Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the combination of molecular replacement and single-wavelength anomalous diffraction phasing for automated structure determination

Santosh Panjikar,* Venkataraman Parthasarathy, Victor S. Lamzin, Manfred S. Weiss and Paul A. Tucker

EMBL Hamburg Outstation, c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany

Correspondence e-mail: panjikar@embl-hamburg.de

Received 3 June 2009
Accepted 24 July 2009

A combination of molecular replacement and single-wavelength anomalous diffraction phasing has been incorporated into the automated structure-determination platform *Auto-Rickshaw*. The complete MRSAD procedure includes molecular replacement, model refinement, experimental phasing, phase improvement and automated model building. The improvement over the standard SAD or MR approaches is illustrated by ten test cases taken from the JCSG diffraction data-set database. Poor MR or SAD phases with phase errors larger than 70° can be improved using the described procedure and a large fraction of the model can be determined in a purely automatic manner from X-ray data extending to better than 2.6 Å resolution.

1. Introduction

As of May 2009, more than 57 000 three-dimensional structures of biological macromolecules had been deposited in the Protein Data Bank (PDB; Berman *et al.*, 2000). With the availability of an ever-increasing number of potential search models among previously determined structures, molecular replacement (MR) has become the predominant technique for the determination of further structures. For the year 2007, it has been reported that more than two thirds of all newly deposited structures in the PDB could be solved using MR (Long *et al.*, 2008). Different approaches for MR have been realised, including the use of Patterson map techniques (*e.g.* Rossmann & Blow, 1962; Huber, 1965; DeLano & Brünger, 1995), structure-factor correlation (Navaza, 1987) and statistical targets (Bricogne, 1992, 1997; Read, 2001). As a consequence, a number of good and easy-to-use MR programs have become available. Examples include *AMoRe* (Navaza, 1994), *MOLREP* (Vagin & Teplyakov, 1997), *CNS* (Brünger *et al.*, 1998), *EPMR* (Kissinger *et al.*, 1999), *QS* (Glykos & Kokkinidis, 2000) and *Phaser* (McCoy *et al.*, 2005).

In principle, MR can lead to a successful structure determination within hours or even minutes. Often, however, the method is not straightforward in practice. The model derived from an MR solution inherently suffers from model bias, which can become severe, especially when the root-mean-square difference (r.m.s.d.) between the search model and the target structure is high. Reduction of the model bias and model completion can become a challenging issue at resolutions lower than 2.3 Å and often requires iterative time-consuming manual correction of the model using computer graphics alternating with model refinement. The standard methods for bias removal include omission of parts of the model, allowance for model errors in the refinement target functions and map coefficients, map-averaging techniques

doi:10.1107/S0907444909029643

Table 1
Description of the test cases.

PDB code†	Residues per subunit	Subunits in the AU‡	Se atoms in the AU‡	$d_{\rm min}$ (Å)	Space group	Search model PDB code§	Sequence identity (%)	R.m.s.d.¶ (Å)	$R_{\rm merge}$ (%)	$R_{\rm anom}$ (%)	Redundancy	$R_{\rm anom}/R_{\rm p.i.m.}$
2hh6	116	1	6	2.04	$P6_522$	2o4t_*A* (72)	44	1.04 (72)	5.6	10.0	7.5	1.43
2gi3	475	1	13	1.80	$P3_221$	2g5i_*A* (466)	50	1.34 (403)	5.8	11.2	6.3	1.19
1vmf	133	3	15	1.90	$P2_12_12_1$	1xbf_*A* (131)	50	0.67 (131)	10.7	13.4	4.1	1.41
1zbt	358	1	8	2.40	$P4_32_12$	2b3t_*B* (299)	49	2.43 (224)	4.3	7.1	6.7	1.45
1vmi	323	1	8	2.32	$P6_322$	1xco_*F* (329)	42	1.76 (311)	4.6	6.2	5.0	1.48
2f4l	283	4	28	2.50	$P2_12_12_1$	2ii1_*C* (277)	36	1.03 (274)	7.5	8.4	3.8	1.49
1vjo	381	2	16	2.00	$P2_12_12_1$	2ch2_*D* (374)	43	1.15 (371)	6.1	6.4	4.1	1.68
1vjf	168	1	8	2.20	$P4_32_12$	1vki_*B* (163)	48	1.28 (160)	5.6	6.7	6.7	2.00
1vjr	259	1	6	2.40	$P4_12_12$	1zjj_*B* (251)	40	1.55 (240)	7.6	7.2	4.6	2.00
1vkn	339	4	32	2.45	$P2_1$	1xyg_*D* (333)	46	1.18 (316)	6.9	6.5	6.1	2.40

† Data sets were taken from the Joint Center for Structural Genomics (JCSG) data depository: 2hh6 (Hoffmüller *et al.*, 2000), 1vjf (Han *et al.*, 2005), 2g5i (Nakamura *et al.*, 2006), 1xbf (Kuzin *et al.*, unpublished work), 2b3t (Graille *et al.*, 2005), 1xco (Xu *et al.*, 2005), 2ch2 (Rossi *et al.*, 2006), 1zjj (Yamamoto & Kunishima, unpublished work), 1xyg (Center for Eukaryotic Structural Genomics, unpublished work), 2gi3, 1vmf, 1zbt, 1vmi, 2f4l, 1vjf, 1vjr, 1vkn, 2o4t, 2ii1, 1vki (JCSG, unpublished work). ‡ AU, asymmetric unit. § PDB code with chain identifier of the search model. The number of residues in the search model is given in parentheses. ¶ R.m.s.d. between target and search model as calculated using the program *SUPERPOSE* (Collaborative Computational Project, Number 4, 1994). The number of C$^\alpha$ atoms aligned is given in parentheses.

(Main, 1967; Bricogne, 1976; Kleywegt & Read, 1997) and free-atom modelling, refinement and model building (Perrakis *et al.*, 1999). During refinement, implementation of maximum-likelihood (ML) targets (Murshudov *et al.*, 1997; Brünger *et al.*, 1998) together with $\sigma_{\rm A}$-weighted map coefficients (Read, 1986) to produce electron-density maps of the form ($2m|F_{\rm obs}| - D|F_{\rm calc}|$, $\alpha_{\rm calc}$) can significantly reduce model bias. 'Classical' OMIT maps (Bhat, 1988; Bhat & Cohen, 1984), $\sigma_{\rm A}$-weighted OMIT maps (Read, 1986, 1990), shake OMIT maps (Zeng *et al.*, 1997) and simulated-annealing OMIT maps (Hodel *et al.*, 1992) are often used for this purpose. The statistical-based reciprocal-space density-modification method (*Prime&Switch*) can be applied to initial experimental maps or model-phased maps. This has been implemented in the program *RESOLVE* and performs well at low resolution with marginal models (Terwilliger, 1999, 2000). Recently, an efficient bias-removal protocol '*Shake&wARP*' has been made available as a web service (Reddy *et al.*, 2003) using a combination of *EPMR* (Kissinger *et al.*, 1999) and the *CCP*4 suite of programs (Collaborative Computational Project, Number 4, 1994). Finally, the direct-method program *OASIS* (Hao *et al.*, 2000) has been extended to perform dual-space molecular-replacement model completion (He *et al.*, 2007).

The second most important phasing method in macromolecular crystallography is single-wavelength anomalous diffraction (SAD). SAD is based on accurately collected anomalous intensity differences arising from the presence of heavy atoms. Naturally, determination of the substructure becomes easier when an anomalous difference Fourier synthesis can be calculated using preliminary phases from an MR solution. The subsequent use of this substructure to generate an unbiased electron-density map (Baker *et al.*, 1995) is often referred to as MRSAD (molecular replacement with single-wavelength anomalous diffraction; Schuermann & Tanner, 2003).

In the past few years, several automated structure-determination pipelines have been developed with varying degrees of automation and often with rather different goals. These include *ACrS* (Brunzelle *et al.*, 2003), *PHENIX* (Adams *et al.*, 2002), *ELVES* (Holton & Alber, 2004), *CRANK* (Ness *et al.*, 2004), *SGXPro* (Fu *et al.*, 2005), *Auto-Rickshaw* (Panjikar *et al.*, 2005), *autoSHARP* (Vonrhein *et al.*, 2006) and *HKL*-3000 (Minor *et al.*, 2006). Most of them are based on experimental phasing approaches. More recently, software aimed at automatically assembling the set of 'best' models for MR has also been developed. Examples are *MrBUMP* (Keegan & Winn, 2007) and *BALBES* (Long *et al.*, 2008). The MR software pipelines make several decisions concerning the actual protocol for sequence alignment and homology modelling, the truncation of the model in regions of uncertain homology and the choice of the MR software engine. The current consensus approach is to derive a variety of models and to try MR for all of them one by one, followed by preliminary refinement and ranking of each potential solution.

Here, we demonstrate that by using some of the above-mentioned developments structure solution by a combination of MR and SAD can be automated and that even poor MR or SAD phases can be significantly improved. This approach is useful for the validation of MR solutions and for the reduction of model/phase bias. It is especially practical in cases in which the anomalous signal is not sufficiently strong to solve the structure by experimental phasing but is good enough to bootstrap the structure starting from a preliminary MR solution. The incorporation of the method into *Auto-Rickshaw* allows the fully automated determination of a large fraction of the structure from X-ray data extending to better than 2.6 Å resolution for most cases studied.

2. Materials and methods

2.1. Test cases

Ten test cases were selected from the JCSG data depository (http://www.jcsg.org/datasets-info.shtml). All of these data sets were collected at the high-energy side of the selenium *K* absorption edge (Table 1). The examples covered maximum resolutions ranging from 1.8 to 2.5 Å, were distributed among various crystal forms and seven different space groups and contained between 116 and 1356 amino-acid residues in the asymmetric unit. The sequence identity of the available search models to the target structures ranged between 36 and 51%.

2.2. Selection of search models for MR

The program *MrBUMP* (Keegan & Winn, 2007) was used for search-model selection based on the sequence identity to the target structure as the main selection criterion. The quality of the search model was assessed by calculating the r.m.s.d. to the homologous part of the target model. When the final refined target structure was superimposed onto the corresponding search model, the r.m.s.d. values ranged from 0.7 to 2.4 Å based on 72–371 superimposed C^{α} atom pairs.

2.3. The MRSAD approach

The process of MRSAD is shown schematically in Fig. 1. The one common entry point to the MRSAD procedure is a set of heavy-atom sites $X_{\rm H}$. These sites can be determined from the observed anomalous differences $\Delta F_{\rm o}$ either *via* heavy-atom substructure determination by Patterson, direct-methods or dual-space techniques or *via* model phases $\alpha_{\rm c,MR}$ resulting from an MR solution or a partial model. The sites are used to compute an initial set of phases $\alpha_{\rm SAD}$, which are improved by density modification, noncrystallographic symmetry (NCS) averaging (where applicable), phase extension *etc.* to yield the modified phases $\alpha_{\rm MOD}$, which in turn are the starting phases for model building. In the second cycle, the model phases $\alpha_{\rm C,SAD}$ derived from the built partial model are used to update the heavy-atom sites and are then combined with the SAD phases derived from the updated heavy-atom substructure. The resultant combined phases $\alpha_{\rm COMB}$ are then used again for density modification and model building. The procedure is repeated until most of the structure has been built.

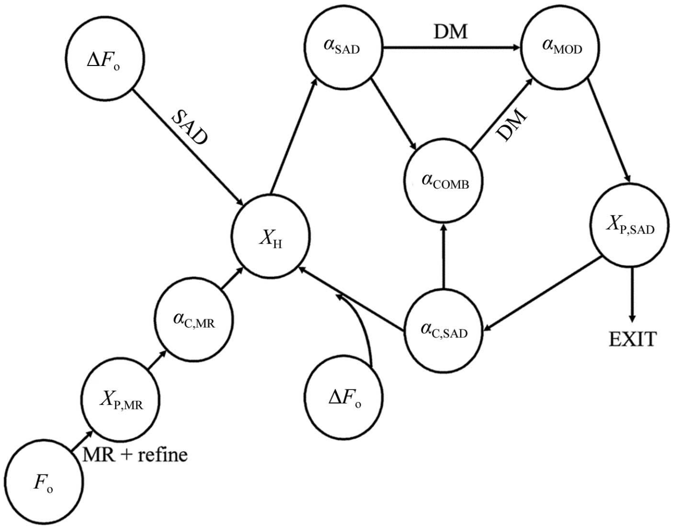

Figure 1
Flowchart for the phase-improvement cycle based on the initial MR or SAD phases. At first, heavy-atom positions ($X_{\rm H}$) are determined either based on MR phases ($\alpha_{\rm C,MR}$) calculated from a refined MR model ($X_{\rm P,MR}$) or by employing standard substructure-determination techniques based on anomalous differences. SAD phases ($\alpha_{\rm SAD}$) are then generated. No phase combination with MR phases takes place in the first cycle and the SAD phases are used for density modification (DM). Once a partial model has been built ($X_{\rm P,SAD}$) in the improved electron density from the first cycle, the model phases ($\alpha_{\rm C,SAD}$) are used to update the heavy-atom positions using an anomalous difference Fourier synthesis and phase combination with the SAD phases is then carried out for all subsequent cycles.

2.4. Evaluation of the results

The success of the MRSAD protocol was judged on the basis of the fraction of the total amino-acid residues built as well as by the $R_{\rm free}$ of the refined partial model. In our experience, for structures traced to a reasonable completeness the fraction of the side chains docked is a good indicator of the overall quality of the model (data not shown).

3. Implementation

The MRSAD approach has been implemented in the automated structure-determination pipeline *Auto-Rickshaw*. The respective crystallographic computer programs invoked at every step are depicted in the MRSAD flowchart (Fig. 2). In *Auto-Rickshaw*, the required input parameters for the MRSAD protocol include only the space group, the number of amino-acid residues per subunit, the number of subunits in the asymmetric unit, the amino-acid sequence of the target structure or a search model and native or anomalous data. The *Auto-Rickshaw* web server (http:/www.embl-hamburg.de/Auto-Rickshaw) allows the user to follow the progress of the structure determination conveniently. It also provides visualization of the resulting model and the possibility to download all relevant files for further inspection. An initial overview of the *Auto-Rickshaw* framework has been described previously (Panjikar *et al.*, 2005). In the following, the various tasks performed in *Auto-Rickshaw* are described in more detail.

3.1. Step 1: molecular replacement

(i) If a search model for MR is provided as input and if the difference in the unit-cell parameters between the search model and the input X-ray data is larger than 1%, MR is performed using the program *MOLREP* (Vagin & Teplyakov, 1997). Otherwise, this step is skipped and the model is refined directly (see below). (ii) If the amino-acid sequence of the target structure is provided by the user, the MR pipeline *BALBES* (Long *et al.*, 2008) is executed, which uses the models of domains from its own database and refines potential solutions using *REFMAC*5. *Auto-Rickshaw* then proceeds to the next step using the best MR model provided by *BALBES*.

3.2. Step 2: refinement of the MR model

This step involves rigid-body refinement of each chain of the MR model using *CNS* (Brünger *et al.*, 1998) at 4 Å resolution. Afterwards, positional and *B*-factor refinement at 3.0 Å resolution is carried out. The resulting model is then used for refinement with *REFMAC*5 (Murshudov *et al.*, 1997) to the maximum resolution of the provided X-ray data. If the asymmetric unit contains more than one molecule and if the resolution is lower than 1.8 Å, NCS restraints are included in the refinement. Once an $R_{\rm free}$ of less than 30% is reached, the process is terminated; otherwise, it continues with the next step.

3.3. Step 3: density modification and model building

When the resolution of the X-ray data is 2.6 Å or higher, phases calculated from the refined model are subjected to statistical density modification using *PIRATE* (Cowtan, 2000). The resultant phases are then used for automated model building using *ARP*/*wARP* (Perrakis *et al.*, 1999). When the resolution is lower than 2.6 Å, '*Prime&Switch*'-based density modification and model building are performed using *RESOLVE* (Terwilliger, 1999, 2000).

3.4. Step 4: anomalous difference Fourier and site selection

This step can only be performed when the input intensity file contains the Friedel pairs. The model phases and the anomalous differences are combined into a single MTZ file using *CAD* (Collaborative Computational Project, Number 4, 1994) and an anomalous difference Fourier map is calculated with *FFT* (Collaborative Computational Project, Number 4, 1994). A peak search is performed with *PEAKMAX* (Collaborative Computational Project, Number 4, 1994) and the site selection is based on the peak-height list produced. Initially, all sites above 5σ (where σ denotes the standard deviation of the anomalous difference Fourier map) are considered. Then, only sites which are above the threshold identified by a drop in the peak height of more than 65% between successive sites are selected. If no such drop can be identified in the peak list, the remainder of the peak list is searched until the peak height reaches 4.5σ. If the substructure model is poor (peak heights between 5σ and 9σ), *RESOLVE* is invoked for '*Prime&Switch*' density modification and the resultant phases are used for the substructure solution. If all peak heights are above 13σ, *SHELXC* and *SHELXE* (Sheldrick, 2008) are used for density modification. *SHELXE* is executed for 400 cycles using the 'free-lunch' algorithm (Caliandro *et al.*, 2007). The phases and structure factors are theoretically extended to 1.5 Å if the resolution of the experimental data is between 2.0 and 1.5 Å. The success of the procedure is gauged by the connectivity of the map. If this approach is successful, the next steps are skipped and the procedure continues with automated model building (see below) using *ARP*/*wARP*.

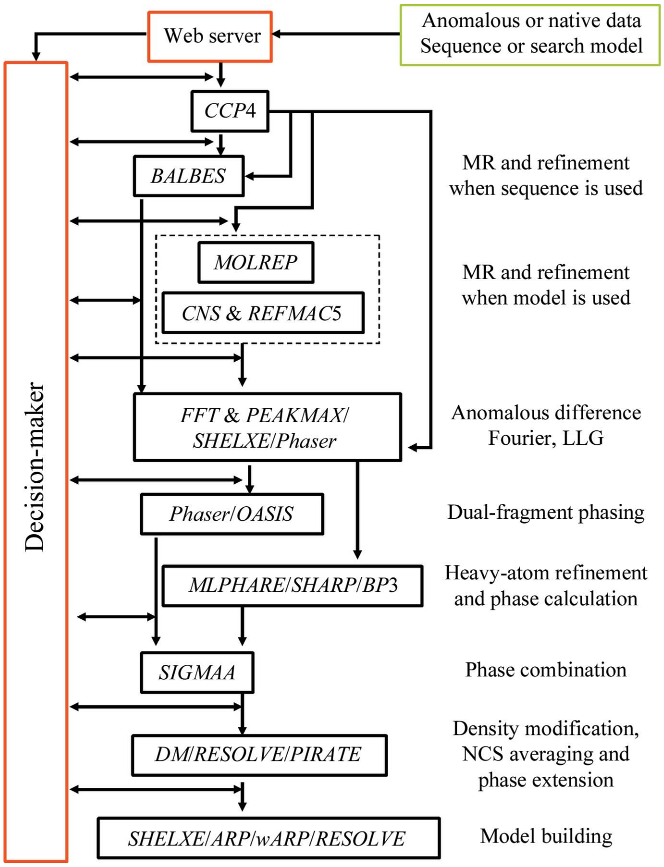
Web server
Anomalous or native data
Sequence or search model
Decision-maker
CCP4
BALBES
MR and refinement when sequence is used
MOLREP
CNS & REFMAC5
MR and refinement when model is used
FFT & PEAKMAX/ SHELXE/Phaser
Anomalous difference Fourier, LLG
Phaser/OASIS
Dual-fragment phasing
MLPHARE/SHARP/BP3
Heavy-atom refinement and phase calculation
SIGMAA
Phase combination
DM/RESOLVE/PIRATE
Density modification, NCS averaging and phase extension
SHELXE/ARP/wARP/RESOLVE
Model building

Figure 2
Architecture of the MRSAD protocol in *Auto-Rickshaw*. The existing crystallographic computer programs invoked by the pipeline are shown in black boxes, the web server and decision makers in red boxes and the user input in the green box. The steps from data reduction through to model building are addressed by the pipeline and run without user intervention. Data collection, processing, manual model completion and structure validation are not included.

3.5. Step 5: fragment phasing

The automatically refined MR model or the partial model resulting from *ARP*/*wARP* or *RESOLVE* and the heavy-atom sites found from the previous step as well as the anomalous data are used to produce a set of phases using *Phaser*. The purpose of this step is to validate the initial heavy-atom sites determined from the anomalous difference Fourier map and to find additional heavy sites which could not be detected in the map. When the MR solution is poor and the anomalous difference Fourier map does not generate heavy-atom sites with peaks higher than 5σ, *Phaser* may still be able to identify some low-occupancy sites. Should *Phaser* not succeed in producing a list of heavy-atom sites or if the heavy atoms are known from the previous step, *OASIS*-2006 (Zhang *et al.*, 2007) is used for dual-space phasing.

3.6. Step 6: heavy-atom refinement and phase calculation

The pipeline can invoke three heavy-atom refinement and phase-calculation programs: *MLPHARE* (Collaborative Computational Project, Number 4, 1994), *SHARP* (de La Fortelle & Bricogne, 1997) and *BP*3 (Pannu *et al.*, 2003; Pannu & Read, 2004). Initially, *MLPHARE* is executed to refine the occupancy of the sites to the maximum resolution of the data. If the figure of merit (FOM) does not exceed 10%, the resolution limit is decreased by 0.2 Å and the sites are refined at the lower resolution. If after this step the FOM has not risen above 15%, *SHARP* is used for refinement and phase calculation. If *SHARP* does not succeed, the refinement is continued using *BP*3.

Table 2
MR *versus* MRSAD phasing.

		Statistics after MR			Refinement of the MR model			Full MRSAD protocol		
PDB code	Residues in the AU	*MOLREP* R/CC† (%)	*CNS* $R/R_{\rm free}$‡ (%)	*REFMAC5* $R/R_{\rm free}$§ (%)	$\rm PC_{MR}$¶	*ARP/wARP* residues/ side chains††	*REFMAC5* $R/R_{\rm free}$§ (%)	$\rm PC_{MRSAD}$‡‡	*ARP/wARP* residues/ side chains††	*REFMAC5* $R/R_{\rm free}$§ (%)
2hh6	116	56.6/33.5	47.0/53.3	46.9/48.9	1	108 (108)	24.2/26.6	1	108 (108)	24.1/25.6
2gi3	475	50.7/32.2	38.0/45.2	39.1/41.2	1	422 (422)	18.5/21.6	1	424 (424)	18.0/21.9
1vmf	399	56.3/22.6	30.9/39.6	33.6/36.9	1	380 (288)	32.9/36.1	1	389 (389)	24.7/28.6
		51.6/35.6			2	377 (350)	26.9/31.4			
		45.0/50.5			3	373 (373)	26.6/29.9			
1zbt	358	54.7/37.3	42.1/47.9	40.3/43.2	1	203 (163)	38.1/40.6	1	216 (204)	29.3/32.6
					2	221 (183)	35.6/38.7	2	212 (201)	27.8/31.3
					3	222 (194)	33.1/35.0	3	225 (225)	25.7/28.9
					4	225 (202)	31.7/34.4			
1vmi	323	55.4/34.2	42.0/50.1	41.3/45.1	1	220 (168)	39.9/42.5	1	280 (275)	30.8/34.3
					2	260 (246)	31.3/36.1	2	281 (281)	24.7/29.7
					3	261 (247)	33.3/38.8			
					4	250 (235)	30.6/34.4			
2f4l	1132	58.8/41.4	34.7/43.8	37.3/38.3	1	782 (474)	41.9/46.4	1	924 (722)	38.4/43.7
		54.1/48.5						2	1002 (951)	24.8/31.1
								3	975 (936)	24.5/28.6
1vjo	762	53.8/28.4	36.1/43.1	39.3/42.5	1	654 (628)	24.5/28.7	1	679 (665)	22.1/27.5
		46.5/47.4								
1vjf	168	53.7/37.0	39.5/52.2	41.9/45.2	1	120 (109)	37.0/40.4	1	155 (155)	23.6/27.2
					2	126 (90)	42.7/48.3			
1vjr	259	54.5/34.9	41.6/46.8	40.1/42.3	1	248 (245)	21.7/24.5	1	250 (250)	21.5/23.4
1vkn	1354	56.5/18.4	42.2/48.9	46.4/47.2	1	515 (55)	49.6/53.0	1	693 (165)	46.4/49.2
		53.4/27.1						2	848 (569)	39.4/45.1
		50.5/35.2						3	871 (579)	40.0/42.3
		48.3/41.7								

† R-factor and correlation statistics after molecular replacement in *MOLREP* to 4 Å resolution for each structure. In the case of 2f4l, R/CC are given for two placed molecules at a time since the data set contains pseudo-translational symmetry. ‡ Working R-factor and free R-factor statistics after positional refinement in *CNS* to 3.0 Å resolution. § Working R-factor and free R-factor statistics after refinement in *REFMAC5* to the maximum resolution of the respective X-ray data. ¶ $\rm PC_{MR}$ stands for phasing cycle for molecular replacement, which consists of MR model completion using *OASIS*-2006 followed by density modification using *PIRATE*/*RESOLVE* and model building using *ARP*/*wARP*. †† Number of residues built and number of side chains docked using *ARP*/*wARP*. ‡‡ $\rm PC_{MRSAD}$ is a phasing cycle for MRSAD, which consists of steps 4–10 described in the text.

3.7. Step 7: phase combination

The two sets of phases calculated in steps 5 and 6 are combined using *SIGMAA* (Read, 1986). This step is skipped for an MR solution, when only native data are available. The resulting phases are improved by density modification and NCS averaging in *PIRATE* or *RESOLVE*.

3.8. Step 8: polyalanine model building

A beta version of *SHELXE* (Sheldrick, 2009) is used to build a polyalanine model using the phases calculated in the previous step. The updated substructure is used in step 5 if the cycle is repeated, for example when the model is not completed in the current cycle.

3.9. Step 9: automated model building and side-chain docking

The choice of programs for model building depends on the resolution of the X-ray data. If the value for the approximate resolution for 50% solvent content d_{50} [according to the formula $d_{50} = d_{\rm min}({\rm sc}^{-1} - 1)^{1/3}$, where $d_{\rm min}$ is the nominal maximum resolution of the X-ray data and sc is the solvent content of the crystal] is higher than 2.6 Å, the initial model building is carried out with *ARP*/*wARP* v.7.0.1. The number of building cycles is dependent on the map quality, which is assessed from the number of residues built in the first building cycle. If $d_{\rm min}$ is less than 2.0 Å and more than 70% of the model is built in the first building cycle, the total number of building cycles is set to five, whereas in all other cases ten building cycles are used. If the maximum resolution is lower than 2.6 Å then *RESOLVE* is used. When a polyalanine model is available from step 8, it is used as a starting model in *ARP*/*wARP* and density-modified phases are used for phased refinement in *REFMAC5* for iterative automated model building and side-chain docking. The benefit of the phased refinement is that *ARP*/*wARP* usually requires fewer building cycles. Similarly, if the model-building path uses *RESOLVE*, the polyalanine model is used as a starting model for further building and side-chain docking.

3.10. Step 10: refinement of the partial model

The model generated in step 9 is now refined to the maximum resolution of the data using *REFMAC5*. If the resultant $R_{\rm free}$ is lower than 30%, the automated procedure is considered to be complete. Otherwise, if the built model is less than 70% complete (using *RESOLVE*) or 90% complete (using *ARP*/*wARP*), an anomalous difference Fourier map is calculated based on the latest phase set and steps 4–10 are repeated. *Auto-Rickshaw* checks the improvement after every big cycle (steps 4–10). The improvement is gauged by the standard deviation of the local r.m.s. of the electron-density map after density modification, the total number of residues built, the $R_{\rm free}$ value from the refinement of the model and the

Table 3
SAD *versus* MRSAD phasing.

		SAD phasing			Refinement of the SAD model		Full MRSAD protocol		
PDB code	Residues in the AU	Resolution cutoff† (Å)	*SHELXD* CC_{all}/CC_{weak}‡	*ARP/wARP* residue/ side chain	*CNS* R/R_{free} (%)	*REFMAC*5 R/R_{free} (%)	PC_{MRSAD}	*ARP/wARP* residue/ side chain	*REFMAC*5 R/R_{free} (%)
2hh6	116	2.80	42.3/26.4	111 (111)	26.7/32.7	28.3/29.8	Skipped	Skipped	Skipped
2gi3§	475								
1vmf§	399								
1zbt	358	3.60	45.0/30.7	208 (165)	36.1/41.4	35.9/39.6	1	217 (196)	30.2/32.7
							2	219 (212)	26.6/29.5
1vmi	323	3.00	25.3/13.4	140 (8)	54.2/56.7	52.7/54.9	1	145 (0)	51.4/53.1
							2	181 (54)	48.7/50.4
							3	208 (103)	45.6/50.8
							4	265 (227)	33.5/38.5
							5	295 (265)	29.4/32.3
2f4l§	1132								
1vjo	762	2.80	47.8/32.3	430 (272)	44.1/48.5	44.8/47.7	1	685 (685)	21.8/25.9
1vjf	168	2.40	49.3/30.5	69 (5)	Skipped	Skipped	1	133 (133)	30.7/36.1
							2	156 (156)	21.1/25.2
1vjr	259	3.20	47.4/29.2	250 (250)	27.4/32.6	28.2/29.1	Skipped	Skipped	Skipped
1vkn	1354	2.45	46.8/28.9	209 (12)	Skipped	Skipped	1	464 (46)	49.7/51.4
							2	682 (215)	46.2/50.6
							3	834 (507)	41.1/45.2
							4	917 (633)	39.1/41.0

† Resolution cutoff for substructure determination. ‡ Correlation coefficients for all reflections and for the weak reflections only as computed by *SHELXD*. The MR step is skipped in the MRSAD procedure for all successful SAD cases. If the quality of a partial model resulting from SAD phases is not sufficiently high for refinement, the first part of the MRSAD protocol is skipped. If a model resulting from SAD phases can be refined in the first part of the MRSAD procedure below an R_{free} value of 30%, the experimental phasing part of the MRSAD protocol is skipped. § Substructure determination for the test cases 2gi3, 1vmf and 2f4l could not be achieved from anomalous differences.

absolute peak heights in the anomalous difference Fourier map. For MR based on native data alone, progress is assessed based on the fraction of the model built and the R_{free} from the refinement of the model.

In the case of MRSAD all of the abovementioned steps are carried out. However, when the substructure cannot be resolved because of a poor-quality model and/or poor anomalous data then the MRSAD protocol switches to a conventional MR recycling protocol. This protocol consists of steps 1–3 and steps 5, 9 and 10. The process is iterated until there is no further improvement from one cycle to the next. The MR recycling protocol can also be invoked using the native data and sequence or model information.

The major goals of the above implementation are to overcome the model bias from an MR solution, to build a more complete model from a partial and possibly fragmented preliminary model, and to use anomalous data in aiding model building in electron-density maps generated from MR phases.

4. Results and discussion

The data sets used to evaluate the MRSAD procedure of *Auto-Rickshaw* are listed in Table 1. The ten examples are sorted by increasing strength of the anomalous signal as indicated by the ratio $R_{anom}/R_{p.i.m.}$. Also shown are the PDB codes of the search models used for MR in each of the cases and the sequence-identity percentages of the search models. In order to evaluate the described MRSAD procedure, a comparison of MRSAD with a purely MR-based structure-determination procedure (Table 2) and with a standard SAD phasing procedure (Table 3) was performed using the ten test cases.

4.1. MR *versus* MRSAD

In Table 2, three approaches based on the primary phase information from an MR solution are compared with each other: the conventional MR procedure, the MR recycling procedure and the MRSAD procedure described above. The conventional MR procedure simply entails structure solution using MR and subsequent model refinement using *CNS* and *REFMAC*5. The MR recycling procedure is based upon iterative improvement of the MR phases. In each phasing cycle (PC_{MR}) the MR phases are improved by model completion using *OASIS*-2006, density modification using *PIRATE*/*RESOLVE* and model building using *ARP*/*wARP*. In the MRSAD procedure a phasing cycle (PC_{MRSAD}) consists of steps 4–10 described above. The numbers presented in Table 2 and graphically in Fig. 3(*a*) demonstrate that the MRSAD procedure yields a larger fraction of automatically built amino-acid residues and equally low or lower R_{free} values in all cases. The number of phasing cycles is also typically reduced, leading to quicker structure determination. This can be a decisive factor when structure determination is invoked whilst a user is at a synchrotron beamline collecting data, when quick answers are required in order to have an influence on further data-collection strategy. A striking example in this respect is the test case 1vkn. This structure contains four molecules of 339 residues each in the asymmetric unit in space group $P2_1$. The maximum resolution of the X-ray data is 2.45 Å. In this case, MR phasing alone was not sufficient to complete the model. Even after a round of MR recycling the free R factor was still above 50% and the model could not be improved any further. In contrast, three rounds of MRSAD cycling produced a model with about 60% of the residues built and about 40% of all side chains docked into the electron density.

4.2. SAD *versus* MRSAD

The described MRSAD procedure was also compared with a purely SAD phasing approach, as well as with a SAD phasing with subsequent model refinement approach (Table 3, Fig. 3*b*). For seven of the ten test cases the substructure could be solved, making them amenable to SAD phasing in the 'Advanced version' of *Auto-Rickshaw*. The remaining three cases (2gi3, 1vmf and 2f4l) were thus not further considered. For two of the seven successful cases, the SAD phases turned out to be so good that most of the structure was built automatically and that the free *R* factor was already below 30% after model refinement, so that no further improvement by MRSAD was anticipated. For the remaining five cases the improvement of MRSAD over SAD is clearly discernible from the numbers in Table 3 and the graphs in Fig. 3(*b*). 1vkn is again a striking case: SAD phasing alone was difficult in spite of the rather high $R_{\rm anom}/R_{\rm p.i.m.}$ ratio of 2.4. The automatically built model from SAD phases alone contained only 209 of the 1356 residues present in the asymmetric unit. This partial model was used as input for the MRSAD protocol and was directly fed into *Phaser* for SAD phasing and substructure completion. *Phaser* produced 40 heavy-atom sites, corresponding to 32 Se and eight S atoms. The sites were refined in *MLPHARE* to a maximum resolution of 2.45 Å. The phases from *Phaser* and *MLPHARE* were then combined and density modification and NCS averaging with *RESOLVE* were carried out followed by model building with *ARP*/*wARP*. In the first phasing cycle, the MRSAD protocol resulted in the building of 464 residues. This model was refined using *REFMAC5* to $R_{\rm work}$ and $R_{\rm free}$ values of 49.7% and 51.4%, respectively. In the second cycle, 682 residues were built and refinement of the model gave $R_{\rm work}$ and $R_{\rm free}$ values of 46.2% and 50.6%, respectively. In the fourth cycle, 917 residues were built and 633 residues were docked into the sequence. Refinement of the model resulted in $R_{\rm work}$ and $R_{\rm free}$ values of 39.1% and 41.1%, respectively. A further round of MRSAD phasing did not improve the total number of residues and $R_{\rm free}$ increased by 3%. Therefore, the procedure was halted at this point. The evolution of the electron density and the model is shown in Fig. 4. This particular example demonstrates that in cases when SAD phases are weak and insufficient to produce a good starting model the MRSAD protocol can rescue the situation.

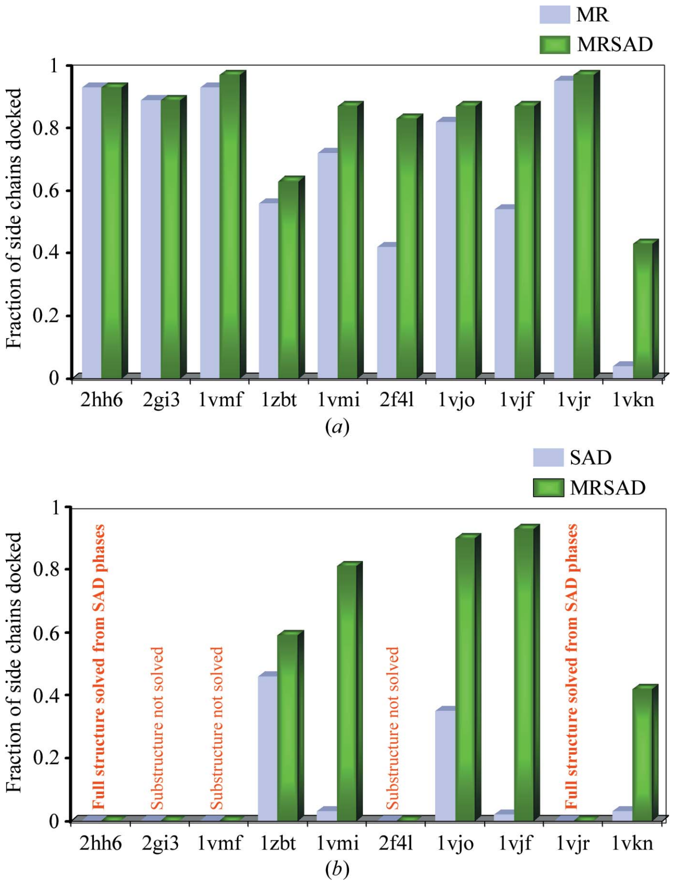

Figure 3
(*a*) MR *versus* MRSAD and (*b*) SAD *versus* MRSAD. (*a*) The fraction of side chains built in the electron density using *ARP*/*wARP* in the MR (blue) and MRSAD (green) phasing protocols for each data set (denoted by its PDB code); (*b*) the same for the SAD (blue) and MRSAD (green) phasing protocols. The fraction always refers to the total number of amino acids present in the asymmetric unit.

Since the implementation of the MRSAD phasing protocol in *Auto-Rickshaw* in August 2007, 84 users have used MRSAD to solve a total of 120 novel structures with resolutions ranging from 2.7 to 1.5 Å and the number of amino-acid residues in the asymmetric unit ranging from 100 to 3000. 46 structures were solved starting from a search model or from sequence information, whilst the remaining structure solutions started from experimental phases. A recent example is the crystal structure of *Plasmodium falciparum* profilin (Kursula *et al.*, 2008), where a partial model (60 residues of 171) was obtained using the three-wavelength Br MAD data sets. This model and the Br peak data set were used as a starting point in the MRSAD protocol, which provided an almost complete model.

The developed protocol can be applied to various kinds of problems. One particularly useful application is the model completion of protein–protein complex structures. As an example, the structure of vascular endothelial growth factor (VEGF-A) in complex with an engineered binding protein was solved using the MRSAD protocol based on a search model available for VEGF and using long-wavelength data (Giese & Skerra, unpublished work). Even for very large structures, such as, for instance, muconate-lactonizing enzyme from *Klebsiella pneumoniae* (3048 residues and eight subunits in the asymmetric unit; PDB entry 3fcp; Fedorov *et al.*, unpublished work), model completion has successfully been achieved.

5. Availability

The *Auto-Rickshaw* platform has been installed on a 68 CPU-core cluster at EMBL Hamburg. It is available *via* a web server

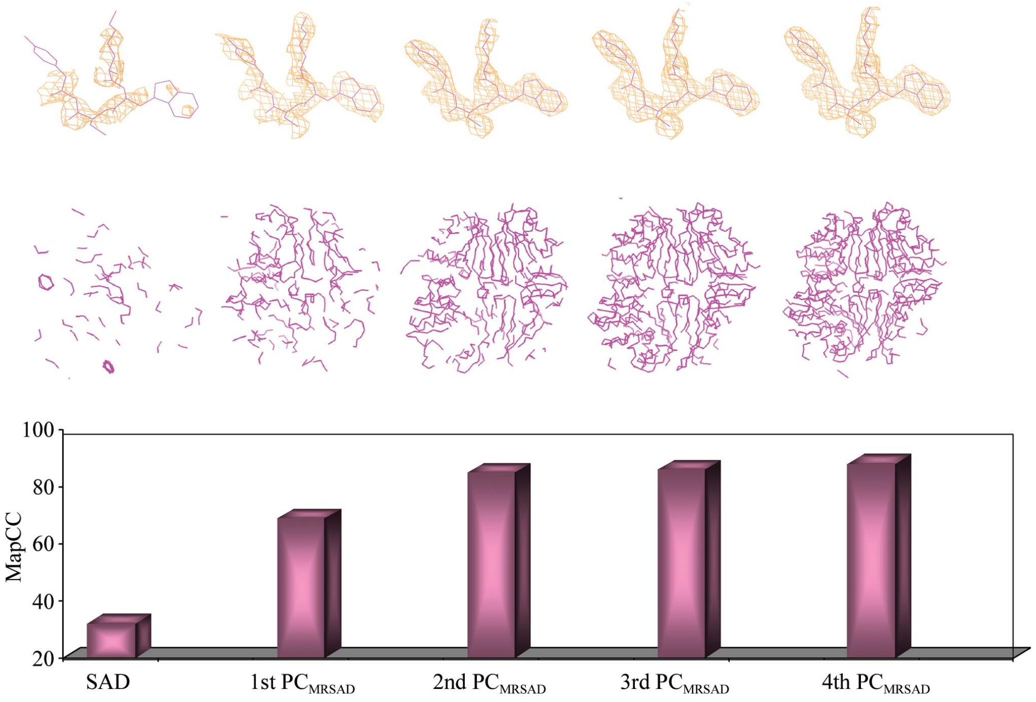

Figure 4
Evolution of the quality of the electron-density maps and the model completeness for the test case 1vkn (Joint Center for Structural Genomics, unpublished work). Part of the final refined model together with some example electron density and the fraction of the model built at each phasing cycle are shown. The electron-density map generated after each cycle after density modification was compared with the electron-density map generated from the final model in terms of map correlation coefficient (MapCC) and phase error. The MapCC values for the SAD phases and for the first to fourth cycle MRSAD phases are 32, 69, 85, 86 and 88% respectively, as shown in the bar-chart part of the figure and the phase errors for the same are 73, 48, 37, 34 and 32°, respectively.

at http://www.embl-hamburg.de/Auto-Rickshaw/. Registration and use of the server are free of charge for academic users.

6. Future perspectives

The *Auto-Rickshaw* platform is undergoing continuous development. This includes the incorporation of new functionalities as well as continuous software upgrades. A number of additional tasks will be incorporated into the MR and MRSAD protocols of the *Auto-Rickshaw* software pipeline in the future. These include the use of other molecular-replacement programs (*e.g. Phaser*), use of the SAD function (Skubák *et al.*, 2004) in refinement and a link to automatic data-collection software such as *DNA* (Leslie *et al.*, 2002) and automated data-processing systems such as *XIA*2 (http://www.ccp4.ac.uk/xia/). Another important aspect is the evolution and improvement of the decision making by evaluating an ever larger number of test cases and by extensive parameter screening in order to increase the efficiency of the coded decision making for the described phasing protocols.

We would like to express our thanks to the developers of the various computer programs for their kind permission to use their software in the *Auto-Rickshaw* pipeline. We also gratefully acknowledge the generous supply of X-ray data from the JCSG data depository by Ashley Deacon. The work was supported in part by the EC-funded BIOXHIT project (contract No. LHSG-CT-2003-503420).

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.
Baker, E. N., Anderson, B. F., Dobbs, A. J. & Dodson, E. J. (1995). *Acta Cryst.* D**51**, 282–289.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bhat, T. N. (1988). *J. Appl. Cryst.* **21**, 279–281.
Bhat, T. N. & Cohen, G. H. (1984). *J. Appl. Cryst.* **17**, 244–248.
Bricogne, G. (1976). *Acta Cryst.* A**32**, 832–847.
Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.
Bricogne, G. (1992). *Proceedings of the CCP4 Study Weekend. Molecular Replacement*, edited by W. Wolf, E. J. Dodson & S. Gover, pp. 62–75. Warrington: Daresbury Laboratory.
Brunzelle, J. S., Shafaee, P., Yang, X., Weigand, S., Ren, Z. & Anderson, W. F. (2003). *Acta Cryst.* D**59**, 1138–1144.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2007). *J. Appl. Cryst.* **40**, 931–937.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. (2000). *Acta Cryst.* D**56**, 1612–1621.
DeLano, W. L. & Brünger, A. T. (1995). *Acta Cryst.* D**51**, 740–748.

Fu, Z.-Q., Rose, J. & Wang, B.-C. (2005). *Acta Cryst.* D**61**, 951–959.

Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.

Graille, M., Heurgue-Hamard, V., Champ, S., Mora, L., Scrima, N., Ulryck, N., van Tilbeurgh, H. & Buckingham, R. H. (2005). *Mol. Cell*, **20**, 917–927.

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

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

Han, G. W. *et al.* (2005). *Proteins*, **58**, 971–975.

Hodel, A., Kim, S.-H. & Brünger, A. T. (1992). *Acta Cryst.* A**48**, 851–858.

Hoffmüller, U., Knaute, T., Hahn, M., Höhne, W., Schneider-Mergener, J. & Kramer, A. (2000). *EMBO J.* **19**, 4866–4874.

Holton, J. & Alber, T. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 1537–1542.

Huber, R. (1965). *Acta Cryst.* **19**, 353–356.

Keegan, R. M. & Winn, M. D. (2007). *Acta Cryst.* D**63**, 447–457.

Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.

Kleywegt, G. J. & Read, R. J. (1997). *Structure*, **5**, 1557–1569.

Kursula, I., Kursula, P., Ganter, M., Panjikar, S., Matuschewski, K. & Schueler, H. (2008). *Structure*, **16**, 1638–1648.

La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

Leslie, A. G. W., Powell, H. R., Winter, G., Svensson, O., Spruce, D., McSweeney, S., Love, D., Kinder, S., Duke, E. & Nave, C. (2002). *Acta Cryst.* D**58**, 1924–1928.

Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *Acta Cryst.* D**64**, 125–132.

Main, P. (1967). *Acta Cryst.* **23**, 50–54.

McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Cryst.* D**61**, 458–464.

Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. (2006). *Acta Cryst.* D**62**, 859–866.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Nakamura, A., Yao, M., Chimnaronk, S., Sakai, N. & Tanaka, I. (2006). *Science*, **312**, 1954–1958.

Navaza, J. (1987). *Acta Cryst.* A**43**, 645–653.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Ness, S. R., de Graaff, R. A., Abrahams, J. P. & Pannu, N. S. (2004). *Structure*, **12**, 1753–1761.

Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst.* D**61**, 449–457.

Pannu, N. S., McCoy, A. J. & Read, R. J. (2003). *Acta Cryst.* D**59**, 1801–1808.

Pannu, N. S. & Read, R. J. (2004). *Acta Cryst.* D**60**, 22–27.

Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.

Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.

Reddy, V., Swanson, S. M., Segelke, B., Kantardjieff, K. A., Sacchettini, J. C. & Rupp, B. (2003). *Acta Cryst.* D**59**, 2200–2210.

Rossi, F., Garavaglia, S., Giovenzana, G. B., Arca, B., Li, J. & Rizzi, M. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 5711–5716.

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

Schuermann, J. P. & Tanner, J. J. (2003). *Acta Cryst.* D**59**, 1731–1736.

Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.

Sheldrick, G. M. (2009). Submitted.

Skubák, P., Murshudov, G. N. & Pannu, N. S. (2004). *Acta Cryst.* D**60**, 2196–2201.

Terwilliger, T. C. (1999). *Acta Cryst.* D**55**, 1863–1871.

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

Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.

Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. (2006). *Methods Mol. Biol.* **364**, 215–230.

Xu, Q. S., Jancarik, J., Lou, Y., Kuznetsova, K., Yakunin, A. F., Yokota, H., Adams, P., Kim, R. & Kim, S.-H. (2005). *J. Struct. Funct. Genomics*, **6**, 269–279.

Zeng, Z.-H., Castaño, A. R., Segelke, B. W., Stura, E. A., Peterson, P. A. & Wilson, I. A. (1997). *Science*, **277**, 339–345.

Zhang, T., He, Y., Gu, Y.-X., Zheng, C.-D., Hao, Q., Wang, J.-W. & Fan, H.-F. (2007). *OASIS-2006: A Direct-Method Program for SAD/SIR Phasing and Reciprocal-Space Fragment Extension.* Institute of Physics, Chinese Academy of Sciences, Beijing, People's Republic of China. http://cryst.iphy.ac.cn/Project/program/oasis.html.