M. & Miller, R. (2000). *Acta Cryst.* D**56**, 238–240.
Yao, J. X., Dodson, E. J., Wilson, K. S. & Woolfson, M. M. (2006). *Acta Cryst.* D**62**, 901–908.
Yao, D., Huang, S., Wang, J., Gu, Y., Zheng, C., Fan, H., Watanabe, N. & Tanaka, I. (2006). *Acta Cryst.* D**62**, 883–890.
Yu-dong, L., Harvey, I., Yuan-xin, G., Chao-de, C.-D., Yi-zong, H., Hai-fu, F., Hasnain, S. S. & Hao, Q. (1999). *Acta Cryst.* D**55**, 1620–1622.
Zhang, T., He, Y., Wang, J. W., Wu, L. J., Zheng, C. D., Hao, Q., Gu, Y. X. & Fan, H. F. (2012*a*). *OASIS*4.2. *A computer program of direct-method phase extension for proteins*. Institute of Physics, Chinese Academy of Sciences, People's Republic of China (available at http://cryst.iphy.ac.cn).
Zhang, T., He, Y., Wang, J. W., Wu, L. J., Zheng, C. D., Hao, Q., Gu, Y. X. & Fan, H. F. (2012*b*). *IPCAS*1.0. *Iterative protein crystal structure automatic solution*. Institute of Physics, Chinese Academy of Sciences, People's Republic of China (available at http://cryst.iphy.ac.cn).
Zhang, T., Wu, L. J., Gu, Y. X., Zheng, C. D. & Fan, H. F. (2010). *Chin. Phys. B*, **19**, 096101.
Zheng, X.-F., Zheng, C.-D., Gu, Y.-X., Mo, Y.-D., Fan, H.-F. & Hao, Q. (1997). *Acta Cryst.* D**53**, 49–55.