

# Extending the resolution and phase-quality limits in automated model building with iterative refinement

Pavol Skubák, Steven Ness and  
Navraj S. Pannu\*

Biophysical Structural Chemistry, Leiden  
Institute of Chemistry, Gorlaeus Laboratories,  
Leiden University, PO Box 9502,  
2300 RA Leiden, The Netherlands

Correspondence e-mail:  
raj@chem.leidenuniv.nl

Received 12 July 2005  
Accepted 10 October 2005

Previously, the direct use of prior phase information from a single-wavelength anomalous diffraction (SAD) experiment with a multivariate likelihood function applied to automated model building with iterative refinement has been proposed [Skubák *et al.* (2004), *Acta Cryst. D* **60**, 2196–2201]. In this approach, the anomalous information from the experimental data is used in refinement to derive phase information in a maximum-likelihood formalism and provided a more theoretically valid way of incorporating prior phase information compared with current approaches. In the present work, the SAD multivariate likelihood function that directly uses prior phase information is tested against currently used functions on many different SAD data sets which exhibit a wide range of resolution limits and anomalous signal. The results clearly show the importance of the more theoretically valid utilization of prior phase information: the SAD function extends the resolution and phase-quality limits needed for successful automated model building with iterative refinement. Indeed, the multivariate likelihood function reduces the overfitting in the refinement procedure and performs consistently better than the current refinement targets in terms of the quality of the models obtained and the number of residues built.

## 1. Introduction

A common occurrence in protein crystallography is a low observation-to-parameter ratio in model refinement. In order to alleviate this problem, all available prior information should be used, thus effectively increasing the ratio of observations to unknowns. Prior information about geometry (Engh & Huber, 1991) has been used as a standard in refinement programs. Several years ago, a refinement target using prior phase information was proposed and implemented in a target function denoted MLHL (Pannu *et al.*, 1998). In this approach, the phase distribution encoded in the form of Hendrickson–Lattman coefficients (Hendrickson & Lattman, 1970) is used. However, the MLHL function has approximations which might not be completely justified, including the assumption that the prior phase information is independent of the model. Furthermore, the MLHL target is strongly dependent on the accuracy and reliability of the phasing program used to generate the Hendrickson–Lattman coefficients. As a consequence, the indirect use of prior phase information may lead to poorer results than not using it at all (Calderone, 2004). Direct incorporation of experimental phase information (Skubák *et al.*, 2004) removes these shortcomings, taking into account the current model and substructure, the measured data (including its anomalous part), the correlations between structure factors and errors in the experiment and in the model within a multivariate likelihood formalism.

**Table 1**  
Data-set statistics.

Molecule	Substructure			$f''$ (approx.)	Resolution	Residues	Map correlation after <i>DM</i>	<i>wARP</i> cycles
	Type	Total	Found					
GerE	Se	12	12	3.9	2.73	384	0.5127	20
MutS	Se	46	45	5.0	3.00	1542	0.6606	20
Bacteriophage T4	Hg	1	1	11.0	2.74	198	0.6566	50
Transhydrogenase	Se	16	16	4.0	2.48	364	0.6506	30
AEP transaminase	Se	66	66	6.5	2.55	2169	0.7660	20
Cyanase	Se	40	39	3.9	2.40	1560	0.7433	30
Ribonuclease	Pt	5	5	6.9	2.50	192	0.4287	30
Crustacyanin	S	12	12	0.72	2.60	362	0.3704	30
$\beta$ -Mannosidase	Se	4	3	5.4	2.00	351	0.7340	10
Thioesterase I	Br	22	20	5.0	1.80	458	0.6984	10
PSCP	Br	9	8	5.0	1.80	371	0.4420	10
Lysozyme (360°)	S, Cl	12	9	0.56	1.64	129	0.7537	10
Lysozyme (270°)	S, Cl	12	8	0.56	1.64	129	0.5924	30
Lysozyme (180°)	S, Cl	12	7	0.56	1.64	129	0.4701	30
Ferredoxin	Fe	8	8	1.25	0.94	55	0.7441	10
Insulin	Zn	2	2	2.23	0.98	102	0.4101	30
Thionein	Cu	8	8	3.84	1.64	36	0.8573	20

Brunger (2005) stated that the use of the MLHL target with iterative and manual improvement of both the heavy-atom model and the experimental phase probability distribution was crucial for the solution of structures with low-resolution data. Since the multivariate likelihood function incorporates phase information directly and dynamically, it should be able to push the resolution limits for automated model building further in a compact and non-iterative fashion. Full automation of macromolecular model building at lower resolutions is an important challenge to high-throughput structure determination since interactive model building of large structures can be both time-consuming and challenging. Significant progress in this field has been achieved at medium to low resolution in automated model building with *MAID* (Levitt, 2002), *RESOLVE* (Terwilliger, 2004) and *ARP/wARP* (Morris *et al.*, 2004). However, as Badger (2003) has shown, there is still room for further improvement, especially at resolutions reaching 3 Å.

Although a refinement target that uses prior phase information directly has been proposed and implemented (Skubák *et al.*, 2004) with promising results, so far no extensive tests of the method have been reported. In this work, we compare the performance of the current refinement targets with the SAD multivariate likelihood function in automated model building and iterative refinement on a wide range of real test data sets with differing resolution limits and quality of initial phase estimates.

## 2. Methods

The multivariate SAD function is compared with the current refinement targets in automated model building with iterative refinement on a set of 15 different previously solved protein structures. This random sample of data sets used to test the SAD function exhibits a wide variety of resolution limits, data quality, anomalous signal and both number and type of

anomalous scatterers. The statistics for all the data sets are presented in Table 1.

All tests were performed on a personal computer with a 2.6 GHz Intel Pentium 4 processor running the GNU/Linux Mandrake 10.1 operating system. We used the automated model-building program *ARP/wARP* (v.6.1.1; Perrakis *et al.*, 1999) employing the model refinement program *REFMAC5* (v. 5.2.0005; Murshudov *et al.*, 1997) from the *CCP4* package (v.5.0.2; Collaborative Computational Project, Number 4, 1994). In the tests, three likelihood targets differing in how they utilize prior phase information are compared: the target lacking any prior phase information (Murshudov *et al.*, 1997; Bricogne & Irwin, 1996; Pannu & Read, 1996), denoted below as the Rice target, the function using the phase information indirectly and encoded in Hendrickson–Lattman coefficients (Pannu *et al.*, 1998), denoted as MLHL, and the target incorporating the information about phases from a SAD experiment directly (Skubák *et al.*, 2004), denoted as the SAD target. The Rice and MLHL targets were present in *REFMAC5*, while the same version of *REFMAC5* was modified to include the SAD function (Skubák & Pannu, unpublished work).

The sequence information for the protein was supplied to *ARP/wARP* and sequence docking was performed for the second half of the building process, as it improved results in a few difficult test cases. The selenomethionine residues in the sequence were replaced by methionines in the input to *ARP/wARP*, since v.6.1.1 does not recognize selenomethionine residues.

For all data sets, we randomly selected 5% of reflections for the free set (Brünger, 1992). Default settings were used in Luzzati error-parameter (Luzzati, 1952) estimation: for the Rice and MLHL functions the working set of reflections was used and for the SAD function the set of free reflections was used: these choices yielded the best results for every function (data not shown).

The *REFMAC5* 'XNON NO' option (Garib Murshudov, personal communication) was used in all runs: this option

causes only the diagonal terms of the Fisher matrix (Steiner *et al.*, 2003) to be used in the minimization. Consistently better results were achieved with this option for all the tested functions. No resolution cutoffs were used for any data set. The number of building cycles was different for every test case according to the rate of model building, but it was the same for all the targets used to refine the same protein (see Table 1). The number of cycles was set to the maximum required by the different targets so that model building was never stopped while the model was still improving (as judged by the improvement in map correlation). All other parameters were used as default in the running of the program for all the tests unless stated otherwise below.

For all test cases, the steps of heavy-atom detection, phasing and density modification were performed using the *CRANK* (v.0.9; Ness *et al.*, 2004) suite. *CRANK* uses the direct-methods program *CRUNCH2* (de Graaff *et al.*, 2001) using difference *E* values calculated by *DREAR* (Blessing & Smith, 1999) or the program *SHELXD* (Schneider & Sheldrick, 2002) for substructure detection, *BP3* (Pannu & Read, 2004) for substructure refinement and phasing and *DM* (Cowtan, 1999) for density modification. In most of the cases, the process of substructure detection, refinement and phasing and density modification was performed automatically using the default *CRANK* values. In a few cases, input parameters such as resolution cutoff for substructure detection, solvent content for density modification or the number of cycles of *CRANK* subprograms were changed in order to obtain better results. Information about non-crystallographic symmetry was not used and only the single-wavelength anomalous diffraction data set was used throughout.

The MLHL function requires the indirect phase information in the form of Hendrickson–Lattman (HL) coefficients. In most of the cases, the coefficients from *BP3* provided better results than HL coefficients from *DM*, so the MLHL results reported in all the following tables will refer to the use of HL coefficients generated by *BP3*. Those cases where the results were better using HL coefficients from *DM* are presented in the text.

The substructure atomic parameters required by the SAD function were obtained from *BP3*. All atoms found were used as input, including any incorrect ones. In most cases, the occupancies of incorrect atoms were refined to very small values by *BP3* and thus did not disturb the subsequent building and refinement by the SAD function. The heavy-atom coordinates were fixed since they are usually well refined by *BP3* and only isotropic temperature factors and occupancies of the substructure were refined by the SAD function.

### 3. Results

To compare models built using refinement against the Rice, MLHL and SAD multivariate likelihood refinement targets, we report the total number of backbone residues built by *ARP/wARP*. Because part of the model built might be traced incorrectly or misplaced, we also include the number of ‘correctly built’ residues. A residue is regarded as ‘correct’ if

its C $\alpha$  atom is placed within 1 Å of a C $\alpha$  position from the final model (*e.g.* Badger, 2003). We also quote the quality of this ‘correct’ part of the model by the root-mean-square (r.m.s.) error of the C $\alpha$  positions of the ‘correctly’ placed residues. The number of correct residues and the r.m.s. error of the correct part of the model are calculated by a compare-protein script (Ness & Skubák, unpublished work) within the *CRANK* suite. The map correlation with the final model is also shown and was computed using the program *SFTOOLS* (Bart Hazes, unpublished program).

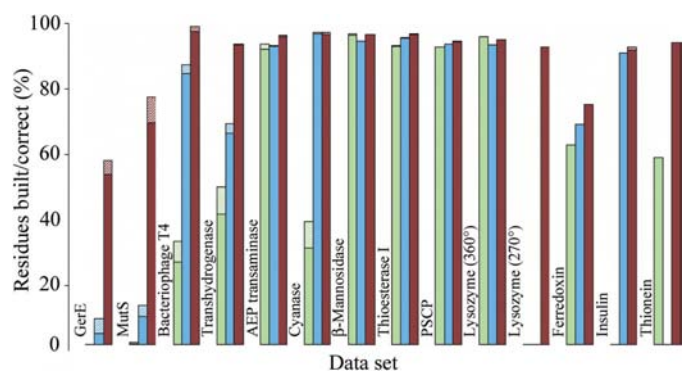
The statistics reported in the following tables are the values from the end of *ARP/wARP* model-building runs. The map correlation before the last model-building cycle is reported. In a few cases, the built models became worse as the building process continued. This is a consequence of problems in the refinement performed between the building cycles. For these cases, we present the number of built residues at the beginning of the run in the text.

Fig. 1 shows a global summary of the performance of all of the target functions for the data sets in terms of the number of residues built. Below, we describe the automated model-building results for each data set in two sections based on resolution.

#### 3.1. Test cases with 2.4 Å or lower resolution

**3.1.1. GerE.** *Bacillus subtilis* regulatory protein GerE derivative consists of six monomers with two selenomethionine residues in each monomer (Ducros *et al.*, 2001). The SAD function with the direct use of prior phase information was essential in improvement of the initial maps and also in building a significant part of the model as can be seen in Fig. 2(a). Both the MLHL and Rice functions failed in automated building and both targets demonstrated a great deal of overfitting as shown in Fig. 2(b) by a plot of the *R* factors as a function of building cycle for all the functions. When refinement was performed against the SAD target, overfitting was significantly reduced.

Despite the low-resolution data (2.73 Å) with relatively poor initial maps and phases (phase error of 62.8° after density



**Figure 1** The number of correctly built residues (solid part) and incorrectly built residues (cross-hatched part) using the Rice function (in green), the MLHL function (in blue) and the SAD function (in red). Only the cases with successfully built models are shown.

**Table 2**

The results of GerE model building.

GerE	Rice	MLHL	SAD
Map correlation	0.3808	0.533	0.6876
Residues built	0	31	221
Correct residues	0	13	204
R.m.s. error of model (Å)	—	0.533	0.547
Total run time (min)	68.1	78.87	117.3

**Table 3**

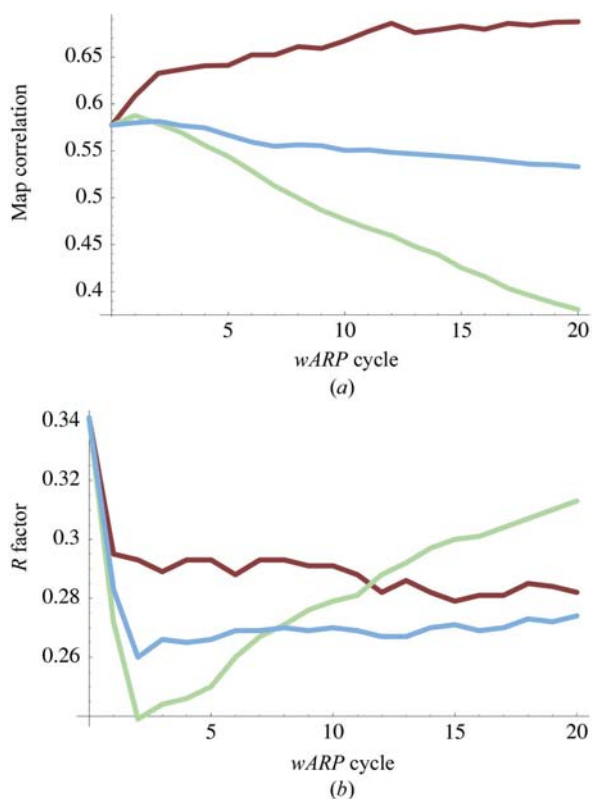
The results of MutS model building.

MutS	Rice	MLHL	SAD
Map correlation	0.5643	0.7074	0.8745
Residues built	11	190	1224
Correct residues	2	139	1092
R.m.s. error of model (Å)	0.759	0.609	0.385
Total run time (min)	327.77	378.33	533.08

**Table 4**

The results of bacteriophage T4 model building.

Bacteriophage T4	Rice	MLHL	SAD
Map correlation	0.6564	0.8811	0.8679
Residues built	64	173	197
Correct residues	51	168	194
R.m.s. error of model (Å)	0.548	0.268	0.247
Total run time (min)	337.74	368.03	429.74

**Figure 2**

The dependence of (a) map correlation and (b) *R* factor on automated model-building cycle for all three likelihood functions in the GerE test case. The green curve shows the results for the Rice function; the blue curve is for MLHL and the red curve is for the SAD function.

modification), a great part of the model was traced by *ARP/wARP* and *REFMAC5* using the SAD function (Table 2). One monomer was built to a high degree (83%), docking part of the side chains correctly. Almost all missing residues of this monomer were present in the fragments of the other built monomers. Although almost 8% of residues built were classified as incorrect, they did not form incorrectly built regions and their displacement was often just slightly above the 1 Å correctness criterion.

Interestingly, even better results with the SAD function were achieved when we did not refine the occupancies of the Se atoms. Over 250 backbone residues were correctly traced with an r.m.s. error of 0.46 Å and over 150 side chains were correctly docked. However, this behaviour was specific to the GerE case, since the refinement of substructure occupancies improved the model building in the other test cases.

**3.1.2. MutS.** The structure of the *Escherichia coli* DNA-repair protein MutS was originally determined from 3 Å resolution MAD diffraction data and a higher resolution (2.2 Å) data set all of which were from a selenomethionine crystal (Lamers *et al.*, 2000). Automated substructure detection, phasing and density modification from the peak 3.0 Å SAD data set with default parameters yielded a good initial map for model building (Table 1). For all refinement targets, around 800 residues were traced in the first building cycle, over 600 of which were placed correctly. Subsequent building cycles with the Rice and MLHL functions worsened the quality of the electron-density map, resulting in only 11 residues built with the Rice function and 190 with MLHL at the end of the building process (Table 3). *ARP/wARP* with the SAD function was able to improve the initial maps further and was able to correctly build over 70% of the backbone (almost 1100 residues). Plots of the change in map correlation and *R* factor with increasing model-building cycles resemble those for GerE (Fig. 2), with larger differences between the MLHL and SAD curves. In contrast to the other test cases, the MLHL refinement was better with static phase information from *DM*. 784 residues were traced with this approach, 659 of which were correct. The quality of the map was slightly improved at the beginning of the automated model-building process and then remained unchanged for the rest of process, with the final map correlation reaching 0.80.

**3.1.3. Bacteriophage T4.** The structure of the receptor-binding domain of bacteriophage T4 was originally solved using the SIRAS method (Thomassen *et al.*, 2003). The anomalous signal from the single mercury turned out to be of sufficient quality to solve the structure using the SAD method. Density modification again played an essential role in the breakdown of phase ambiguity as it improved the phase error from 73.4° to 48.1° owing to the very high solvent content.

For this protein, the process of model building was significantly slower than in the other test cases. *ARP/wARP* was only able to build a small part of the structure without prior phase information (Table 4), although the overfitting was not as strong as in the MutS case and the refinement did not cause worsening of map quality. Very complete models were built with targets utilizing prior phase information. Building with

the SAD function was more rapid than for MLHL, but approximately 35 rebuilding cycles were still required to build the model. However, at the end of building, 98% of the model was built correctly with an r.m.s. error of less than 0.25 Å.

**3.1.4. Transhydrogenase.** The structure of domain III of human heart transhydrogenase [NADP(H)-binding component; White *et al.*, 2000] was determined using an SeMet-derivative crystal diffracting to 2.48 Å. All 16 Se atoms were correctly identified and a reasonable map (with a correlation of 65%) was output by *CRANK*. The differences between the functions in terms of the quality of the maps produced are large, as can be observed in Fig. 3. These differences are reflected in the number of traced residues and the r.m.s. of models as stated in Table 5. With the SAD function, over 95% of the model was built correctly, including correct docking of all side-chain residues. One of the two monomers in the asymmetric unit was built almost completely, with only two outer residues missing of the total 182 residues. Using the Rice or MLHL functions, no monomer was completely built; at least 55 residues were missing.

**3.1.5. AEP transaminase.** The crystal structure of 2-aminoethylphosphonate (AEP) transaminase was originally determined by a multiwavelength anomalous diffraction experiment (Chen *et al.*, 2002). The automated process of substructure determination, phasing and density modification yielded a very good density map which was successfully traced by *ARP/wARP* using all the refinement functions, despite the moderate resolution of 2.55 Å.

The number of residues built was greater than 2000 and similar for all functions, but there were significant differences in the quality of the models built. The differences are reflected in the map correlations after building and in the r.m.s. statistics of the model built (Table 6): the correlation coefficient produced with the SAD function was significantly higher and the r.m.s. error of the model was lower.

**3.1.6. Cyanase.** *E. coli* cyanase was originally solved using MAD with an SeMet derivative at 2.4 Å and a 1.65 Å resolution native data set (Walsh *et al.*, 2000). In this case, the performance of automated model building was strongly

**Table 5**  
The results of transhydrogenase model building.

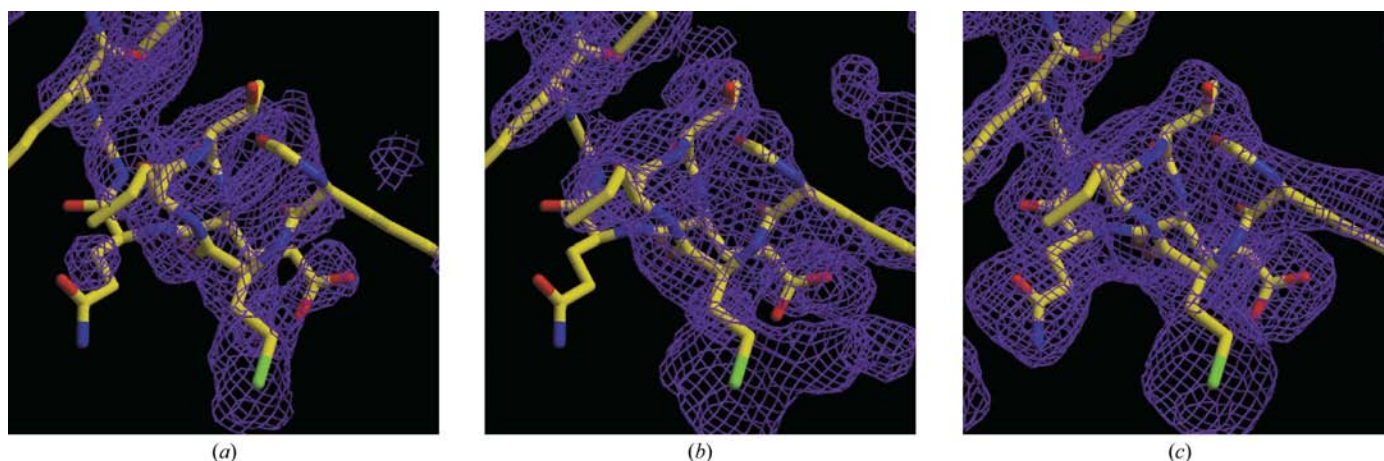
Transhydrogenase	Rice	MLHL	SAD
Map correlation	0.6697	0.7686	0.892
Residues built	182	255	348
Correct residues	151	244	347
R.m.s. error of model (Å)	0.523	0.376	0.219
Total run time (min)	178.87	189.05	206.94

**Table 6**  
The results of AEP transaminase model building.

AEP transaminase	Rice	MLHL	SAD
Map correlation	0.8415	0.8982	0.9322
Residues built	2056	2049	2117
Correct residues	2024	2043	2105
R.m.s. error of model (Å)	0.397	0.243	0.213
Total run time (min)	658.88	650.17	719.07

dependent on the use of experimental phase information. With the default *ARP/wARP* Rice function (*i.e.* using no phase restraints) the overall map correlation decreased from the initial 0.77 to 0.56 after 20 building cycles, probably owing to high overfitting in refinement. Therefore, the best models were built during the first building cycles with approximately 800–900 backbone residues reported, ~90% of which were placed correctly. In contrast, both functions using prior phases build very complete high-quality models with 97% of the backbone of all ten subunits with more than 90% of the correct side residues docked. The only significant difference between the SAD and MLHL functions was in the speed of building: *ARP/wARP* with the SAD function built the model significantly faster than with MLHL (Table 7).

**3.1.7. Ribonuclease and crustacyanin.** Automated model building failed for both the *Streptomyces aureofaciens* ribonuclease (Sevcik *et al.*, 1996) 2.5 Å platinum-derivative data set and crustacyanin (Gordon *et al.*, 2001) 2.6 Å sulfur SAD data set regardless of the refinement target used. The map correlation after phasing and density modification were only around 0.4 owing to small anomalous signal, although all substructure atoms were correctly identified.



**Figure 3**  
Electron-density maps of a small transhydrogenase region at the end of *ARP/wARP* procedures superimposed on the final deposited model. The maps (all contoured at 1σ) obtained when using (a) Rice, (b) MLHL and (c) SAD targets for refinement are shown.

**Table 7**

The results of cyanase model building.

Cyanase	Rice	MLHL	SAD
Map correlation	0.5592	0.9328	0.9333
Residues built	600	1524	1522
Correct residues	472	1515	1512
R.m.s. error of model (Å)	0.686	0.178	0.181
Total run time (min)	608.05	826.38	727.41

**Table 8**The results of model building of  $\beta$ -mannosidase, thioesterase, PSCP and lysozyme (360°).

	Rice	MLHL	SAD
<b><math>\beta</math>-Mannosidase</b>			
Map correlation	0.9528	0.9509	0.9546
Residues built	341	333	340
Correct residues	339	333	340
R.m.s. error of model (Å)	0.097	0.091	0.08
Total run time (min)	73.8	76.72	78.47
<b>Thioesterase I</b>			
Map correlation	0.8902	0.8948	0.8986
Residues built	428	440	445
Correct residues	427	439	444
R.m.s. error of model (Å)	0.157	0.153	0.157
Total run time (min)	58.78	61.21	65.45
<b>PSCP</b>			
Map correlation	0.6042	0.6106	0.6066
Residues built	349	352	356
Correct residues	349	352	355
R.m.s. error of model (Å)	0.355	0.351	0.351
Total run time (min)	84.13	85.71	94.84
<b>Lysozyme (360°)</b>			
Map correlation	0.9576	0.9533	0.9527
Residues built	124	121	123
Correct residues	124	121	123
R.m.s. error of model (Å)	0.077	0.104	0.081
Total run time (min)	31.39	32.21	35.52

### 3.2. Test cases with resolution higher than 2.4 Å

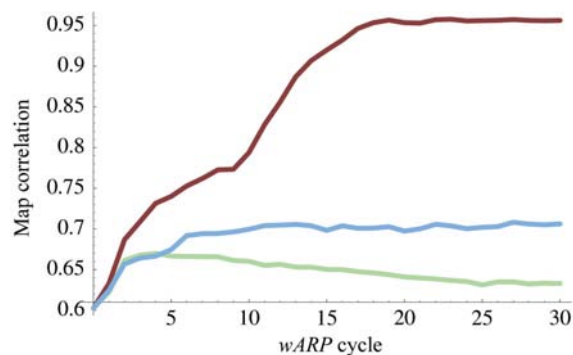
**3.2.1.  $\beta$ -Mannosidase, thioesterase I, PSCP and lysozyme (360°).** For most of the test data, significant differences existed between the results of automated model building when comparing the results of any target function with another. However, for the cases of  $\beta$ -mannosidase (Boraston *et al.*, 2003), thioesterase I (Devedjiev *et al.*, 2000), PSCP (*Pseudomonas* serine-carboxyl proteinase; Dauter *et al.*, 2001) and lysozyme collected over 360° (Weiss, 2001), we found no significant differences when using *ARP/wARP* with different target functions (except using MLHL with HL coefficients from *DM* as described below). The model building was always very rapid, with very complete models containing correct side chains (more than 90% of the model) obtained in less than ten building cycles. The results of the automated model building for all four test cases are summarized in Table 8.

Unexpected results were obtained using the MLHL target with Hendrickson–Lattman coefficients from *DM*. Of these four data sets, only the PSCP model was built with a similar number and quality of built residues as yielded by the other three approaches (Table 8), although a larger number of building cycles were required. The building provided only 191 residues for  $\beta$ -mannosidase, 107 residues for thioesterase and 34 for lysozyme, with significantly poorer r.m.s. errors and map

correlations. The complete models of all these proteins can be built using *DM* phases applying blurring factors with low scales as suggested by Murshudov (Pannu *et al.*, 1998). Blurring factors with scales lower than 0.4 were required to build the complete thioesterase model; higher blurring scales were sufficient to build the  $\beta$ -mannosidase and lysozyme models. However, when applying blurring factors with such a low scale, it is questionable whether these truly qualify as phase restraints. Since the models can be built without any prior phase information, the success of building with very low blurring scales is expected regardless of the prior phase distributions used.

**3.2.2. Lysozyme (270°).** The automated building of lysozyme from data collected over 270° (Weiss, 2001) was significantly different than building from the data collected from the same crystal over the whole sphere. Fewer S atoms were identified, with a greater r.m.s. error (most likely owing to the smaller redundancy of the data, leading to a smaller signal-to-noise ratio). This caused *CRANK* to produce a map for model building with a correlation over 15% less than in the case of 360° data, as can be seen in Table 1. The map turned out to be of insufficient quality to be traced by *ARP/wARP* at the beginning of the building process. Thus, building relied on the improvement of maps in the refinement process. With either of the current *REFMAC5* refinement targets, the improvement of maps stopped after a certain number of cycles, as shown in Fig. 4, and was not sufficient to build any residues. In contrast, rapid improvement was achieved by direct incorporation of phase information into the refinement target, yielding an almost complete lysozyme model. It is interesting to observe that the quality of the model built is outstanding and similar to that obtained from the 360° data (Tables 8 and 9), although the initial map and phase error before building were much poorer.

The fact that only the SAD function was able to build the model using this data is no longer true if a significantly better map is available before the initial model building. Changing the input parameters for substructure detection in *CRANK* for either *CRUNCH2* or *SHELXD* led to a more complete sulfur substructure and higher quality input maps. These maps could then have been traced using all the target functions, as in the case of the lysozyme 360° data.

**Figure 4**

The improvement of map correlation during the building process of lysozyme collected over 270°. The green curve stands for use of no prior phase information, the blue curve is for its indirect use and the red curve shows its direct use by SAD function.

**Table 9**

The results of lysozyme (270°) model building.

Lysozyme (270°)	Rice	MLHL	SAD
Map correlation	0.633	0.7062	0.9562
Residues built	0	0	120
Correct residues	0	0	120
R.m.s. error of model (Å)	—	—	0.068
Total run time (min)	39.2	40.53	92.92

**Table 10**

The results of ferredoxin model building.

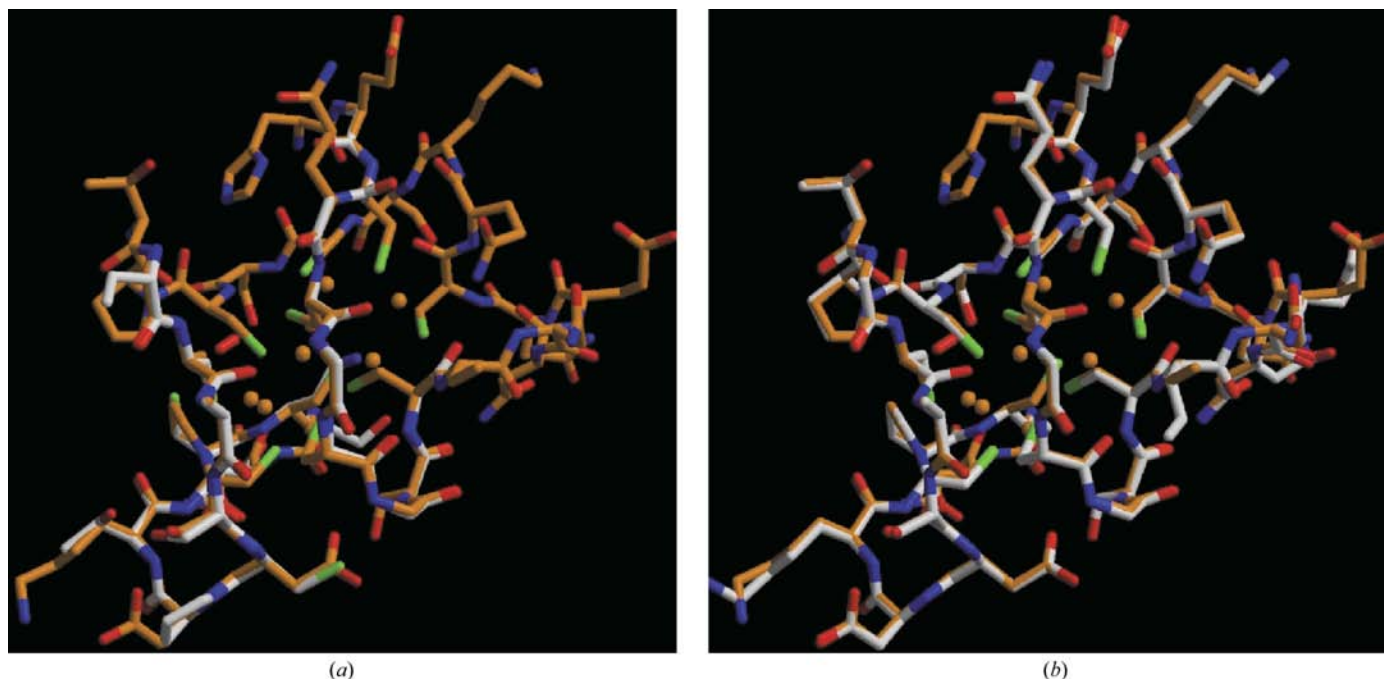
Ferredoxin	Rice	MLHL	SAD
Map correlation	0.8429	0.8799	0.9129
Residues built	40	44	48
Correct residues	40	44	48
R.m.s. error of model (Å)	0.05	0.045	0.045
Total run time (min)	20.99	23.23	25.18

**3.2.3. Lysozyme (180°).** The lower redundancy of lysozyme data collected over 180° (Weiss, 2001) caused a lower success rate of substructure determination. In this run, seven atoms were correctly identified, providing a map with 0.47 correlation to the final map. Although there were clear differences in the map improvement using the the automated building with different refinement targets, the maps were never of sufficient quality to automatically trace the model. As in the case of the lysozyme 270° data set, changing the input parameters in *CRANK* for substructure detection in either *CRUNCH2* or *SHELXD* led to a more complete sulfur substructure and higher quality input maps that could be traced.

**3.2.4. Ferredoxin.** The crystal structure of ferredoxin from *Clostridium acidurici* has been solved using data extending to

0.94 Å (Dauter *et al.*, 1997). Owing to a very strong anomalous signal from the Fe atoms from two [4Fe–4S] clusters, the map correlation after *BP3* was already very high (0.726). As *ARP/wARP* reported an error building with the default building algorithm, we used the other available algorithm for model building. Despite good maps and high resolution, all backbones built are less than 90% complete. *ARP/wARP* using the SAD functions can build several more residues than with MLHL, which is again several residues better than the Rice function (Table 10). The quality of all the models built is excellent; all C $\alpha$  atoms are correctly placed with an r.m.s. deviation of ~0.05 Å.

**3.2.5. Insulin.** Although the insulin data set was collected to 1.0 Å, with less than 75% Friedel pair completeness, it is possible to phase this protein by the SAD method using the anomalous signal of the two Zn atoms (Dauter *et al.*, 2002). However, in this single test case we were not able to phase the structure using only the automated procedures of *CRANK*. The complication was caused by the fact that both atoms lie on special positions: the threefold symmetric axis of the *R3* (hexagonal setting) space group. Although in *CRANK* runs *CRUNCH2* was able to find the approximate positions of both Zn atoms, they were displaced from the symmetry axis and their position could not be corrected in heavy-atom refinement. Thus, phasing under these conditions was unsuccessful, yielding a phase error of more than 80°. Since the atoms were close to special positions, we edited the scripts produced by *CRANK* and manually shifted the atoms to the exact special position and kept them fixed on the threefold axis, only allowing them to move along the axis in heavy-atom refinement, as is consistent with the location of the special position. This treatment provided phases with a phase error of 74.0°.



**Figure 5**

Models of thionin as built by *ARP/wARP* using the (a) Rice and (b) SAD functions, shown in white, superimposed on the final deposited model in orange. No model was built using the MLHL function.

**Table 11**

The results of insulin model building.

Insulin	Rice	MLHL	SAD
Map correlation	0.7119	0.8891	0.903
Residues built	0	93	95
Correct residues	0	93	94
R.m.s. error of model (Å)	—	0.123	0.117
Total run time (min)	111.9	209.99	212.91

Subsequently, *DM* was run from a script produced by *CRANK*, improving the phases slightly to 72.0°.

In spite of poor initial phases, *ARP/wARP* could trace almost the complete model by using prior phase information in the refinement. Using the Rice function with a default working set of reflections for Luzzati error-parameter refinement, the maps were greatly improved, having a map correlation of approximately 0.71 (Table 11). However, further refinement did not yield further improvement and no residues could be traced from the density map obtained. An interesting improvement was achieved using the Rice function by using the free set of reflections for Luzzati parameter refinement only. Using this approach, the map improvement continued for significantly longer, allowing over 60 residues to be built. Usually, the correct tracing of additional residues increases the map quality and allows tracing of more residues in the next building cycles. However, this trend was not observed in this case, as the map correlation was not improved after a great number of residues was traced.

Just ten building cycles were required to improve the phases by more than 40° and to build the majority of the model with the MLHL and SAD functions. The quality of models was further improved by correct docking of all residues in the second part of the building process. However, no model could be built using the MLHL function with HL coefficients from density modification. The refinement process with this approach was the worst of all and gave no significant increase in map correlation, which was 0.428 at the end of refinement. Although the refinement was improved by applying blurring factors, the improvement was not sufficient to build the model. The best results achieved using *DM* phases with any combination of blurring scale and blurring *B* factor were similar to those produced without any use of prior phase information.

**3.2.6. Thionein.** The truncated copper thionein from yeast (Calderone *et al.*, 2004) consists of eight Cu atoms surrounded by 36 protein residues (Fig. 5), with the anomalous signal being approximately 15% of the total scattering. The very high anomalous signal and small number of residues are similar to the ferredoxin test case and the results obtained are also very similar: *CRANK* was able to find all heavy atoms and produced an outstanding map with a correlation of 0.857 with the final map which was surprisingly difficult to trace. Using the Rice function for refinement, 21 residues were traced at the end of the building. However, the number of traced residues fluctuated rapidly between 0 and 25 during building (Fig. 6), which is very unusual and was not observed in any of the other test cases. No residues were traced if we used *ARP/wARP* with MLHL for refinement, although the map corre-

**Table 12**

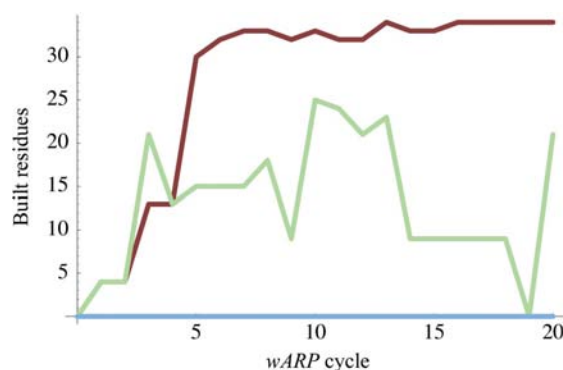
The results of thionein model building.

Thionein	Rice	MLHL	SAD
Map correlation	0.7156	0.7827	0.8736
Residues built	21	0	34
Correct residues	21	0	34
R.m.s. error of model (Å)	0.143	—	0.071
Total run time (min)	39.81	37.51	48.35

lation in the first cycles was comparable with that of the Rice function and did not significantly decrease during building. The reason for this behaviour is not understood, but the results are in accordance with those obtained in the original structure determination. *ARP/wARP* with the SAD function built a very good model of low r.m.s. error consisting of 34 residues (Table 12), all of which were correctly docked.

#### 4. Discussion and conclusions

From the test cases, the SAD refinement target is shown to extend the limits of phase quality and resolution needed compared with currently used functions. In difficult cases, such as GerE, 270° lysozyme, transhydrogenase, thionein or MutS, the direct use of this information in the more theoretically justified SAD function is shown to generate more complete models of higher quality. The majority of the 3 Å MutS model (Table 3) or ~2.75 Å bacteriophage T4 and GerE models (Tables 2 and 4) were successfully built by *ARP/wARP* using the SAD function. This is beyond the currently recommended *ARP/wARP* resolution limit of 2.6 Å (Morris *et al.*, 2004). However, if the resolution and the phase quality of the data are lower, automated building is likely to fail. The current limits of phase quality and resolution for successful model building using *ARP/wARP* with the SAD function can be estimated from Fig. 7. Besides extending the phase quality and resolution limits, the figure also demonstrates that the use of the SAD function can be surprisingly beneficial in some cases of high map correlation and resolution (the upper left part of the plots) that are not traced completely using current methods.

**Figure 6**

The number of built residues in the thionein test case during the *ARP/wARP* run for the Rice function (green), MLHL (blue) and SAD (red) functions.



Table 13 summarizes the results on the basis of the resolution of the data set.

The performance of the indirect use of prior phase information turned out to be strongly dependent on the source of the static phase distributions used. Taking the phase distributions from a phasing program usually provided better models than those obtained with the phase distributions obtained from density modification. However, this behaviour may be dependent on the programs used for phasing (*BP3*) and density modification (*DM*), as well as on the input parameters to these programs. In some cases, refinement with the phases from *DM* was improved when the blurring factors with low scales or high *B* factors were applied. However, in the insulin case, when the proper prior phase information was required to build the complete model, the blurring of phase distributions from density modification was not sufficient to build the model, although it could be built using the phase distributions from Hendrickson–Lattman coefficients from *BP3*. These results show that it is crucial that the phase distributions used by refinement with the MLHL target are as precise as possible and not biased. Therefore, the performance of the indirect use of prior phase information without the refinement of substructure parameters and/or Luzzati error parameters may be less than optimal in some cases. All these

**Table 13**

The mean value of the ratio of correctly built to the total number of residues in the final model from the 14 test cases that were successfully built as a function of resolution.

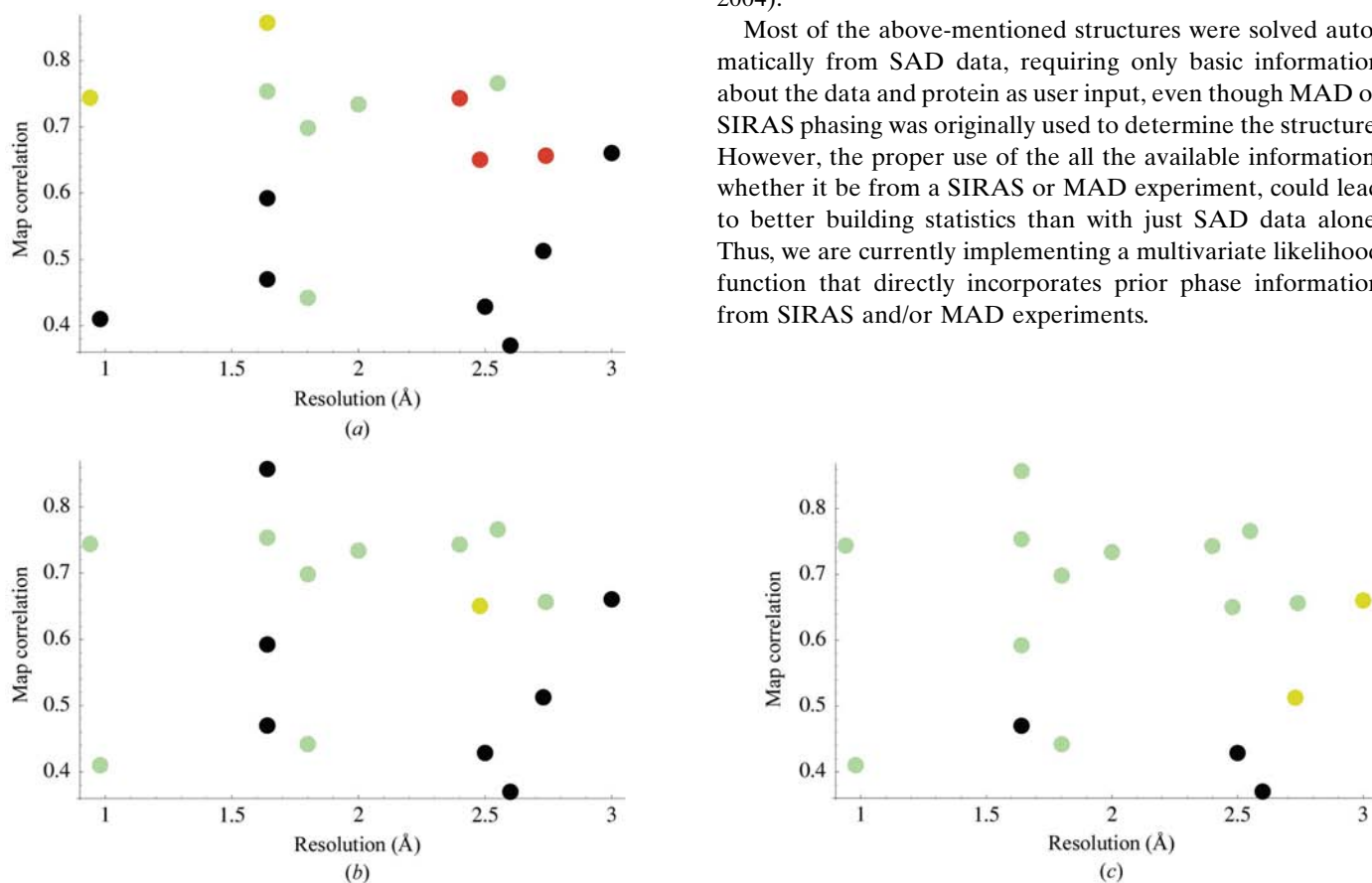
Correct part of model	Rice	MLHL	SAD
Resolution better than 2.4 Å (%)	63.9	68.8	93.9
Resolution 2.4 Å or lower (%)	27.2	50.7	73.1

problems are avoided if the SAD function is used, which directly reconstructs the phase information from the available information and allows for the refinement of substructure and error parameters.

The use of the SAD function did not significantly slow down the building-program runtime, which was approximately 10% longer than without any prior phase information, as judged from the cases in §3.2.1. The greater differences between building runtime with different functions in some other cases are caused by a different number of residues built and treated when different functions are used.

The SAD function is certainly not limited for use with any particular refinement-performing program, but can be incorporated straightforwardly for use with other autobuilding programs employing refinement, such as *RESOLVE* or recent developments using conditional dynamics (Scheres & Gros, 2004).

Most of the above-mentioned structures were solved automatically from SAD data, requiring only basic information about the data and protein as user input, even though MAD or SIRAS phasing was originally used to determine the structure. However, the proper use of the all the available information, whether it be from a SIRAS or MAD experiment, could lead to better building statistics than with just SAD data alone. Thus, we are currently implementing a multivariate likelihood function that directly incorporates prior phase information from SIRAS and/or MAD experiments.



**Figure 7**

The percentage of model built as a function of map correlation before building and the resolution of data for all tested data sets. The results obtained using (a) Rice, (b) MLHL and (c) SAD target for refinement are shown using ‘traffic light colours’: green represents an almost completely built model (80–100% of backbone residues built), yellow and red represent partially built models (50–80 and 20–50%, respectively) and black shows unsuccessful model building (0–20%).

We thank G. Murshudov, R. A. G. de Graaff and A. Perrakis for useful discussions, M. Weiss for his valuable insights into the manuscript, and A. Boraston, V. Calderone, C. Chen, Z. Dauter, Y. Devedijev, V. Ducros, E. Gordon, J. Ševčík, T. Sixma, E. Thomassen, M. Walsh, M. Weiss, S. White and their colleagues for making available the diffraction data used in the test cases. Funding for this work was provided by Leiden University and the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). The log files of above described runs and more information on the availability of the SAD function can be found at <http://www.bfsc.leidenuniv.nl/software/>.

## References

- Badger, J. (2003). *Acta Cryst.* **D59**, 823–827.
- Blessing, R. H. & Smith, G. D. (1999). *J. Appl. Cryst.* **32**, 664–670.
- Boraston, A. B., Revett, T. J., Boraston, C. M., Nurizzo, D. & Davies, G. J. (2003). *Structure*, **11**, 665–675.
- Bricogne, G. & Irwin, J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. J. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.
- Brünger, A. T. (1992). *Nature (London)*, **355**, 472–474.
- Brunger, A. T. (2005). *Structure*, **13**, 171–172.
- Calderone, V. (2004). *Acta Cryst.* **D60**, 2150–2155.
- Calderone, V., Dolderer, B., Hartmann, H. J., Echner, H., Luchinat, C., Del Bianco, C., Mangani, S. & Weser, U. (2004). *Proc. Natl Acad. Sci. USA*, **102**, 51–56.
- Chen, C. C., Zhang, H., Kim, A. D., Howard, A., Sheldrick, G. M., Mariano-Dunaway, D. & Herzberg, O. (2002). *Biochemistry*, **41**, 13162–13169.
- Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* **D50**, 760–763.
- Cowtan, K. (1999). *Acta Cryst.* **D55**, 1555–1567.
- Dauter, Z., Dauter, M. & Dodson, E. J. (2002). *Acta Cryst.* **D58**, 494–506.
- Dauter, Z., Li, M. & Wlodawer, A. (2001). *Acta Cryst.* **D57**, 239–249.
- Dauter, Z., Wilson, K. S., Sieker, L. C., Meyer, J. & Moulis, J. M. (1997). *Biochemistry*, **36**, 16065–16073.
- Devedijev, Y., Dauter, Z., Kuznetsov, S. R., Jones, T. L. & Derewenda, Z. S. (2000). *Structure Fold. Des.* **8**, 1137–1146.
- Ducros, V. M., Lewis, R. J., Verma, C. S., Dodson, E. J., Leonard, G., Turkenburg, J. P., Murshudov, G. N., Wilkinson, A. J. & Brannigan, J. A. (2001). *J. Mol. Biol.* **306**, 759–771.
- Engl, R. A. & Huber, R. (1991). *Acta Cryst.* **A47**, 392–400.
- Gordon, E. J., Leonard, G. A. & Zagalsky, P. F. (2001). *Acta Cryst.* **D57**, 1230–1237.
- Graaff, R. A. G. de, Hilge, M., van der Plas, J. L. & Abrahams, J. P. (2001). *Acta Cryst.* **D57**, 1857–1862.
- Hendrickson, W. A. & Lattman, E. E. (1970). *Acta Cryst.* **B26**, 136–143.
- Lamers, M. H., Perrakis, A., Enzlin, J. H., Winterwerp, H. H., de Wind, N. & Sixma, T. K. (2000). *Nature (London)*, **407**, 711–717.
- Levitt, D. G. (2002). *Acta Cryst.* **A58**, C28.
- Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
- Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A. & Lamzin, V. S. (2004). *J. Synchrotron Rad.* **11**, 56–59.
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* **D53**, 240–255.
- Ness, S. R., de Graaff, R. A. G., Abrahams, J. P. & Pannu, N. S. (2004). *Structure*, **12**, 1753–1761.
- Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst.* **D54**, 1285–1294.
- Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* **A52**, 659–668.
- Pannu, N. S. & Read, R. J. (2004). *Acta Cryst.* **D60**, 22–27.
- Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
- Scheres, S. H. W. & Gros, P. (2004). *Acta Cryst.* **D60**, 2202–2209.
- Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* **D58**, 1772–1779.
- Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). *Acta Cryst.* **D52**, 327–344.
- Skubák, P., Murshudov, G. N. & Pannu, N. S. (2004). *Acta Cryst.* **D60**, 2196–2201.
- Steiner, R. A., Lebedev, A. A. & Murshudov, G. N. (2003). *Acta Cryst.* **D59**, 2114–2124.
- Terwilliger, T. C. (2004). *J. Synchrotron Rad.* **11**, 49–52.
- Thomassen, E., Gielen, G., Schutz, M., Schoehn, G., Abrahams, J. P., Miller, S. & van Raaij, M. J. (2003). *J. Mol. Biol.* **331**, 361–373.
- Walsh, M. A., Otwinowski, Z., Perrakis, A., Anderson, P. M. & Joachimiak, A. (2000). *Structure*, **8**, 505–514.
- Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.
- White, S. A., Peake, S. J., McSweeney, S., Leonard, G., Cotton, N. P. & Jackson, J. B. (2000). *Structure Fold. Des.* **8**, 1–12.