Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

SPINE workshop on automated X-ray analysis: a progress report

M. Bahar,[a] C. Ballard,[b] S. X. Cohen,[c] K. D. Cowtan,[d] E. J. Dodson,[d] P. Emsley,[d] R. M. Esnouf,[a] R. Keegan,[b] V. Lamzin,[e] G. Langer,[e] V. Levdikov,[d] F. Long,[d] C. Meier,[a] A. Muller,[d] G. N. Murshudov,[d] A. Perrakis,[c] C. Siebold,[a] N. Stein,[b] M. G. W. Turkenburg,[d] A. A. Vagin,[d] M. Winn,[b] G. Winter[b] and K. S. Wilson[d]*

[a]The Division of Structural Biology and Oxford Protein Production Facility, The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, England, [b]CCLRC, Daresbury Laboratory, Warrington WA4 4AD, England, [c]Netherlands Cancer Institute, Molecular Carcinogenesis Department, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands, [d]York Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, York YO10 5YW, England, and [e]EMBL Hamburg, Building 25A, DESY, Notkestrasse 85, 22603 Hamburg, Germany

Correspondence e-mail: keith@ysbl.york.ac.uk

Received 22 February 2006
Accepted 14 August 2006

The Structural Proteomics In Europe (SPINE) consortium contained a workpackage to address the automated X-ray analysis of macromolecules. The aim of this workpackage was to increase the throughput of three-dimensional structures while maintaining the high quality of conventional analyses. SPINE was able to bring together developers of software with users from the partner laboratories. Here, the results of a workshop organized by the consortium to evaluate software developed in the member laboratories against a set of bacterial targets are described. The major emphasis was on molecular-replacement suites, where automation was most advanced. Data processing and analysis, use of experimental phases and model construction were also addressed, albeit at a lower level.

1. Introduction

The Structural Proteomics In Europe (SPINE) project, initiated in 2002, aimed to introduce new technologies and approaches to the complete set of processes required to determine the three-dimensional structures of biomedically relevant proteins. It was envisaged that the majority of these structures would be determined using X-ray crystallography and a distinct section of the programme was devoted to this method. The stated aims of this work package were

> to address the problems of automated X-ray analysis of macromolecules. To achieve throughput in keeping with genome sequencing projects, macromolecular crystallography (MX) procedures must be streamlined, and work in a number of laboratories in Europe, including several SPINE partners, is directly addressing this. Scripting will link the various stages, and better algorithms will be formulated in key areas such as molecular replacement (MR), experimental phasing, automated generation of atomic models, molecular graphics and quality assessment. The software will ensure that high quality will accompany high throughput.

(from the SPINE Contract QLG2-CT-2002-00988).

Within the SPINE project, most of the resources were devoted to major bottlenecks for structural biology, namely in protein cloning, overexpression and crystallization. Hence, SPINE had only limited resources to contribute to the development of high-throughput crystallographic computing, but by bringing together major users and providers of code it was in a good position to gain access and provide some input to developments. This problem is being addressed worldwide. It is clear that contacts and coordination are essential to optimize the output of developers and that such contacts must be maintained. Early in the programme SPINE held two workshops to discuss automation, attended by people both within the project and from associated groups. This report

Table 1
Targets used during the workshop.

N_{Res} is the number of residues per molecule. N_{Mol} is the expected number of molecules in the asymmetric unit. MR, molecular replacement. SAD, single-wavelength anomalous dispersion. MAD, multiple-wavelength anomalous dispersion.

ID	N_{Res}	N_{Mol}	Space group	Resolution (Å)	Method	PDB code
York						
BA0288	161	8	$C2$	1.80	MR	1xmp
BA0296	346	2	$P4_1$	2.31	MR	
BA0592	377	6	$C2$	2.84	MR	
BA1071	311	1	$P2_12_12_1$	2.60	MR	
BA1483	235	8	$P2_12_12_1$	2.24	MR	1xe3
BA1563	282	2	$P2_1$	2.20	MR	
BA3935_1	292	4	$P2_12_12_1$	1.94	MR	
BA3935_2	292	4	$P2_12_12_1$	2.23	MR	
BA4499	203	2	$P2_1$	1.80	MR	
BA4508	298	1	$C2$	2.57	MR	
BA5696	208	2	$P2_1$	1.80	MR	
BA5705	327	2	$P4$	1.80	MR	
BSAppA	543	1	$P2_12_12_1$	2.28	MR	1xoc
BSYloQ	298	1	$P4_32_12$	2.51	MR	1t9h
CJ0982	292	2	$C2$	2.00	MR	
Peb3	230	2	$P2_12_12_1$	1.65	MR	
SiaP	306	2	$P2_12_12$	2.60	MAD	
Oxford						
OPPF651	294	2	$P2$	2.40	MR	
OPPF1294	193	2	$P4_12_12$	2.70	MAD	
OPPF1311	255	4	$P6_122$	2.72	SAD	
OPPF1314	200	2	$P1$	1.50	MR	
OPPF2088	139	1	$P3_121$	2.20	MAD	
OPPF2153	222	4	$P2_1$	2.70	MAD	
OPPF2245	229	2	$P2_1$	3.30	MR	

summarizes the activities at a third workshop where the current methodology was tested against targets selected from SPINE partner laboratories in Oxford and York. It does not describe in detail the software being developed or the structures of the individual targets, as these will be published elsewhere.

The SPINE project has tried to follow the traditional CCP4 (Collaborative Computational Project, Number 4, 1994) approach of linking contributions from a number of sources [such as *ARP*/*wARP* (Perrakis *et al.*, 1999) and *SHELX* (Schneider & Sheldrick, 2002)] to form a set of modular tools. This requires agreement on exchange protocols, which can be hard to establish, but will result in more robust and flexible software which can be easily upgraded in the years to come. The aim of the workshop was to assess progress towards this within the SPINE team and its associates.

2. The target data sets

23 data sets were selected from a group of bacterial (mainly *Bacillus anthracis* and *Campylobacter jejuni*; Alzari *et al.*, 2006) targets under study in Oxford and York (Table 1). Merged structure factors and the amino-acid sequence data were the basis for most of the activity; however, for a subset of targets raw images were made available for assessment of an automated processing protocol. In the event, this effort was essentially restricted to a single problem data set (OPPF1314).

The basic selection parameters were that target proteins should be less than 50 kDa, not part of a complex, contain no signal peptides and have no transmembrane regions. Most were candidates for MR and were straightforward targets for the subsequent application of the *ARP*/*wARP* packages of automated electron-density interpretation.

During the workshop two structures were examined in greater detail to pinpoint problems in the structure-automation pipelines. These were OPPF1314 (Oxford) and SiaP (York).

2.1. OPPF1314

OPPF1314 data were used both for testing the data processing and analysis pipeline and for the automated model-building procedures. OPPF1314 is a 5-formyltetrahydrofolate cycloligase (BA4489) and has a molecular weight of 22.3 kDa (292 residues). The protein catalyses the ATP-dependent formation of 5,10-methenyltetrahydrofolate from 5-formyltetrahydrofolate (folinic acid; Huennekens *et al.*, 1984).

The full details of the structure determination will be described elsewhere (Meier *et al.*, 2006). Briefly, crystals were obtained from cocrystallizations of OPPF1314 with the substrates ATP and 5-formyltetrahydrofolate and data extending to 1.5 Å resolution were measured on ID14EH1 at the ESRF from a crystal belonging to space group $P1$ containing two molecules in the crystallographic asymmetric unit. Data were acquired in a high-resolution pass (in which many of the low-resolution reflections were overloaded) followed by a low-resolution pass. The diffraction showed a high degree of mosaicity. Data reduction with *DENZO*/*SCALEPACK* (Otwinowski & Minor, 1997) prior to the workshop gave an apparently reasonable merged data set, but it had proved impossible to phase this satisfactorily by MR using either a medium-resolution model structure obtained previously in a different space group or a related structure (PDB code 1ydm) with 47% sequence identity.

2.2. SiaP

The SiaP protein, a candidate for MAD phasing (Table 1), was used during the workshop to test using experimental phasing to kick-start automated model building. One structure in the PDB, 1k7k, had some (25%) sequence identity, but only over a third of the molecule. MAD data sets had been collected for SeMet-labelled protein at three wavelengths (0.97907, 0.90778 and 0.97920 Å) on BM14, the UK MAD beamline at the ESRF. The SeMet crystals diffracted to 2.6 Å resolution with high merging R factors in the outer ranges and belonged to space group $P2_12_12$, with two molecules in the crystallographic asymmetric unit. Although the data between 2.9 and 2.6 Å resolution were especially weak [$I/\sigma(I) = 1.5$ in the outer shell], they proved essential for structure solution.

16 Se atoms were expected in the asymmetric unit and *SHELXC* and *SHELXD* (Schneider & Sheldrick, 2002) found 14 sites. Phases had been calculated with *SHELXE* but automated construction of a model using *REFMAC-ARP*/*wARP* from these phases had failed: the procedure built many

short disconnected peptides with no side chains docked. The same heavy-atom solution was used with greater success in *RESOLVE* (Terwilliger, 2003), which built 468 residues in ~44 chains, but with only 75 side chains docked. This model was in turn fed to *REFMAC-ARP/wARP* with the 'default' options, but the procedure dismembered the model rather than adding additional features. The situation was improved by using the *RESOLVE* model with phase restraints imposed during *REFMAC* refinement cycles. For this purpose, the reference phase set was based on the original *SHELXE* phases from the selenium substructure, improved by solvent flattening to give a single 'best' phase estimate with an associated figure of merit. This gave a better result, with R converging at around 30.4% ($R_{\rm free}$ was not used) for a model with 260 backbone residues in 27 chains and 65 side chains docked. This again reflected some degradation of the *RESOLVE* model. All this work was carried out prior to the workshop.

3. Software suites and developers involved

Extensive use was made of *CCP*4 modules and utilities. The *XIA-DPA* system was used for the processing of X-ray images and their subsequent analysis and quality assessment. *XIA-DPA* provides wrappers for software packages including *LABELIT* (Sauter *et al.*, 2004), *MOSFLM* (Leslie, 1999), *XDS* and *XSCALE* (Kabsch, 1993), *SCALA* (Evans, 1993), *TRUNCATE* (French & Wilson, 1978) and *SFCHECK* (Vaguine *et al.*, 1999).

Sequence analysis and putative MR model identification used well established software packages available on the web, particularly *MSDtarget* and *MSDfold* from EBI (http://www.ebi.ac.uk/msd) and *BLAST* (http://www.ncbi.nlm.nih.gov/BLAST/).

MR was carried out by three teams. All used established core software, but different scripted protocols. One group (RK and MW) has developed *MrBUMP*, which uses existing web servers (described below) to select multiple models, modifies them using *CHAINSAW* (Stein, unpublished work), *MOLREP* (Vagin & Teplyakov, 1997) or *PDBCLIP* (a local utility), followed by application of *MOLREP* or *Phaser* (McCoy *et al.*, 2005) for the MR search. Models were assessed using *REFMAC* (Murshudov *et al.*, 1997). A second group (NS and CB) has developed *AutoAMoRe*, incorporating *CHAINSAW* and *AMoRe* (Navaza, 1994). The third group (GNM, FL and AAV) has developed a package *BALBES* that utilizes a pre-constructed database for model selection, *SFCHECK* for data-quality analysis, *MOLREP* for model preparation and molecular replacement, and *REFMAC* for initial refinement and final quality assessment. Sequence analysis and model identification use programs and procedures under development by the authors (Murshudov, private communication).

With the exception of SiaP, there was little work on experimental phasing, as the automated pipelines are at an earlier stage of development. However, several of the Oxford structures requiring experimental phasing had been solved prior to the workshop using *SHELXD* and *SHELXE* (Schneider & Sheldrick, 2002).

Model construction and rebuilding for both MR and experimentally phased maps was primarily based on the *REFMAC–ARP/wARP* pipeline (SC, GL; Perrakis *et al.*, 1999). Maps were visualized using *Coot* (Emsley & Cowtan, 2004) and development versions of *Pirate* (Cowtan, 2000) and *Buccaneer* (Cowtan, 2001) were tested. Ligand fitting was attempted using both *ARP/wARP* and *Coot*.

4. Data processing

4.1. Data integration

XIA-DPA was applied to targets where images were available. *XIA-DPA* is an automated wrapper for existing data-processing and analysis software. It aims to combine independently developed functionality in a modular manner so that it will be straightforward to replace individual functions. *XIA-DPA* incorporates these into an expert system capable of making decisions about how to process data without user intervention.

The user interface to XIA-DPA is simple: the filename of an image is sufficient to initiate data processing tasks for either two-dimensional or three-dimensional integration: `xia-autoprocess-2d /path/to/data/set/foo_1_001.img` or `xia-autoprocess-3d /path/to/data/set/foo_1_001.img`.

The current software distribution uses *LABELIT* to perform the autoindexing followed by two-dimensional integration with *MOSFLM* or three-dimensional integration using *XDS*. *POINTLESS* (Evans, 2006) is used to select the most likely point group. Scaling and merging can be performed with *SCALA* and *TRUNCATE* or by *XSCALE*. Images are processed to provide reduced and scaled, merged and unmerged reflection files in the commonly useful formats (MTZ with I^+, I^-, I, F^+, F^-, F and *SCALEPACK*) along with estimates of the resolution and lists of possible space groups from an analysis of the systematic absences. The objective is to provide the initial stages of a 'data-to-structure' pipeline to generate machine-readable information to be used in the subsequent steps of structure solution.

4.2. Data analysis and quality assessment

It was realised very early in the workshop that the experimental data as provided often did not carry all of the necessary crystal information in a form accessible by user or computer. Some of the information such as wavelength should be recorded in the reflection-file header. We suggest that a simple solution would be to define an accepted exchange format and record tagged information conforming to this format in an exchange file.

Decisions to be taken in automatic procedures fall into four categories.

(i) Sample parameters, *e.g.* the sequence, molecular weight and expected numbers of 'heavy' atoms.

(ii) Details of the X-ray experiment: direct parameters such as wavelength, beamline and temperature, derived parameters

including unit cell, point group, the likely number of molecules in the crystallographic asymmetric unit and the presence of any non-crystallographic translational operator and quality indicators including nominal resolution, estimated *B* factor and anisotropy plus completeness at both low and high resolution (the former being important for MR), multiplicity, $I/\sigma(I)$ and merging *R* factor, all as functions of resolution.

(iii) Intensity statistics to be tested against expectation values including cumulative intensity distributions and moments. These are sensitive indicators of problems in the experiment such as twinning or local errors in the processing such as saturation of substantial numbers of low-resolution terms (Fig. 1).

(iv) The identification of special features of the crystal such as pseudo-symmetry or potential alternative indexing.

This list is certainly not complete and requires agreement amongst the community on a formal definition of needs.

Most of the necessary information is already available in the output from various programs, but is not yet encoded into an accepted exchange file. The data sets used for the workshop were evaluated retrospectively using *TRUNCATE* and *SFCHECK* (Table 2).

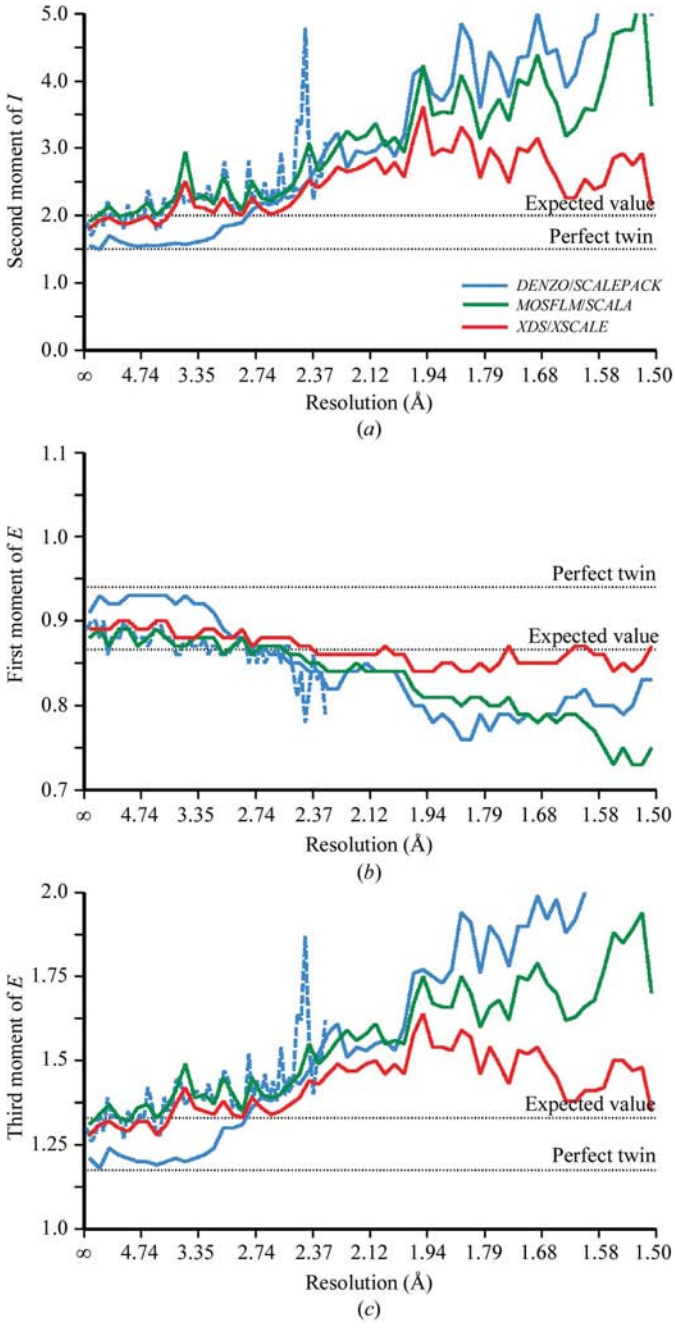

Figure 1
Results of data reduction for OPPF1314 images using different data-processing packages (see Table 3) analysed with *TRUNCATE*. (*a*) Second moments of *I* in resolution shells. (*b*) First moments of *E* in resolution shells. (*c*) Third moments of *E* in resolution shells. In all panels the combined data sets are shown with solid lines (blue, *DENZO*/*SCALEPACK*; green, *MOSFLM*/*SCALA*; red, *XDS*/*XSCALE*), while results for the low-resolution pass processed with *DENZO*/*SCALEPACK* are shown by blue dashed lines.

4.3. Test applications

4.3.1. OPPF1314. Analysis of the merged OPPF1314 data had shown an unusual distribution of reflection intensities, in particular in the low-resolution range (Fig. 1). The original images were reprocessed during the workshop with *XIA-DPA*, trying both the two-dimensional and three-dimensional options. The two-dimensional pipeline (*i.e.* using *MOSFLM* and *SCALA*) produced similar results to those from *DENZO*/*SCALEPACK*; again the intensity statistics were unusual. However, the three-dimensional pipeline (*XDS* and *XSCALE*) gave a well merged data set which led to a successful structure solution in a straightforward manner. More detailed analysis and careful reprocessing with *MOSFLM*/*SCALA* subsequent to the workshop suggested that the first failures were a consequence of poor relative scaling of the low- and high-resolution passes and that the high mosaic spread proved difficult to handle with the current two-dimensional software. The scaling *R* factor between the *XDS* and reprocessed *MOSFLM* amplitudes is ∼4% to 2.3 Å, rising to 16% at 1.5 Å. The quality assessment (Tables 2 and 3) suggested that the two-dimensional integration with *MOSFLM* was unsatisfactory in the higher resolution ranges. *SFCHECK* showed that the high mosaicity reduced the completeness of the two-dimensional data set in certain zones.

4.3.2. Other targets. Diffraction images for a further six targets were reprocessed at the workshop using *XIA-DPA* using the two-dimensional option (Table 4). The agreement with the data provided for the workshop for BA0592 appears to be satisfactory and the quality assessment for the other data sets was acceptable (Table 2). BA0296 illustrates a situation where rapid automated data processing should have been performed during data collection to select the best strategy. Autoindexing was satisfied by a cubic crystal class. However, subsequent analysis showed that the point group was tetragonal. This unfortunately resulted in a rather incomplete data set. This workshop greatly stimulated the further development of the *XIA-DPA* pipeline.

4.4. Discussion; lessons for automated data processing

The data-processing and analysis step is critical for the structure-solution pipeline. The *XIA-DPA* pipeline performs adequately for many data sets, but it is essential that it flags aberrant cases and alerts the crystallographer in the more challenging cases, such as OPPF1314. In the case of BA0296 the information derived from initial indexing was later found to be incorrect and this highlights the need for easy and reliable determination of the point group from limited data during data collection.

A final assessment of data quality at the end of an automated procedure is good practice and should always be carried out. Automated procedures have two features to offer, reproducibility and standardization, which should allow for more objective summary statistics. They may also perform the tedious transformations between different packages, enabling all statistics to be calculated by the same program and hence be more comparable.

Finally, in any automation effort, the success of the pipeline depends on the cumulative success of the individual steps. As data-processing programs become more reliable and sophisticated, the overall success rate of the pipeline should improve.

Table 2
Data-quality evaluation based on the output of the *TRUNCATE* and *SFCHECK* programs.

$N(I)$ is the cumulative intensity distribution from *TRUNCATE*, flagged as 'OK', 'sig' (sigmoidal, indicating twinning) or 'odd'. $B_{\rm est}$ are the overall B-factor estimates by each program (*TRUNCATE* and *SFCHECK*, respectively). Aniso represents the degree of anisotropy in the merged data: this is described in *SFCHECK* by the three eigenvalues of the overall anisotropic scaling ellipsoid. An off-origin peak in the native Patterson synthesis indicates the presence of pseudo-translational symmetry (PseudoT) and *SFCHECK* flags this. The likelihood of twinning is estimated by *SFCHECK*, after taking account of pseudo-translation and anisotropy. Substantial deviation from linearity of the Wilson plot is flagged by *TRUNCATE*.

ID	$N(I)$	$B_{\rm est}$ (Å^2)	Aniso	PseudoT	Twin	Wilson	Comments
York							
BA0288	OK	18/27	N				Low-res. weak
BA0296	Sig	58/66	N		T		High-res. incomplete
BA0592	Odd	71/69	Y				95% complete
BA1071	OK	54/55	Y				95% complete, low-res. weak
BA1483	OK	27/35	Y?				Low-res. weak, high-res. incomplete
BA1563	OK	35/44	Y	Y			Low-res. weak, high-res. incomplete
BA3935_1	OK	20/29	N			Nonlinear	Low-res. weak
BA3935_2	OK	42/49	Y				Low-res. weak, better than previous
BA4499	Sig	25/34	Y				
BA4508	Odd	53/56	Y				94% complete, poor strategy?
BA5696	OK	27/36	Y				
BA5705	Sig	24/32	Y		T?		
BSAppA	OK	32/39	Y				Low-res. weak
BSYloQ	OK	44/66	Y				
CJ0982	Sig	2738	Y			Nonlinear	Incomplete and very anisotropic
Peb3	OK	17/24	Y			Odd	Low-res. odd, probably OK
SiaP	OK	34/42	Y				Low-res. weak
Oxford							
OPPF651	Wild	41/47	Y	Y			Low-res. weak
OPPF1294	Sig	105/105	N		T?		Very weak overall
OPPF1311	Odd	72/70	Y	Y			An odd distribution
OPPF1314	OK	37/47	Y				See text; incomplete, problems in merging
OPPF2088	Sig	42/51	Y		T?		
OPPF2153	Wild	49/50	Y	Y	?	Nonlinear	Data very weak
OPPF2245	OK	76/54	Y		T?	Nonlinear	Low resolution
XIA-DPA processing							
OPPF1314-2D	Odd	24/19	Y			OK	5% of reflections rejected
OPPF1314-3D	OK	22/19	Y			OK	Fewer reflections rejected, data quality better
BA0296	Sig	49/53	N	N	T		Missing wedge of data, incomplete
BA0592	OK	69/58	Y	N			Very anisotropic
BA1071	Sig	34/44	Y	N	T?		Indications of twinning
BA2236	Sig	26/39	Y	N	T?		Wilson poor for high res.
BA4525	OK	21/33	Y	N			Data missing from 2.7 Å
BA5505	OK	57/66	N	N			Ice ring

5. Molecular-replacement pipelines

Of all the automated structure-solution pipelines, those addressing MR are the most advanced at present and were the core activity at the workshop. Each pipeline has five basic tasks to address.

(i) Is there a suitable model structure in the PDB? This requires the use of sequence-matching tools, such as *BLAST* and *FASTA* (Brenner *et al.*, 1998), which can scan the PDB (Bernstein *et al.*, 1977; Berman *et al.*, 2000) for homologous structures to use as templates. The results of these alignments have to be viewed critically. For example, sequence similarity over a short segment is likely to be of limited value; what is required is a fit over the whole length of the sequence or at least over an extensive fraction such as a domain. Decisions must be made about the optimum search unit based on knowledge of the biology: this ranges from identifying possible oligomers for the search to breaking down a single chain into individual domains. Models can often be usefully modified based on the sequence alignment, *e.g.* the deletion of gaps and pruning of side chains. A practical advantage of such pruning is to introduce the 'correct' residue numbering and naming into the model, which is useful during rebuilding.

(ii) What is the information content of the X-ray data? The reduction and analysis of the X-ray images should both provide a set of essential information and make it available to the user/program/pipeline (see §4.2).

(iii) Does the MR search indicate a satisfactory solution? Each program provides a scoring function for potential solutions and a contrast between the best and the others is required. In some cases, the crystal space group is ambiguous; the automated translational searches must then cover all

Table 3
Data sets obtained by processing of the ESRF ID14EH1 images obtained from OPPF1314 cocrystals (space group *P*1) using different data-processing packages.

Images for high- and low-resolution (LOW) passes were collected and these were also merged to form combined (CMB) data sets. Data were processed using *DENZO*/*SCALEPACK* (DEN), *MOSFLM*/*SCALA* (MOS) and *XDS*/*XSCALE* (XDS). Values in parentheses are for the outer (highest resolution) data shells, *i.e.* 2.38–2.30 Å for the low-resolution pass (note that the low completeness results from processing into the corners of a square detector) and 1.55–1.50 Å for the combined low- and high-resolution passes.

	DEN-LOW	MOS-LOW	XDS-LOW	DEN-CMB	MOS-CMB	XDS-CMB
Resolution limit (Å)	2.30	2.30	2.30	1.50	1.50	1.50
Unique reflections	11585	10987	11455	59140	56913	60271
Completeness (%)	63.5 (11.4)	62.1 (13.8)	64.2 (14.4)	93.0 (90.2)	90.5 (89.1)	95.2 (93.7)
Multiplicity	2.4 (2.2)	2.4 (2.3)	2.4 (2.3)	2.1 (1.7)	2.4 (1.9)	2.4 (2.0)
$I/\sigma(I)$	20.6 (5.5)	18.3 (4.0)	17.2	20.9 (1.1)	8.7 (1.3)	11.1 (2.1)
$R_{\rm merge}$† (%)	4.0 (17.5)	3.9 (17.7)	3.2 (15.5)	5.3 (83.5)	5.7 (63.9)	4.3 (46.7)

† $R_{\rm merge} = \sum I - \langle I \rangle / \sum \langle I \rangle$.

Table 4
Summary of other data sets processed with *XIA-DPA*.

For BA0592 a previously processed data set had also been supplied to the workshop (Table 1) and $R_{\rm fac}$ gives the scaling *R* factor between this data set and the *XIA-DPA* processing. The suggested space group is that predicted from systematic absence analysis.

Protein	Resolution (Å)	Space group	$R_{\rm merge}$	Completeness (%)	Multiplicity	$I/\sigma(I)$	$R_{\rm fac}$
BA0296	2.5	$P4_1/4_3$	11.3	81.7 (70.9)	2.5	7.7 (2.3)	
BA0592	3.0	*C*2	20.1	99.5 (99.3)	4.0	8.4 (1.9)	0.087
BA1071	2.0	$P2_12_12_1$	14.0	96.7 (81.5)	4.1	6.6 (1.7)	
BA2236	2.1	$P2_12_12_1$	15.1	88.1 (56.8)	6.0	7.6 (1.8)	
BA4525	2.0	*C*2	7.8	69.1 (17.4)	3.5	13.0 (3.0)	
BA5055	2.4	$P3_n21$	14.8	96.0 (84.6)	7.3	10.2 (1.0)	

possibilities and a clear result in just one of these is also an indicator of likely success.

(iv) Is the solution likely to be correct? Firstly, for a correct solution the model molecule must not clash too severely with symmetry-related copies in the unit cell. Secondly, preliminary automated refinement should reduce both *R* and $R_{\rm free}$. More sophisticated tests can address such questions as to whether the solution makes good biological sense (*e.g.* residues in suitable electrostatic environments, sensible lattice contacts), but these are harder to automate.

(v) Can the resulting model be satisfactorily rebuilt? The best criterion for a suitable MR solution remains the quality of the resulting electron density. If new correct features are visible in the map and incorrect features of the model are not, then it can be considered a solution. This is discussed in §7.

5.1. Individual pipelines

In this section, we report on the results obtained by the three teams. All teams used some of the same tools for the various tasks. *CHAINSAW* (next release of the *CCP*4 suite) is a new utility developed for manipulating models. It examines the sequence alignment between target and template provided in a standard format and modifies the template PDB file by pruning non-conserved side chains back to the γ atom while leaving conserved residues unchanged. Atom and residue names and numbers are matched to those in the target. The result is what Schwarzenbacher *et al.* (2004) have termed a 'mixed model', since more atoms are preserved than in a polyalanine model, but parts of the model which are unlikely to be present in the crystal structure, and thus may degrade the signal, are pruned.

MOLREP also contains many model-preparation tools. It aligns a given sequence with the model and prepares a truncated model. It can also carry out locked translation searches using a given non-crystallographic symmetry transformation. *AMoRe* is a well established package which separates the rotation, translation and rigid-body refinement modules, allowing great flexibility in tailoring protocol to problem. *Phaser* has a sophisticated scoring scheme and also recalculates the orientation search for multiple molecules, taking account of the contribution from any model positioned previously. *REFMAC*5 refinement was used to assess the quality of the solutions; if both the *R* and $R_{\rm free}$ fell, then a solution was judged to be substantially correct.

5.1.1. The *MrBUMP* package. The *MrBUMP* package has been developed as part of the eHTPX (http://www.e-htpx.ac.uk/) and CCP4 projects: eHTPX provided extensive computing resources in the form of an 18-CPU cluster accessed *via* an eHTPX web service. This mode is particularly valuable for marginal cases with low sequence homology where a parallel approach can be used with a range of putative trial models and methodologies being investigated contemporaneously. Alternatively, the *MrBUMP* package can be customized to run on a desktop and it was tested in both modes during the workshop. It is designed to make use of web-accessible databases of sequences and structures, rather than relying on local databases. This guarantees that the information is up to date, but has the drawback that queries are submitted across a public network and may be slower. Since the workshop, *MrBUMP* has been extended to allow this search to be performed locally.

In brief, the *MrBUMP* pipeline comprises the following steps: the properties of the target are generated from the reflection and sequence files provided (for example, the expected number of molecules in the asymmetric unit) and then a *FASTA* search of the current PDB is made to generate a list of possible homologous structures, which are then downloaded. For each, the PQS server (http://pqs.ebi.ac.uk/) is queried to ascertain whether the model exists as a multimer. If so, and if the multimer could fit into the target unit cell, it is added to the list of templates. A recent addition, not available for the tests described here, is to add domains identified by *SCOP* (Murzin *et al.*, 1995) to the list of templates.

Table 5
Summary of results from *MrBUMP*.

Results in italics were obtained after the workshop, otherwise results are as obtained from 'live' runs at the workshop. The columns are as follows. Protein ID, identification tag as used in Table 1; N_{mol}, number of molecules expected in the asymmetric unit; Model, PDB code and chain ID of the template structure; % ID, percentage sequence identity; Pruned, trial model-generation method used, see text; MR prog, program used for molecular replacement; initial R/Rf, *R* factor and R_{free} reported for cycle 0 of *REFMAC*; final R/Rf, *R* factor and R_{free} reported for last cycle of *REFMAC*; Success, whether probable solution (Y), possible solution (P) or no solution (N); Rebuilt, whether rebuild attempted in *ARP/wARP* and if so whether successful.

ID	N_{mol}	Model	% ID	Pruned	MR Prog	Initial R/Rf	Final R/Rf	Success	Rebuilt
BA0288	8	1u11_*A*	66	*MOLREP*	*MOLREP*	0.45/0.45	0.30/0.33	Y	Y
BA0288	8	1u11_0	66	Multimer	*MOLREP*	0.46/0.45	0.32/0.36	Y	–
BA0592	6	1pjb_*A*	54	*PDBCLIP*	*MOLREP*	0.44/0.43	0.32/0.39	Y	N
BA1071	1	1c9e_*A*	73	*CHAINSAW*	*MOLREP*	0.43/0.44	0.29/0.36	Y	Y
BA1483	*6*	*1pr1_A*	*57*	*CHAINSAW*	*MOLREP*	*0.51/0.52*	*0.35/0.41*	*Y*	*Y*
BA1563	2	1ufv_*B*	49	*CHAINSAW*	*MOLREP*	0.51/0.50	0.40/0.47	P	P
BA3935_1	4	1dhp_*A*	42	*MOLREP*	*MOLREP*	0.52/0.52	0.37/0.41	Y	Y
BA3935_2	*4*	*1s5t_B*	*42*	*MOLREP*	*MOLREP*	*0.49/0.49*	*0.33/0.39*	*Y*	*Y*
BA4499	2	1jr9_0	71	Multimer	*MOLREP*	0.44/0.46	0.28/0.33	Y	Y
BA4508	1	1qtw_*A*	32	*CHAINSAW*	*MOLREP*	0.52/0.53	0.40/0.48	P	N
BA4508	1	1qtw_*A*	32	*CHAINSAW*	*Phaser*	0.52/0.51	0.40/0.48	P	–
BA5696	2	1jr9_0	56	Multimer	*MOLREP*	0.46/0.47	0.29/0.33	Y	Y
BA5705	2	1vrd_*B*	35	*CHAINSAW*	*MOLREP*	0.55/0.56	0.43/0.49	P	–
BA5705	2	1vrd_B	35	*CHAINSAW*	*Phaser*	0.54/0.55	0.42/0.45	Y	Y
BSAppa	1	1dpe	28	*CHAINSAW*	*MOLREP*	0.57/0.57	0.56/0.55	P	–
BSYloQ	1	1u0l_*A*	40	*CHAINSAW*	*MOLREP*	0.55/0.52	0.41/0.48	P	–
BSYloQ	1	1u0l_*A*	40	*CHAINSAW*	*Phaser*	0.54/0.50	0.42/0.49	P	–
CJ0982	2	1qok_*A*	44	–	–	–	–	N	–
OPPF651	2	1v6s_*A*	51	*MOLREP*	*MOLREP*	0.66/0.65	0.42/0.50	Y	N
OPPF651	2	1php	77	*MOLREP*	*MOLREP*	0.65/0.64	0.33/0.39	Y	Y
OPPF1314	2	1ydm_*C*	47	*CHAINSAW*	*MOLREP*	0.53/0.52	0.47/0.48	P	P
OPPF2245	2	1lfp_A	37	*CHAINSAW*	*MOLREP*	0.54/0.55	0.45/0.52	P	N

The next step is to generate trial models from the templates. Currently, three methods are used. Firstly, in the *PDBCLIP* method the raw template coordinates are used, after various tidying steps such as removal of waters, removal of alternative conformations *etc.* Secondly, the alignment and model-improvement method of *MOLREP* is used. Thirdly, the *CCP*4 program *CHAINSAW* described above is used to provide a mixed model.

The top models are passed to *MOLREP* for the first attempt at MR. If *MOLREP* produces a solution (irrespective of score), the positioned model is passed to *REFMAC* for 30 cycles of restrained refinement. The free *R* factor is used as the criterion for success. If the R_{free} drops significantly, the script stops and reports details of the solution. Marginal solutions are identified without stopping the script. Unless a clear solution is obtained, the *MrBUMP* script continues to process all trial models through *MOLREP*, after which the process is repeated using *Phaser* as the MR engine. For the cluster implementation of *MrBUMP*, trial models are processed in parallel, with a success on any cluster node stopping the script on all nodes.

A summary of *MrBUMP* results is given in Table 5. Unless explicitly indicated otherwise, these results were obtained during the workshop, without previous knowledge of the structures and without any *ad hoc* customization of the default script. The only exception to this is that if the actual target structure had been deposited, the script was run with this structure excluded. It became clear that the criterion for a good solution was too strict and in many cases the script continued processing trial models after a good solution had been found. In these cases, Table 5 shows one or more of the solutions identified as 'marginal' rather than as 'success'. For comparison, in some cases, the solution from the *Phaser* loop is shown alongside that from the *MOLREP* loop. The result of the MR step is a positioned but inaccurate and incomplete model. The cycles of restrained refinement often indicate that the model will refine and that a final model is likely to be realised (see, for example, BA4499). In other cases, there is no such clear-cut indication (for example, BA1563) and conclusions on the success of the procedure must await model rebuilding.

The second column in Table 5 shows the actual number of molecules in the asymmetric unit. In three cases, the automated script overestimated the correct number: BA1483, BA3935_2 and BA0592. For the last two this did not matter, since *MOLREP* failed to find the predicted last molecule and the refinement proceeded with the correct number of molecules. In the first case, *MOLREP* found seven molecules, so that the final model has a spurious extra molecule. At present, the *MrBUMP* pipeline does not deal explicitly with translational NCS, as occurs for example in OPPF651. In the majority of cases, there were several models and methods that could be used to solve the structure and the choice selected by the script (after identifying a 'success') or the author (after examining 'marginal' solutions) is largely arbitrary. Often, the structure could be solved both with the monomer search model and with a multimer. For example, BA0288 could be solved with chain *A* of 1ul1 or with the octamer downloaded from the PQS server. Both refine quickly to adequate *R* values, though the constrained geometry of the octamer leads

Table 6
Summary of results obtained by running *AutoAMoRe* on 18 target structures.

The column labels correspond to those in Table 5. If the R/Rf columns are blank, the solution was rejected owing to excessive clashing. The information in italics was obtained after the workshop.

Target	N_{mol}	Model	% ID	Initial R/Rf	Final R/Rf	Success	Rebuilt
BA0288	8	1u11	65	0.41/0.42	0.32/0.35	Y	Y
BA0296	2	1cli	53	0.46/0.47	0.35/0.48		
BA0592	6	1pjc	55				
BA1071	1	1ak1	73	0.42/0.41	0.26/0.41	Y	*Y*
BA1483	*6*	*1ecp*	*56*	*0.51/0.51*	*0.39/0.46*	*Y*	
BA1563	*2*	*1v8f*	*48*	*0.52/0.54*	*0.43/0.52*	*P*	*Y*
BA3935_1	4	1dhp	43	0.51/0.50	0.41/0.46	Y	*Y*
BA3935_2	4	1dhp	43	0.48/0.47	0.36/0.43	Y	*Y*
BA4499	1	1jr9	70	0.43/0.49	0.30/0.34	Y	*Y*
BA4508	1	1qum	31	0.52/0.51	0.39/0.52		
BA5696	*2*	*1jr9*	*55*	*0.44/0.44*	*0.30/0.34*	*Y*	*Y*
BA5705	2	1vrd	41				
BSAppa	1	1dpp	26				
BSYloQ	1	1uol	40	0.56/0.56	0.39/0.60		
CJ0982	2	1wdn	27				
OPPF 651	*2*	*1php*	*77*	*0.49/0.51*	*0.34/0.48*	*Y*	
OPPF1314	2	1ydm	47	0.49/0.46	0.37/0.46	Y	Y
OPPF 2245	2	1kon	36				

to a slightly poorer result. The advantage of the multimeric search is speed and potentially signal-to-noise ratio and in general it would be tried first.

Of the 17 structures attempted (ignoring duplicate entries in Table 5), ten were essentially solved, six were possibly solved but require further investigation and one was clearly unsolved. It should be stressed that these conclusions are based on statistics from the MR and refinement programs and in some cases from rebuilding in *ARP*/*wARP* and are provisional pending structure completion. In the case of CJ0982, which is deemed unsolved, the initial homology search yields only one hit which has a sequence identity of 44% but with an alignment length of only 70 residues. As discussed elsewhere, there were problems with the data processing of OPPF1314 and the result quoted in Table 5 is against the original problematic data truncated to 2.3 Å resolution.

The *MrBUMP* package is still under development. The current version is expected to automate structure solution *via* MR in straightforward examples. For the current test cases, it identified solutions or likely solutions in the majority of cases. Many of these cases had good homologues available in the PDB and could be solved by any reasonable method. For these, the advantage of *MrBUMP* is simply one of convenience, in particular when several homologues need to be tried and compared.

MrBUMP requires the *CCP*4 package plus a small number of helper applications and it was installed at York without problems. It is currently run from a simple shell script and users at the workshop found it easy to run the package for themselves. *MrBUMP* provides a framework within which further developments can be made in order to tackle more difficult cases. Ongoing work addresses both the algorithms applied in each step and the connecting work flow. Since the workshop, *MrBUMP* has been made available at http://www.ccp4.ac.uk/martyn/BMP/mrbump.php and feedback is encouraged.

5.1.2. Automated molecular replacement with *AutoAMoRe*. The advantages of *AMoRe* are its speed and flexibility. *AutoAMoRe* is a Python script produced as part of the *CCP*4 automation project to automate the numerous steps in solving a structure by MR using *AMoRe*. The *AutoAMoRe* script calls various *CCP*4 utility programs. It checks the native Patterson for translational NCS and, if appropriate, positions molecules on a pairwise basis. The final coordinates are generated using *PDBSET* and checked for clashes with *DISTANG*. *AutoAMoRe* generates a concise summary file of important parameters. The methodology adopted was to feed the target sequence into the *BLAST* server and choose as the template the solved structure with highest homology, after excluding any with 100% identity. The coordinates were downloaded from the EBI (http://www.ebi.ac.uk/) and passed through *CHAINSAW*. This selection and manipulation was performed manually and only the subsequent MR calculations with *AMoRe* were automated. However, the *AutoAMoRe* script has since been incorporated as a module in the *MrBUMP* pipeline. *AutoAMoRe* was run on 18 of the target data sets using a monomer model in each case. Solutions were scored by inspecting the final correlation coefficient for all molecules in the asymmetric unit and any solution with greater than 20 clashes was rejected. The highest scoring solution was subjected to ten cycles of refinement using *REFMAC* with the default parameters from the *CCP*4*i* GUI. If the R_{free} fell by more than 5% during refinement, the solution was deemed to be successful. Six structures were solved successfully using the version of the *AutoAMoRe* software available at the workshop. A number of subsequent improvements to the software resulted in solutions to another four cases, an overall success rate of 55%. The results are summarized in Table 6.

5.1.3. *BALBES*. *BALBES* is a system for automatic MR developed by FL, AAV and GNM. It has three main components: the PDB (Bernstein *et al.*, 1977; Berman *et al.*, 2000), which has been reorganized to aid model selection, a Python script which controls the work flow and makes decisions and scientific programs to perform the actual calculations. The script uses the following programs: *SFCHECK* for structure-factor analysis, *MOLREP* for molecular replacement and *REFMAC* for refinement. Several other programs for purposes such as alignment and searching in the reorganized PDB have been developed.

The ∼30 000 structures in the PDB have been reorganized and classified according to sequence and three-dimensional structure. Redundant entries are removed if two proteins have a sequence identity above 90% or if the root-mean-square

Table 7
Summary of results obtained using *BALBES*.

Column labels correspond to those in Table 5, with the addition of columns indicating the presence of pseudo-translational symmetry and more detail on the multimeric state and number of copies expected (N_{mol} expected) and found (N_{mol} found).

Target	N_{res}	Model	% ID	Pseudotranslation	N_{mol} expected	N_{mol} found	Initial R/Rf	Final R/Rf	Success†
BA0288	161	1xmp, 1o4v	100	N	4 × 2	4 × 2	0.31/0.31	0.24/0.28	Y
BA0592	377	1pjc	54	N	7	6	0.42/0.42	0.30/0.36	Y
BA1071	311	1doz	72	N	1	1	0.45/0.45	0.29/0.39	Y
BA1483	235	1xe3	100	N	4 × 2	4 × 2	0.44/0.44	0.22/0.27	Y
BA1563	282	1ufv	48	Y	2	2	0.58/0.56	0.43/0.52	Y/M
BA3935_1	292	1dhp	41	N	4	4	0.52/0.52	0.39/0.41	Y
BA3935_2	292	1dhp	41	N	4	4	0.49/0.47	0.33/0.38	Y
BA4499	283	1jr9	70	N	2	2	0.44/0.44	0.29/0.34	Y
BA4508	298	1qtw	32	N	2	1	0.52/0.54	0.40/0.50	Y/M
BA5696	208	1jr9	55	N	2	2	0.45/0.44	0.29/0.33	Y
BA5705	327	1eep	33	N	2	2	0.55/0.56	0.41/0.47	Y
OPPF651	394	1php	77	Y	2	2	0.65/0.65	0.31/0.39	Y
OPPF1294	193	1yby	52	N	2	None			N
OPPF1311	255	1lm4	39	Y	2 × 2	2	0.60/0.60	0.56/0.63	N
OPPF1314	200	1ydm	47	N	2	2	0.49/0.48	0.36/0.43	Y
OPPF2088	139	1oqq	26	N	1	1	0.58/0.55	0.49/0.59	N
OPPF2153	222	1qu0	23	Y	2 × 2	2	0.69/0.70	0.51/0.57	P/M
OPPF2245	229	1kon	37	N	2	2	0.54/0.55	0.41/0.53	P/M

† Y, definite solution; P, probable solution; N, no solution; M, checked manually.

deviation between matched atom pairs after superposition is less than 1 Å. This reduces the number of entries to a reference set of ∼10 000 structures, which are organized into a hierarchical database based on similarity. For each entry, potential multimers and domain structures are established and catalogued.

The Python script reads the experimental data and sequence information for the protein under investigation. The reorganized PDB is searched for related sequences, candidate models are identified and coordinates returned with multimers and domains where appropriate. The whole search takes around 10 s on a Macintosh G5 computer. The putative models are modified according to sequence identity and surface accessibility. The experimental data are analysed using *SFCHECK*, which indicates problem features such as pseudo-translation, twinning or anisotropy and suggests the best resolution for the MR search. Information from these analyses is passed to *MOLREP*. Several protocols are tested in order: first with multimers, then with individual subunits and then with domains. After each protocol, a decision is made as to whether the 'solution' is correct. If the number of expected monomers is not yet satisfied, then MR is continued. During the workshop models were passed directly to *ARP*/*wARP* for rebuilding. Subsequently, a better protocol has been implemented: the MR solution is first passed to *REFMAC* for a number of cycles of rigid-body followed by restrained refinement. This has led to substantially better results for the rebuilding.

BALBES is at the development stage and the database is updated automatically and regularly. For the current tests the database includes PDB entries released by the end of 2004.

Only a subset of the available protocols was necessary for the workshop examples. These included simple MR with one subunit (two cases, one successful), a search with dimers (four cases, two successful, one probably successful), a stepwise search for multiple subunits (13 cases, ten successful and one probable) and a use of pseudo-translation (four cases, two successful, one probable). More sophisticated protocols such as domain searches, multi-copy searches, iterative refinement and MR were not required for the current tests. Experience at the workshop suggested that implementation of several protocols needed to be faster. The results are summarized in Table 7.

5.2. Summary of molecular-replacement pipelines

Out of the structures considered, the majority have a close homologue available and are straightforward to solve by MR. The minority that are not straightforward to solve are the interesting examples for methods development and will be the focus of further work. The difficulty may arise from problems in the data processing. In other cases, more sophisticated model generation may be needed or experimental phasing is required.

The solution of the OPPF1314 structure was one of the key achievements of the workshop. All three pipelines were able to determine the solution with either the low-resolution data integrated using *DENZO*/*SCALEPACK* or that processed using the three-dimensional XDS option of *XIA-DPA*. Refining and building a complete model proved more challenging and was only possible after reprocessing the data.

Comparison of the results with *MrBUMP* and *BALBES* at the workshop showed that even for identical models *ARP*/*wARP* performed better with the *MrBUMP* solutions. A key difference between the two approaches was the absence of an automated refinement step for *BALBES* prior to attempting to rebuild the model with *ARP*/*wARP*. The introduction of refinement into the *BALBES* protocol subsequent to the

Table 8
Rebuilding with *REFMAC-ARP/wARP*.

For each MR solution, both *ARP/wARP* and *pyWARP* were tried: *ARP/wARP* as an assessment of the quality of each MR solution and *pyWARP* for its own evaluation. Values correspond to the number of residues traced in the asymmetric unit, with the number of residues for which side chains are built in parentheses.

				MrBUMP		*AutoAMoRe*		*BALBES*	
ID	Resolution (Å)	N_{mol}	Total residues	*ARP/wARP*	*pyWARP*	*ARP/wARP*	*pyWARP*	*ARP/wARP*	*pyWARP*
BA0288	1.80	8	1288	1246 (1084)	1252 (1252)	1241 (1145)	1249 (1249)	1250 (1235)	1248 (1166)
BA0592	2.84	6	2262	549 (35)	429 (162)	—	—	616 (28)	475 (123)
BA1071	2.60	1	311	201 (95)	181 (155)	199 (140)	189 (162)	206 (141)	145 (125)
BA1483	2.24	6	1410	1364 (1364)	1334 (1328)	357 (43)	1382 (1269)	1369 (1359)	1389 (1230)
BA1563	2.20	2	564	281 (132)	248 (216)	5 (0)	51 (27)	11 (0)	85 (26)
BA3935_1	1.94	4	1168	1137 (1130)	1141 (1141)	1109 (1061)	1136 (1136)	1143 (1143)	1139 (1139)
BA3935_2	2.23	4	1168	955 (674)	1057 (1044)	721 (327)	305 (182)	1040 (907)	1081 (1038)
BA4499	1.80	2	406	373 (373)	384 (384)	378 (378)	349 (349)	371 (371)	379 (379)
BA4508	2.57	2	596	105 (21)	108 (45)	28 (0)	118 (80)	142 (39)	110 (91)
BA5696	1.80	2	416	386 (386)	391 (391)	391 (391)	396 (396)	396 (396)	390 (390)
BA5705	1.80	2	654	533 (533)	547 (547)	—	—	531 (526)	528 (528)
OPPF651	2.40	2	788	233 (44)	245 (159)	—	—	598 (484)	656 (646)
OPPF1294	—	2	386	—	—	—	—	—	—
OPPF1311	2.72	4	1020	—	—	—	—	Crash	83 (23)
OPPF1314	2.30	2	400	200 (96)	197 (163)	—	—	184 (0)	179 (158)
OPPF2153	2.69	2	444	—	—	—	—	11 (0)	173 (43)
OPPF2245	3.30	2	458	43 (0)	102 (37)	—	—	19 (0)	82 (41)

workshop now gives highly comparable results for the two pipelines (Tables 5 and 7).

At the simplest level, the role of automation is one of convenience, providing a solution with little user effort. However, it is likely that automation can provide more objective criteria of relative success for different models and also optimize the solution to minimize the effort required later for model rebuilding and completion.

6. Experimental phasing

Experimental phasing was not addressed in any depth as part of the automation testing, as the pipelines are at an earlier stage of development. Again, it became clear that many key items of information were not directly available, *e.g.* several reflection files did not provide correct wavelength information or document MAD data sets. OPPF1294, OPPF1311, OPPF2088 and OPPF2153 had all been solved previously using the *SHELX* suite. However, the SiaP structure was phased and largely built during the workshop, providing insight into how best to use weak phase information for automated model building at moderate resolution.

The principal software vehicle for this investigation was *Pirate*, a statistical phase-improvement program (Cowtan, 2000). *Pirate* classifies the electron density by sparseness/denseness and order/disorder without requiring knowledge of the solvent content. Statistical targets are constructed, from which the distribution of probable density values is inferred on the basis of local density mean and variance. These targets are optimized to the problem at hand by the use of a known 'reference' structure which is manipulated by a process of scaling and error simulation to produce a map which is statistically similar to the map under examination. The software, which is still under development, is designed to be used in a fully automated manner.

The SiaP structure was phased using only the peak data from the Se sites found using *SHELXD* and initial phasing performed by *SHELXE*. This uses a solvent-flattening procedure to refine the initial SAD estimates and outputs phases and associated figures of merit. During the workshop, phases were recalculated using *MLPHARE* to record Hendrickson–Lattman (HL) coefficients. The average figure of merit was 0.43, falling to 0.15 at 2.7 Å. *Pirate* was used to improve these SAD phases, reducing the overall phase error from 64 to 48° (evaluated against the final model fully refined after the workshop) and providing better and much more realistic estimates of the figure of merit. Model completion was attempted from both these starting phase sets (§7.4).

7. Constructing and completing models

There was not sufficient time at the workshop to explore fully the best approach to the problem of model preparation. In cases where the resolution of the data set extends to ∼2.3 Å, packages such as *ARP/wARP* can often build a model automatically, providing that there are sufficiently good quality starting phases. *ARP/wARP* was run on all suitable MR solutions and the results are given in Table 8; this exercise was subsequently repeated using *pyWARP* (Cohen *et al.*, 2004; Table 9). However, model building is a real stumbling block for lower resolution data sets and models with low sequence identity (typically less than 30%). Full automation of the MR pipeline for low-resolution data may require the incorporation of new modules such as *Buccaneer* (Cowtan, 1998, 2001), which is designed to recognize larger structural features. In all cases it is necessary to complete the model using a graphical display and *Coot* (Emsley & Cowtan, 2004) can provide this functionality. These modules are now described in more detail.

Table 9
Rebuilding OPPF1314 with *REFMAC-ARP/wARP*.

After the OPPF1314 data were reprocessed, both *ARP/wARP* and *pyWARP* were applied to the two-dimensional (OPPF1314M) and three-dimensional (OPPF1314X) data sets at three different resolution cutoffs. The best results at 1.5 Å are obtained from the three-dimensional integration, demonstrating better treatment of high mosaicity. For all tests, the extra functionality of *pyWARP* proved more successful. Column labels are equivalent to those in Table 8.

				BALBES	
ID	Resolution (Å)	N_{mol}	Total residues	*ARP/wARP*	*pyWARP*
OPPF1314X	1.5	2	400	193 (109)	294 (279)
OPPF1314M	1.5	2	400	131 (34)	257 (246)
OPPF1314X	1.65	2	400	277 (268)	322 (316)
OPPF1314M	1.65	2	400	253 (208)	303 (303)
OPPF1314X	1.85	2	400	271 (249)	323 (323)
OPPF1314M	1.85	2	400	261 (226)	322 (322)

7.1. *ARP/wARP*

ARP/wARP was used to evaluate the quality of the solutions obtained from the different MR pipelines. Each solution was input to the currently distributed version of *ARP/wARP* (v.6.1.1) and ten rebuilding cycles were performed (each comprising five update cycles) starting from the positioned model (using the mode described in Perrakis *et al.*, 1999, 2001). In this procedure, most of the stereochemical information is preserved as long as possible during the building, giving the refinement program *REFMAC* more restraints.

Some of the data sets highlighted an error in the sequence-docking/side-chain building module of *ARP/wARP*, which occurred when a main-chain fragment became longer than the provided sequences. This was fixed following the meeting and the corrected version used to generate Tables 8 and 9. All solutions now run through to the preset end, with the exception of the *BALBES* solution for OPPF1311, which still causes problems.

The available MR solutions were used to evaluate *pyWARP*, a new control system for *ARP/wARP* currently under development (Cohen *et al.*, 2004). This control system makes run-time decisions based on the current status of the model. Tables 8 and 9 also show the results from *pyWARP*, which appears to perform a little better than *ARP/wARP* and clearly docks a greater portion of the traced main chain into the sequence. Indeed, in difficult cases (OPPF2153 and OPPF2245) *pyWARP* significantly improved the completeness of the autotraced model, showing the value of using variable parameterization during the procedure.

7.2. *Buccaneer*

Buccaneer is a new model-building program which makes repeated application of a single optimized feature-recognition technique. The process involves the construction of an optimal likelihood density target for the electron density in a 4 Å sphere around a typical C^{α} atom (the idea, but not the target function, is similar to that in Ioerger & Sacchettini, 2002). The 'optimization' of the target is similar to that described in §6. A six-dimensional search is made in the unsolved map for likely C^{α} atom positions using Fast Fourier Feature Recognition (Cowtan, 2001). Once initial candidate positions are obtained, a 'growing' procedure is applied to find chain fragments. New residues are added in each direction using the Ramachandran plot to constrain chain geometry and a two-residue-deep search to rank the fit to density of the new positions. It is computed using the same likelihood density target, but now calculated in real space. The procedure continues until the fit to density falls below some threshold. The next stage is to combine and merge overlapping chain fragments, using the *Coot* utility GLOBULARISE-PROTEIN. The approach is implemented using the CLIPPER libraries (Cowtan, 2003). Despite its simplicity, it quickly rebuilt missing features for two of the lower resolution MR structures, BA1071 and BA4508. Its performance with the SiaP structure is described below. It can be integrated into a recycling scheme including density modification and refinement. Other improvements are possible, such as using bi-residue groups in common conformations.

7.3. *Coot*

Coot is a molecular-graphics application for protein map interpretation and structure validation. The workshop highlighted a number of its strengths, but also some missing features. It proved very powerful for rebuilding initial models from automated model building. The validation tools identify poorly built regions of the model, both by a variety of geometrical indicators and by fit-to-density analysis. These regions can be rapidly improved by interactive real-space refinement and regularization. Although *Coot* was not used greatly in its role as a model-completion and validation tool by this workshop, missing features were highlighted including a means to reverse a baton-built C^{α} trace, better tools for joining fragments, a user interface for the automatic restoration of side chains and tools to correct out-of-register errors (a more substantial problem). Some of these deficiencies have now been addressed. *Coot* presently implements all these algorithms from a graphical interface, but it should be possible to incorporate the underlying functionality into a command-line-driven program suitable for an automated pipeline.

7.4. Experimental phasing case history: SiaP

Several attempts were made to rebuild this structure using *ARP/wARP*. In §2.2 we describe the results obtained before the workshop. The breakthrough came from the procedures described in §6, namely when the SAD experimental phase distributions in the form of the Hendrickson–Lattman coefficients calculated using the *MLPHARE* program were provided as restraints to *REFMAC-ARP/wARP*. Firstly, the procedure was started from the *RESOLVE* partial model and the *MLPHARE* phases and within 25 cycles it had built 560 of the expected 612 residues, with 545 side chains docked. Secondly, the more straightforward approach of feeding the Hendrickson–Lattman coefficients directly into the *ARP/wARP* procedure used for building the initial model was

attempted. This took longer, but effectively reached the same solution. The third and fourth tests used *Buccaneer* to construct an initial model. The third test began from the *MLPHARE* phases, from which *Buccaneer* built a polyalanine model of 288 residues (47% of the total). The fourth test used the improved phases from *Pirate* and with these *Buccaneer* was able to construct a 384-residue polyalanine model (63%). Both these models were able to kick-start the *ARP*/*wARP* procedure and speeded up its convergence considerably. Starting from the *Pirate*/*Buccaneer* model, *ARP*/*wARP* completely built 578 residues (94%).

With the current state of developments this is an impressive result: the automated building and refinement of an essentially complete protein structure with rather weak 2.6 Å SAD data. The application of any specific program in this solution is certainly less important than the retention of the full experimental phase distribution as restraints in the model-building/refinement stage.

This result has influenced several developments within the *CCP*4 and York automation pipelines currently under construction.

(i) Experimental phase restraints, in the form of Hendrickson–Lattman coefficients, were essential to keep the *ARP*/*wARP*-*REFMAC* cycles on target.

(ii) Phase improvement using *Pirate* improved this. It is important that the weighting of the experimental phases is realistic.

(iii) While initial model building using feature recognition such as *Buccaneer* or *RESOLVE* was not able to generate a complete structure, starting models were created which considerably speeded up the *ARP*/*wARP*-*REFMAC* process.

7.5. Molecular-replacement case history: OPPF1314

Before the workshop, the *DENZO*/*SCALEPACK* data set from the high-resolution data-collection pass alone had given a clear MR solution with the expected two molecules in the asymmetric unit. Structure completion with *ARP*/*wARP* was partly successful, but convergence was slow: missing data at low-resolution inevitably degrade the electron density which *ARP*/*wARP* requires for selection and rejection of atomic sites.

After initial reprocessing with the *XIA*-*DPA* three-dimensional option, the data from the combined high- and low-resolution passes showed an expected distribution in reflection intensities (§4.3; Fig. 1). This allowed *ARP*/*wARP* to produce a structure containing 329 residues in nine chains out of the 400 residues expected in the asymmetric unit (for the final cycle of rebuilding the R factor was 0.234 with an $R_{\rm free}$ of 0.289). The maps showed that one part of each molecule was poorly ordered, explaining the missing residues. Further inspection of the maps revealed additional features in the electron density that were not part of the protein and could be attributed to bound ligands (§7.6).

The two-dimensional integrated data also led to a similarly successful result (Table 9) after reprocessing using the improved *XIA*-*DPA* software (§4.3). However, with data from both the three-dimensional and two-dimensional integration *ARP*/*wARP* built more residues when the data were restricted to 1.85 Å resolution rather than using the full range to 1.5 Å. This may reflect the optimization of *ARP*/*wARP*, the poorer quality of the outer shells or be the result of residual errors arising from the effect of the mosaicity on the highest resolution data. In all cases, the extra functionality of *pyWARP* proved its worth.

Since the workshop, refinement of this structure has been completed, giving a model containing 192 residues from each chain, one ADP bound to each chain and 256 modelled waters. The current R factor is 0.219 (with an $R_{\rm free}$ of 0.265).

What conclusions can be drawn from this case study?

(i) The importance of data quality and flagging of unexpected values of the quality-assessment parameters at the data-processing stage (see §4.3.1).

(ii) The importance of completeness of low-resolution data for electron density.

(iii) The robustness of MR methods even with substandard data.

(iv) The value of *CHAINSAW*-type procedures before the MR search and after MR solution but before rebuilding.

(v) The automated updating of parameters during the course of model rebuilding using *pyWARP* led to significantly better performance compared with the normal *ARP*/*wARP* procedure.

(vi) A good test for ligand fitting (§7.6).

7.6. Ligand binding to OPPF1314

7.6.1. *ARP*/*wARP LigandBuild*. Once a model is close to completion, the remaining density can be searched for small-molecule ligands. This procedure can be accomplished using the *ARP*/*wARP* suite (v.6.1.1) with *CCP*4 and a text editor. The *ARP*/*wARP LigandBuild* graphical user interface requires structure-factor amplitudes, protein coordinates without any HETATM entries (used to generate a mask) and a set of coordinates for the known ligand (Zwart *et al.*, 2004).

For OPPF1314, the extra density after model building (§7.5) was assumed to be attributable to one or both of the cofactors in the crystallization screen, ATP and 5-formyltetrahydrofolate, which have a somewhat similar shape. Automated ligand fitting was attempted for both cofactors. The first and second trials failed; the map was still too noisy and wrong sites with impossible conformations were found. In each case, the volume covered by these was then added to the mask and the procedure was repeated. The correct sites were found in the third and fourth attempts and were verified by inspection of the electron density. The fit with ATP was clearly superior and after inspection of the electron density it was concluded that each of the two molecules in the asymmetric unit bound a well ordered ADP (Fig. 2). Although there were other residual density features, they could not be unambiguously attributed to 5-formyltetrahydrofolate.

The result showed the need for updating masks. The PDB file was modified automatically to add extra 'atoms' at each

cycle. This is also needed when searching for multiple ligands, where it is recommended to build the largest first. It also showed how the recognition capabilities of the software are limited by noise in the map, as the search currently only checks a limited number of features for a potential ligand site.

7.6.2. Ligand building with *Coot*. *Coot* has the ability to search a map for likely ligand sites. It uses the *REFMAC* monomer dictionary to provide a description of the ligand geometry and also needs a set of coordinates for the known ligand, which can be provided by the *CCP*4 program *LIBCHECK* (Vagin *et al.*, 1998). As with *ARP/wARP LigandBuild*, the density is masked by selected coordinate sets. After *ARP/wARP* for OPPF1314, *Coot* found seven putative ligand sites matching the expected size and shape for ADP. On visual inspection, several were found to be protein structure missing from the model, but the first and second sites ordered on the density correlation corresponded to the two nucleotide sites. The fit to density was then optimized using *Coot*'s real-space refinement option.

7.7. Summary of model rebuilding

From Tables 8 and 9, it is noticeable that the success of *ARP/wARP* varies substantially depending on how the positioned model was obtained. For target BA1563, for example, *ARP/wARP* was able to rebuild about half of the model using the output of *MrBUMP*, while very little could be rebuilt from the other putative solutions. Both *MrBUMP* and *BALBES* used 1ufv as the template and *MOLREP* for MR. At the workshop, for target BA3935_2 *ARP/wARP* rebuilt about 80% of the model from *MrBUMP*, but only about 50% of the model from *BALBES*. Post mortem analysis showed that the difference was that *MrBUMP* carried out 30 cycles of *REFMAC* refinement before *ARP/wARP*. This refinement step has subsequently been activated in *BALBES*.

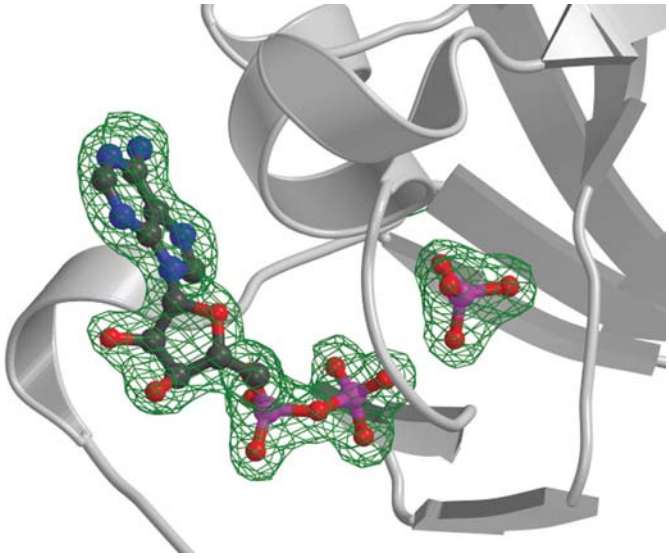

Figure 2
Part of the model for OPPF1314 showing the bound cofactor ADP (with a separate phosphate group presumably resulting from ATP hydrolysis) after automatic fitting of the cofactor by either *ARP/wARP* or *Coot* and further refinement. Green contours show the OMIT-map density for the cofactor contoured at 3σ. This figure was drawn using *BobScript* (Esnouf, 1999) and rendered with *Raster3D* (Merritt & Murphy, 1994).

We have not carried out a systematic investigation of the factors which are important for successful rebuilding and the differences noted may be coincidental. However, where there is no clear-cut MR result, there is a clear advantage in taking several putative solutions through to model rebuilding. In the BA3935_2 example, both templates 1dhp and 1s5t have 42% sequence identity with the target and both should be tried. The high-resolution limit of the data for BA1563 and BA3935_2 is in both cases 2.2 Å and it is in this regime where subtle differences may affect the *ARP/wARP* procedure. Automated schemes are particularly useful for investigations of such multiple models.

8. Conclusions

Protein crystallography has a series of potential bottlenecks, including protein overexpression, solubility, crystallization and structure solution. Recently, rapid advances in the first three of these have been made (see other contributions to this issue). Considerable progress has been and is being made worldwide in the automation of structure analysis. The automation of image processing and data reduction was addressed at the workshop using *XIA-DPA*. The results obtained emphasized the importance of this step and showed that while it was in principle subject to a high level of automation, great care needs to be taken in establishing protocols and in passing the appropriate information to subsequent steps.

The MR step in the structure-solution pipeline is presently closest to full automation. Three emerging procedures were extensively tested with a high success rate and this software should be released for general use within the next year. Lessons learnt at the workshop included (i) the apparent advantage of running several cycles of restrained refinement on correctly positioned MR models before starting the rebuilding procedure and (ii) generally trying a multimer as a model when appropriate before trying individual subunits. Based on the results obtained, automated MR procedures are likely to be successful, at least for crystals which diffract to 2.5 Å or better and satisfy a number of defined criteria (overall $R_{\rm merge} \simeq 6\%$, low-resolution shell $\simeq 4\%$, high-resolution shell $\simeq 35\%$, completeness $\simeq 90\%$). At present, twinning poses real problems, but this should be resolved in the near future: for crystals with merohedral twinning, diffraction to ∼2.1 Å or better may be necessary.

Models for MR should satisfy one of the following criteria.

(i) ∼30% identity for one molecule in the asymmetric unit and no significant domain movement.

(ii) ∼45% identity for multiple molecules in the asymmetric unit and no significant domain movement.

(iii) >50% identity for two or more molecules in the asymmetric unit, where there is significant domain movement.

For structures solved by experimental phasing, there are already modules such as the *SHELX* suite, *AUTOSHARP* and *SOLVE*/*RESOLVE* which integrate parts of the pipelines.

These were tested on a couple of examples at the workshop and the importance of making full use of the Hendrickson–Lattman coefficients for low-resolution data became clear. *Pirate* was tested for density modification and appears to provide a more realistic set of Hendrickson–Lattman coefficients than earlier software. Pipelines being developed are still at the early stages, but considerable insight was gained as to the direction which these developments should take. The restrictions on data quality are quite different from those stated above for MR; experimental phasing is effective at substantially lower resolutions but requires much more accurate estimates of intensity, usually achieved by measuring high-multiplicity data sets.

ARP/*wARP* was the only automated model-building tool used extensively. It proved to be very powerful for structures with data extending to 2.3 Å or better (see Table 8). At lower resolutions problems were encountered and the feature-recognition tools within the *Buccaneer* and *Coot* programs, briefly tested during the workshop, would need to be exploited for these problems. However, crystals with diffraction limits in the greyzone (2.7–3.3 Å) still require a lot of time and effort and sometimes this effort fails. High-throughput structure determination means a limited amount of time can be dedicated to an individual project, resulting in a need for automation. We have encountered several examples, BA0592 (from the workshop set) and BA4525 (collected recently), where crystals have been obtained, data collected and a MR solution found, but the project has had to be abandoned because automated model building and refinement failed. The more sustained effort of project-oriented research may have led to success.

Taken altogether, the outlook is very promising for modules for fully automated solution of protein crystal structures in the near future, provided the data are of sufficient quality and resolution.

The SPINE project is funded by the European Commission as SPINE (Structural Proteomics In Europe) contract No. QLG2-CT-2002-00988 under the Integrated Programme 'Quality of Life and Management of Living Resources'. GW and RK are supported by the BBSRC e-HTPX grant (BEP17782) and CB, NS, MGWT and MW by CCP4. KDC is supported by The Royal Society (grant No. 003R05674). PE and AAV are funded by BBSRC grant No. 87/B17320. GNM is supported by the Wellcome Trust, FL by the EU BIOXHIT contract under the Sixth Framework Programme thematic area 'Life Sciences, Genomics and Biotechnology for Health' contract No. LHSG-CT-2003-503420. *ARP*/*wARP* algorithm development at the NKI (AP, SXC) and the EMBL (VL, GL) is funded by the NIH (grant R01 GM62612-01) and the EU BIOXHIT contract. AP and SXC thank Marouane Ben Jelloul for his work in the development of *pyWARP*.

References

Alzari, P. M. *et al.* (2006). *Acta Cryst.* D**62**, 1103–1113.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
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.
Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
Cohen, S. X., Morris, R. J., Fernandez, F. J., Ben Jelloul, M., Kakaris, M., Parthasarathy, V., Lamzin, V. S., Kleywegt, G. J. & Perrakis, A. (2004). *Acta Cryst.* D**60**, 2222–2229.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. (1998). *Acta Cryst.* D**54**, 750–756.
Cowtan, K. (2000). *Acta Cryst.* D**56**, 1612–1621.
Cowtan, K. (2001). *Acta Cryst.* D**57**, 1435–1444.
Cowtan, K. (2003). *IUCr Comput. Commun. Newsl.* **2**, 4–9.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Esnouf, R. M. (1999). *Acta Cryst.* D**55**, 938–940.
Evans, P. (1993). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by L. Sawyer, N. Isaacs & S. Bailey, pp. 114–122. Warrington: Daresbury Laboratory.
Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.
French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517–525.
Huennekens, F. M., Henderson, G. B., Vitols, K. S. S. & Grimsha, C. E. (1984). *Adv. Enzyme Regul.* **22**, 3–13.
Ioerger, T. R. & Sacchettini, J. C. (2002). *Acta Cryst.* D**58**, 2043–2054.
Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.
Leslie, A. (1999). *Acta Cryst.* D**55**, 1696–1702.
McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Cryst.* D**61**, 458–464.
Meier, C., Carter, L. G., Esnouf, R. M., Owens, R. J. & Stuart, D. I. (2006). In preparation.
Merritt, E. A. & Murphy, M. E. P. (1994). *Acta Cryst.* D**50**, 869–873.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Perrakis, A., Harkiolaki, M., Wilson, K. S. & Lamzin, S. (2001). *Acta Cryst.* D**57**, 1445–1450.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Sauter, N., Grosse-Kunstleve, R. & Adams, P. (2004). *J. Appl. Cryst.* **37**, 399–409.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Schwarzenbacher, R., Godzik, A., Grzechnik, S. K. & Jaroszewski, L. (2004). *Acta Cryst.* D**60**, 1229–1236.
Terwilliger, T. C. (2003). *Methods Enzymol.* **374**, 22–37.
Vagin, A. A., Murshudov, G. N. & Strokopytov, B. V. (1998). *J. Appl. Cryst.* **31**, 98–102.
Vagin, A. A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.
Vaguine, A. A., Richelle, J. & Wodak, S. (1999). *Acta Cryst.* D**55**, 191–205.
Zwart, P. H., Langer, G. G. & Lamzin, V. S. (2004). *Acta Cryst.* D**60**, 2230–2239.