Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Automatic structure determination based on the single-wavelength anomalous diffraction technique away from an absorption edge

Gordon A. Leonard, Germaine Sainz, Maaike M. E. de Backer and Seán McSweeney*

Macromolecular Crystallography, European Synchrotron Radiation Facility, BP 220, F-38043, Grenoble CEDEX, France

Correspondence e-mail: seanmcs@esrf.fr

Received 26 August 2004
Accepted 5 January 2005

The phasing of macromolecular structures based on the use of the single-wavelength anomalous diffraction method has recently enjoyed a revival. Here, additional evidence is provided that the method may be successfully applied at wavelengths remote from the absorption edge of interest and that it is in principle applicable to a large number of systems. This opens up the possibility of rapid and reliable automatic *de novo* structure determination using simple experimental configurations with no need for wavelength tunability or absorption-edge scanning. The method should therefore be exploitable at most synchrotron beamlines. The effects of data completeness and multiplicity on the quality of the phases obtained are discussed as are the prospects for the automation of macromolecular structure solution based on the experimental protocols described.

1. Introduction

The burgeoning of many structural genomics initiatives requires that many hundreds, perhaps thousands, of macromolecular structures are determined rapidly and reliably. Increasing attention is thus being focused on the use of automation in all aspects of macromolecular structure determination (Arzt *et al.*, 2005). Progress is being made in the areas of automation of sample changing (Abola *et al.*, 2000) and sample characterization (Leslie *et al.*, 2002) and methods have been available for some time that address the automation of phasing (Terwilliger & Berendzen, 1999*a*) and model-building (Perrakis *et al.*, 1999) procedures. However, the automatic production of usable experimental data to feed these latter processes remains problematic.

The multiwavelength anomalous dispersion (MAD) method (Hendrickson, 1991, 1999; Hendrickson & Ogata, 1997) has risen to a position of pre-eminence amongst experimental phasing methods and it is now a straightforward and widely accepted technique for producing *de novo* phase information for use in macromolecular structure determination. Indeed, in suitable cases, structure determination using the MAD method can be effected in a matter of hours (Walsh *et al.*, 1999). However, the use of the techique is not without its drawbacks. There is a need for special equipment, such as energy-dispersive fluorescence detectors on beamlines and the requirement for careful accurate data collection at a number (typically three) of wavelengths means that MAD experiments place great demands on instrumentation reliability, reproducibility and stability. Owing to the efforts of many scientists over the last decade, these problems have largely been over-

come. In tandem with developments in beamline technology, software developments have also kept pace with interest in MAD structure determination and the elucidation of large anomalous scattering substructures is now routine (von Delft *et al.*, 2003). However, a potentially more serious problem for the technique has recently emerged with the increased use of highly brilliant undulator beamlines for data collection. This is the phenomenon of radiation damage, which can severely limit the amount of data collected from the crystal sample and also removes one of the prime advantages of MAD: the nearly perfect isomorphism of the data collected at the different wavelengths (Ravelli & McSweeney, 2000). One of the consequences of radiation damage is that in many cases the data-collection protocol required for a successful MAD experiment is no longer straightfoward.

Possibly as a result of this drawback, there has recently been a great deal of interest in using single-wavelength anomalous diffraction (SAD) data in the elucidation of macromolecular structures (Harvey *et al.*, 1998; Brodersen *et al.*, 2000; Rice *et al.*, 2000; Dauter *et al.*, 2002), with investigations showing that the SAD technique may be applied to many diverse problems, ranging from weak anomalous signals to highly complex substructures (von Delft *et al.*, 2003). In principle, the SAD method is used with data collected close to an absorption edge of the anomalous scatterer in the sample under investigation. However, the work of Hendrickson & Teeter (1981) and later developments by Wang (1985) demonstrated the possibility of exploiting the weak anomalous signals available away from the edge features. This work was revisited when the usefulness of the anomalous scattering of sulfur at the Cu $K\alpha$ wavelength to solve macromolecular crystal structures was demonstrated (Dauter *et al.*, 1999). These experiments have since been complemented by various investigations at longer wavelengths (Weiss *et al.*, 2001; Gordon *et al.*, 2001; Cianci *et al.*, 2001; Micossi *et al.*, 2002; Ramagopal *et al.*, 2003*a*,*b*) and together they have demonstrated that it is possible to successfully utilize anomalous signals as low as 0.6% for SAD phasing.

Although one needs to overcome the problem of bimodal phase-probability distributions inherent to the SAD technique (see Hauptman, 1996; Langs *et al.*, 1999; Hao *et al.*, 2000; Dauter *et al.*, 2002, for discussion), the basic simplicity of the SAD method and the success of SAD structure determinations using very small anomalous signals tempted us to investigate the possibility of employing a further simplification of the SAD technique; namely SAD structure determination on a synchrotron beamline without wavelength tunability. This idea complements work described by Brodersen *et al.* (2000) and our interest in this experimental approach for structure determination addresses many of the problematic issues described previously. Most obviously, use of data at a single wavelength means that the data-collection process is greatly simplified Additionally, as data collection will not normally be close to an absorption edge, one may expect reduced radiation damage to the anomalous scatterer. The resulting simplification of data-collection protocols would make them suitable for incorporation into existing phasing and model-building packages, thus allowing automatic procedures in which all aspects of the structure-determination process are carried out without the need for manual intervention.

Once experimental intensity data have been collected and processed, in the majority of cases structure determination using the SAD technique proceeds *via* a three-step process. Firstly, the determination of the positions of the anomalous scatterers is carried out; phases are then developed in order to produce electron-density maps and, in the final stage, these are interpreted using either manual or automatic methods to produce a starting model for refinement procedures. An automatic system must successfully pass each of these stages and therefore in addition to the potential simplification of data-collection procedures, we were also interested in the amount of data that was necessary for each step to be successfully completed and in the correlation between the completion of one stage and the optimization of the next.

2. Experimental methods

Two systems for which the structures were already known were chosen as test cases. Both contain anomalously scattering atoms for which the absorption edges are at a much longer wavelength than that at which data were actually collected. The two systems have somewhat different characteristics. They contain different scatterers, one is of relatively high molecular weight the other rather low, the Laue symmetry of the two systems is different (one rather higher than than other) and one contains two monomers in the asymmetric unit, the other a single copy of a monomer. Given these differences and the rather small anomalous signals available (see below), we are thus confident that our assessment of the potential exploitation of the techniques outlined in the paper results in conclusions that may be generally applicable.

The first test case was shrimp alkaline phosphatase (SAP) (Nilsen *et al.*, 2001; Olsen *et al.*, 1991; de Backer *et al.*, 2002; PDB code 1k7h), provided by Biotec Pharmacon ASA (Tromsø, Norway). The protein monomer has a molecular weight of 53 kDa and contains three Zn^{2+} ions in its active site. f'' for Zn^{2+} at the wavelength of our experiment (0.933 Å) is 2.3 e^- (calculated with the *CCP*4 program *CROSSEC*) and the anomalous signal calculated according to Hendrickson *et al.* (1985) is approximately 1.95%. The enzyme crystallizes in space group $P4_32_12$, with unit-cell parameters a = 171.07, c = 84.32 Å. The crystals contain a dimer in the asymmetric unit, resulting in a Matthews coefficient (Matthews, 1968) of 2.9 $Å^3$ Da^{-1} and a solvent content of 57%. Cryoprotection during data collection was obtained by soaking in the crystallization solution with 30%(*v*/*v*) glycerol.

The second test case was ferredoxin VI (FdVI) from *Rhodobacter capsulatus* (Armengaud *et al.*, 2001; PDB code 1e9m). This protein has a molecular weight of 11.6 kDa and each molecule contains a single [2Fe–2S] cluster, yielding an anomalous signal of 1.45% (λ = 0.933 Å, f'' for Fe^{2+} is 1.4 e^-). Crystals belong to the orthorhombic space group $P2_12_12_1$, with unit-cell parameters a = 45.28, b = 49.2, c = 54.33 Å and one molecule in the asymmetric unit. The solvent content is 53% and the Matthews coefficient is 2.7 $Å^3$ Da^{-1}.

Data were collected from one crystal of each of the test cases at 100 K and at a wavelength of 0.933 Å. After indexing, these initial frames were integrated to assess the signal-to-noise ratio, mosaicity and the diffraction limit of the sample. For each crystal a highly redundant data set (300° of oscillation in 0.3° slices for SAP, 360° of oscillation in 0.5° slices for FdVI) was collected to a resolution limit normally sufficient for successful automatic model building and that was readily achievable with the available samples (2.1 Å for SAP and 2.2 Å for FdVI).

In order to assess the effects of data completeness and multiplicity on the success of the experiment, each full data set was then used to generate data sets of limited rotation (0–n°), with each range being separately processed and scaled using *DENZO* (Otwinowski & Minor, 1997) and *SCALEPACK*. The resulting data files were converted to mtz format (*COMBAT*) and analysis of data quality (see Tables 1 and 3) was carried out with the program *SCALA*. Structure factors and anomalous differences were obtained using *TRUNCATE* (all programs from the *CCP*4 suite; Collaborative Computational Project, Number 4, 1994). In order to simplify notation, the data sets will be referred to by the sample name and the rotation range included in the data set. Thus, SAP45 refers to the data set for SAP with total rotation range 45°, Fd360 the FdVI data set with total rotation range 360° *etc*.

Table 1
Data-collection statistics for SAP.

All statistics derived from a *SCALA* analysis of data processed and scaled with *DENZO*/*SCALEPACK*. Figures in parentheses are for data in the highest resolution shell (2.21–2.09 Å).

Data set	SAP300	SAP180	SAP90	SAP65	SAP45
R_{sym} (%)	5.9 (26.3)	4.8 (19.6)	3.8 (17.0)	3.4 (16.3)	3.4 (16.4)
R_{meas} (%)	6.1 (28.6)	5.2 (22.3)	4.5 (21.5)	4.3 (22.0)	4.4 (22.8)
Completeness (%)	98.1 (93.1)	97.9 (92.4)	97.0 (88.1)	96.1 (84.7)	93.0 (79.0)
Anomalous completeness (%)	97.2 (90.8)	96.8 (86.9)	93.5 (73.6)	89.2 (61.1)	79.0 (49.0)
Slope†	1.24	1.22	1.16	1.13	1.09
$\langle I\rangle/\langle\sigma(I)\rangle$	36.9 (6.4)	33.1 (6.2)	25.6 (5.0)	22.0 (4.2)	17.9 (3.6)

† Slope of anomalous difference normal probability plot produced by *SCALA*.

For each rotation range, searches to determine the substructure of the anomalously scattering atoms were carried out using direct methods as implemented in a beta test version of the program *SHELXD*-2001 (Sheldrick *et al.*, 2001;

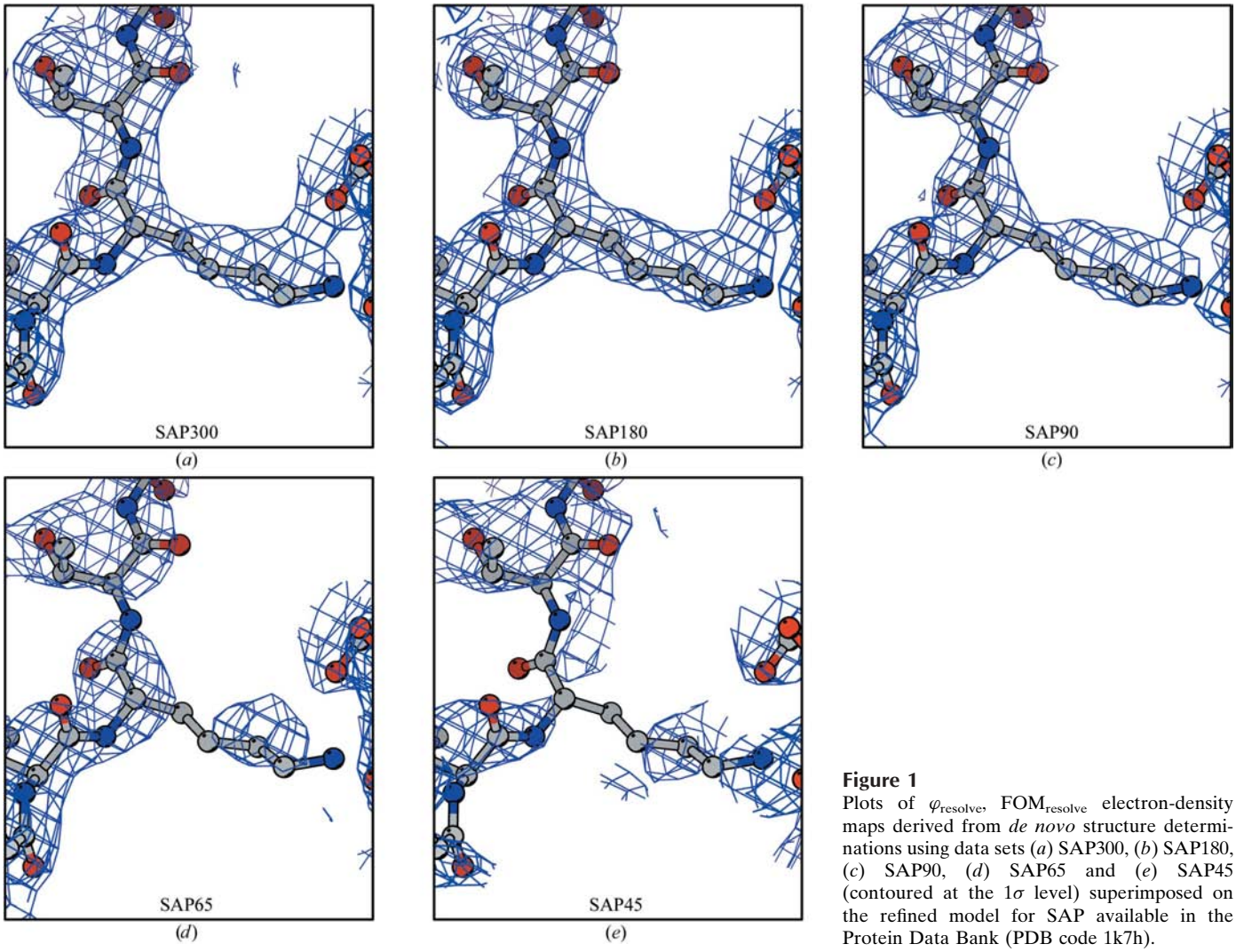

Figure 1
Plots of $\varphi_{resolve}$, $FOM_{resolve}$ electron-density maps derived from *de novo* structure determinations using data sets (*a*) SAP300, (*b*) SAP180, (*c*) SAP90, (*d*) SAP65 and (*e*) SAP45 (contoured at the 1σ level) superimposed on the refined model for SAP available in the Protein Data Bank (PDB code 1k7h).

Table 2
Results from phasing and automatic model-building procedure for SAP.

Data set	SAP300	SAP180	SAP90	SAP65	SAP45
Multiplicity	19.5	12	6	4.3	3.1
No. of sites found	6	6	6	6	6
CC_{best}	47.0	45.6	37.4	26.6	26.7
FOM†	0.21 (0.49)	0.19 (0.49)	0.15 (0.48)	0.13 (0.48)	0.11 (0.48)
Contrast difference (*SHELXE*)	0.24	0.24	0.12	0.14	0.1
$CC_{r.m.s.}$‡	0.26	0.24	0.18	0.15	0.12
MCC§	0.80	0.8	0.74	0.64	0.37
$\langle\Delta\varphi\rangle$ (°)	49.3	47.5	52.8	59.9	75.2
Main-chain residues¶	831	821	481	142	70
No. of chains in model	31	30	56	31	16
Connectivity index	0.92	0.91	0.73	0.51	0.46

† Figure of merit after phasing using the heavy-atom substructure with *SOLVE*. The values in parentheses are after density modification in *RESOLVE*. ‡ Correlation of local r.m.s. density maps before density modification taken from the program *SOLVE* and as defined in Terwilliger & Berendzen (1999*b*). § Map correlation coefficient $MCC = \langle xy\rangle\langle x\rangle\langle y\rangle/(\langle x^2\rangle - \langle x\rangle^2)^{1/2}(\langle y^2\rangle - \langle y\rangle^2)^{1/2}$, where x represents the density values from one map and y the values from the other. ¶ Number of main-chain residues built automatically using *ARP*/*wARP*.

Schneider & Sheldrick, 2002). Input command and data files were prepared using the Bruker–Nonius program *XPREP*. For all the SAP data sets input data were truncated at d_{min} = 3.0 Å, while for FdVI data were truncated at the resolution where the ratio $\langle\Delta F/F\rangle$, as indicated by *XPREP*, fell below 1.5. In all cases default search parameters were used and 100 trials were performed. The correct enantiomorph was identified implicitly using *SOLVE* (Terwilliger & Berendzen, 1999*a*) (see below) and explicitly using *SHELXE* (Sheldrick, 2002), using standard settings modified only for the solvent content of each test case.

For each data set, the program *SOLVE* was used to develop phase-probability distributions from the previously determined anomalous scatterer substructure and subsequent density modification was performed using *RESOLVE* (Terwilliger, 2000) to the diffraction limit of the data collection. Electron-density maps ($\varphi_{resolve}$, $FOM_{resolve}$) were then calculated using the *CCP*4 program FFT. Attempts at automatic model building were made using *ARP*/*wARP* in *warpNtrace* mode, using the slow option. Ten cycles were run with ten refinement cycles between rebuilding. The output of the program was checked and judging from the quality of the model produced the program was restarted with the previously determined model until the model did not improve any further. Comparisons of the ($\varphi_{resolve}$, $FOM_{resolve}$) electron-

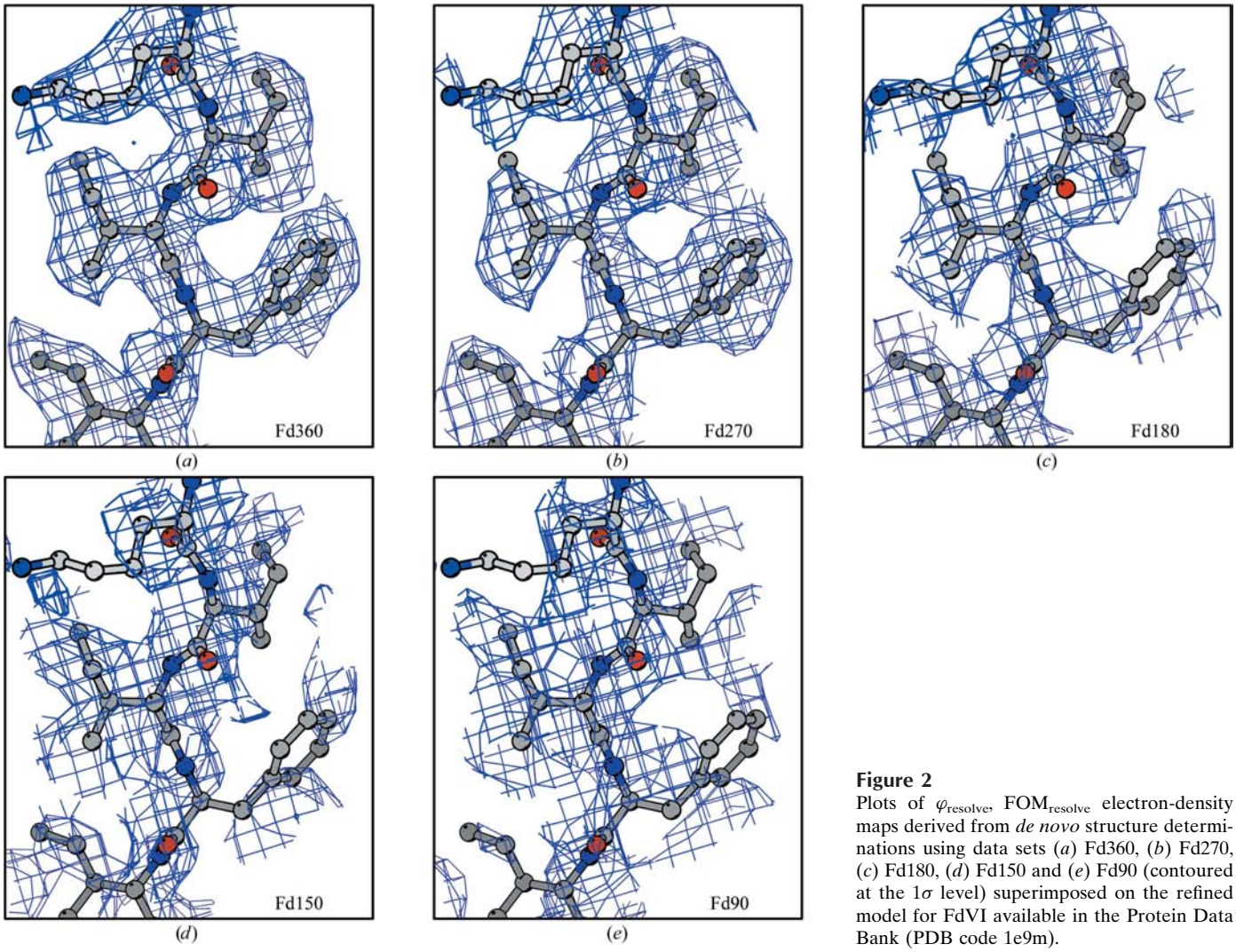

Figure 2
Plots of $\varphi_{resolve}$, $FOM_{resolve}$ electron-density maps derived from *de novo* structure determinations using data sets (*a*) Fd360, (*b*) Fd270, (*c*) Fd180, (*d*) Fd150 and (*e*) Fd90 (contoured at the 1σ level) superimposed on the refined model for FdVI available in the Protein Data Bank (PDB code 1e9m).

Table 3
Data-collection statistics for FdVI.

All statistics derived as indicated in Table 1. Values in parentheses are for data in the highest resolution shell (2.32–2.20 Å).

Data set	Fd360	Fd270	Fd180	Fd150	Fd90
R_{sym} (%)	4.2 (13.3)	3.8 (11.8)	3.5 (10.4)	3.3 (10.7)	3.2 (13.0)
R_{meas} (%)	4.5 (14.3)	4.1 (13.2)	4.1 (12.2)	4.3 (13.1)	4.0 (16.2)
Completeness (%)	99.4 (99.4)	99.4 (99.4)	99.4 (99.4)	99.2 (97.7)	87.6 (87.6)
Anomalous completeness (%)	99.7 (98.4)	99.6 (98.0)	99.6 (98.0)	99.4 (97.2)	79.5 (80.7)
Slope	1.76	1.76	1.53	1.46	1.4
$\langle I\rangle/\langle\sigma(I)\rangle$	48.5 (20.2)	43.7 (18.8)	36.4 (16.0)	33.2 (14.4)	26.5(10.5)

density maps resulting from the *SOLVE*/*RESOLVE* procedure with (F_c, φ_c) maps calculated from the coordinates associated with the relevant PDB entries (see above) were carried out using the *CCP*4 program *OVERLAPMAP*. Mean phase differences at various points in the phasing/model-building procedure were calculated using the program *PHISTATS*.

3. Results

3.1. SAP

Data-collection and processing statistics for SAP are shown in Table 1. Multiplicity ranges from about 3 for SAP45 through to 19.5 for SAP300. The data were of good quality throughout, although it is possible that the crystal was starting to show signs of radiation damage during the later stages of data collection since the multiplicity-weighted R factor, R_{meas} (Diederichs & Karplus, 1997), for SAP300 and SAP180 had increased compared with that for the other lower multiplicity data sets. The normal probability plot (Howell & Smith, 1992; Table 1) of the anomalous differences shows a slope of 1.24 for the SAP300 data set. This indicates a detectable intensity difference within the Bijvoet pairs and is indicative of the presence of some anomalous signal in the data. For each of the subsequent SAP data sets this falls off gradually until for SAP45 a slope for the normal probability plot of 1.09 suggests little or no anomalous signal in the data. A more detailed analysis of the anomalous differences using *XPREP* showed that for all the SAP data sets there was significant anomalous signal [*i.e.* $\langle\Delta F/\sigma(\Delta F)\rangle > 1.5$] to a resolution of around 4.5 Å. At higher resolution $\langle\Delta F/\sigma(\Delta F)\rangle$ dropped off, with values being generally lower for the low-multiplicity data sets (SAP45, SAP65) than for those with higher multiplicity (SAP180, SAP300).

Using the protocols described above, it was possible to determine the positions of the anomalous scatterers in the asymmetric unit for all the SAP data sets. In each case, the best solution chosen by *SHELXD* corresponded to the correct solution (or its enantiomorph). It is noticable, however, that the degree of confidence that the solution was correct increased with the multiplicity of the data sets. The correlation between E_{obs} and E_{calc} for the best solutions (CC_{best}) increased markedly from SAP45 to SAP300 (Table 2), with CC_{best} for SAP180 and SAP300 unambiguously indicating the solutions for these data sets to be correct. Enatiomorphs were correctly identified by *SOLVE* and were revealed by examination of the contrast in the solvent-flattened map as reported by *SHELXE* when run in both hands (Table 2).

For each SAP data set, phasing was performed and electron-density maps (Fig. 1) were calculated as described above. The electron-density maps obtained from data sets with larger angular ranges (SAP300, SAP180, SAP90) are readily interpretable by eye, those with smaller angular ranges (SAP65, SAP45) less so. The main effect of reduced data multiplicity is a loss of connectivity of main-chain electron density and loss of electron-density definition for the side chains. This degradation in the quality of electron-density maps is also evident from an examination of the correlation coefficients between the ($\varphi_{resolve}$, $FOM_{resolve}$) maps and the map obtained from the final refined model. The mean phase differences, $\langle\Delta\varphi\rangle$, observed between those derived experimentally and those calculated from the final refined model (Table 2) also reflect the degradation in map quality. For SAP300, SAP180 and SAP90, map correlation coefficients (MCC) are significantly above 0.7 and mean phase errors significantly below 60°, while for SAP65 and (particularly) SAP45 the values of both MCC and mean phase error suggest that the resulting maps would be difficult to interpret.

The indications regarding map/phase quality were confirmed by attempts to automatically trace models using *ARP*/*wARP*. Almost the entire protein main chain could easily be traced for SAP300 and SAP180 (Table 2), while for SAP90 approximately 50% of the protein main chain was built and the result would have served as an excellent starting point for further manual model building. Not surprisingly, for both SAP65 and SAP45 automatic model building struggled to produce models with high connectivity, probably because of the loss of connectivity in the main-chain electron density.

3.2. FdVI

The FdVI data sets are of high quality (Table 3) but owing to the lower space-group symmetry the multiplicity of the data sets is in general lower, reaching a maximum of 14.2 for Fd360. Examination of normal probability plots of the anomalous differences for each of the data sets (Table 3) shows slopes with a significant deviation from 1.0 in all cases. Surprisingly, given the smaller anomalous signal expected, the slopes are steeper than for SAP but, as for SAP, this steepness of slope increases with multiplicity. In keeping with observations regarding the steepness of the slopes of normal probability plots, detailed analysis of the anomalous differences with *XPREP* showed significant differences between the magnitudes of Bijvoet pairs to a resolution of around 2.9 Å for the higher multiplicity data sets and to around 3.1 Å for Fd90, where the multiplicity is only 4.1. Searches for the heavy-atom substructure using *SHELXD* appeared to be successful for all rotation ranges tested, as in each case the list of peak heights produced by the program clearly indicates two potential Fe^{2+} sites. As for SAP, CC_{best} (Table 4) is increased for the higher multiplicity data sets. Interestingly, only for the Fd360 and

Table 4
Results of phasing and automatic model-building procedure for FdVI.

All statistics defined as in Table 2. No chain tracing was attempted for Fd150 and Fd90.

Data set	Fd360	Fd270	Fd180	Fd150	Fd90
No. of sites found	2	2	2	2	2
CC_{best}	35.1	35.6	30.9	27	30.2
FOM	0.27 (0.55)	0.25 (0.54)	0.22 (0.52)	0.17 (0.50)	0.19 (0.50)
Contrast difference (*SHEXLE*)	0.07	0.2	0.11	0.14	0.01
$CC_{r.m.s.}$	0.18	0.17	0.16	0.18	0.14
MCC	0.78	0.7	0.63	0.56	0.31
$\langle\Delta\varphi\rangle$ (°)	44.1	51.5	57.4	62.8	73.2
Main-chain residues	105	68	17		
No. of chains in model	1	8	4		
Connectivity index	0.98	0.79	0.6		

Fd270 data sets is the Fe—Fe distance the same (3.6 Å) as in the final refined structure of FdVI. For Fd180 and Fd150 this distance is 3.8 Å, while for Fd90 the Fe—Fe distance for the substructure atoms as determined by *SHELXD* increases markedly to 4.6 Å. This demonstrates that in this case increased multiplicity plays a crucial role in the determination of the correct heavy-atom substructure. The correct enantiomorph could be identified from inspection of the *SHELXE* statistics. From Fd150 onwards a clear distinction between hands can be observed (Table 4); the effect is less marked than in the case of SAP. Indeed, the discrimination reduces for Fd360, perhaps indicating the onset of radiation damage.

Not surprisingly, the degree of correctness of the heavy-atom positions used to develop phases then propagates through the whole phasing and model-building procedure with predictable consequences (Fig. 2, Table 4). The ($\varphi_{resolve}$, $FOM_{resolve}$) electron-density maps are of excellent quality for both Fd360 and Fd270 and would easily allow manual building of almost the entire molecule. Mean phase differences (compared with the final refined model) of 57.4 and 62.8° for Fd180 and Fd150 indicate that interpretation of the resulting electron-density maps for these data sets would be problematical, while a value of 73.2° for Fd90 suggests the electron-density map here would be of limited use in structure determination. These impressions are reinforced by our attempts at automatic model building for FdVI using *ARP*/*wARP* (Table 4). Only for Fd360 (the entire polypeptide chain) and Fd270 (68 residues out of 106) could useful models be built.

4. Discussion

The aims of the work we describe here were twofold: (i) to demonstrate the potential of fixed-wavelength beamlines for *de novo* macromolecular structure determination using the remote SAD technique and (ii) to derive a standard set of protocols (coupled with suitable indicators) that would facilitate routine automatic structure dermination using this method.

4.1. Applicability

The two systems studied are metalloproteins which at the wavelength of our experiments yield rather small (<2%) anomalous signals from which to develop experimental phases. For both systems, we have been successful in producing experimental phases that have allowed the automatic building of complete models. We are thus confident that it should be possible to solve many macromolecular structures using the SAD method when wavelength tunability is not available or the absorption edge of interest is for some reason not accessible. It should be emphasized that around 30% of all proteins are metalloproteins and a large number of them will have anomalous signals of a similar level to those exploited here. Since metalloproteins have well ordered metal centres as a rule, this class of macromolecules should be readily amenable to this technique. Selenomethionine derivatives are equally attractive prospects since at a wavelength of 0.933 Å we would expect an anomalous signal of 4.0% for the average protein, which contains around one Met per 42 amino-acid residues. Other possible targets for the technique are crystals in which proteins have been derivatized using compounds containing third-row transition metals. For such derivatives containing one fully occupied heavy-atom site per 300 amino acids one would expect for data collected at 0.933 Å anomalous signals of around 4–5% for the series Ta–Ir and around 3.5–4% for the series Pt–Pb.

The only common heavy-atom derivatives that could not be easilly targeted at 0.933 Å wavelength are those resulting from soaking in either NaBr (Dauter *et al.*, 2000, 2001) or RbBr (Korolev *et al.*, 2001). In these cases a wavelength of 0.933 Å is on the long-wavelength side of the *K* absorption edge of both Br and Rb.

At first sight, the major problem facing the routine use of this technique in macromolecular structure determination is the apparent need for high-multiplicity data sets with the consequent risk of increased radiation damage to samples. However, for both the systems studied the amount of data required to yield almost complete automatic structure solution (180° for SAP and 360° for FdVI) is not significantly higher than that one would collect in a three-wavelength MAD experiment. In fact, the prospect of reduced radiation damage could be an advantage of avoiding the absorption edge. The smaller absorption cross-section for many anomalously scattering atoms at 0.933 Å wavelength then compared with that at the appropriate *K* or *L* absorption edges, particularly in the presence of so-called 'white lines', would reduce localized changes in the vicinity of the anomalous scatterer positions (see, for example, Rice *et al.*, 2000), resulting in higher data quality and unabated anomalous signal.

4.2. Prospects for automation

The data-collection method described is particularly amenable to automation because the essential information is acquired in the most straightforward and simple fashion possible. In particular, the removal of the need to perform and analyse XANES absorption-edge scans greatly simplifies the

experimental protocol. Data collection would thus be more robust and the steady improvement of statistics as the data collection proceeds offers the possibility to develop software systems that generate 'feedback' such that only sufficient data are collected to achieve the desired goal. Possible points of feedback concern (i) the ability to determine the anomalous scatterer positions, (ii) the determination of the correct enantiomorph for the substructure and (iii) the quality of the experimentally phased electron-density maps at various points during the data collection.

We noticed during our analyses that approximate positions for the atoms comprising the anomalous scatterer substructure could be determined with relatively little data (Tables 2 and 4). For both test cases, all substructure atoms could be found

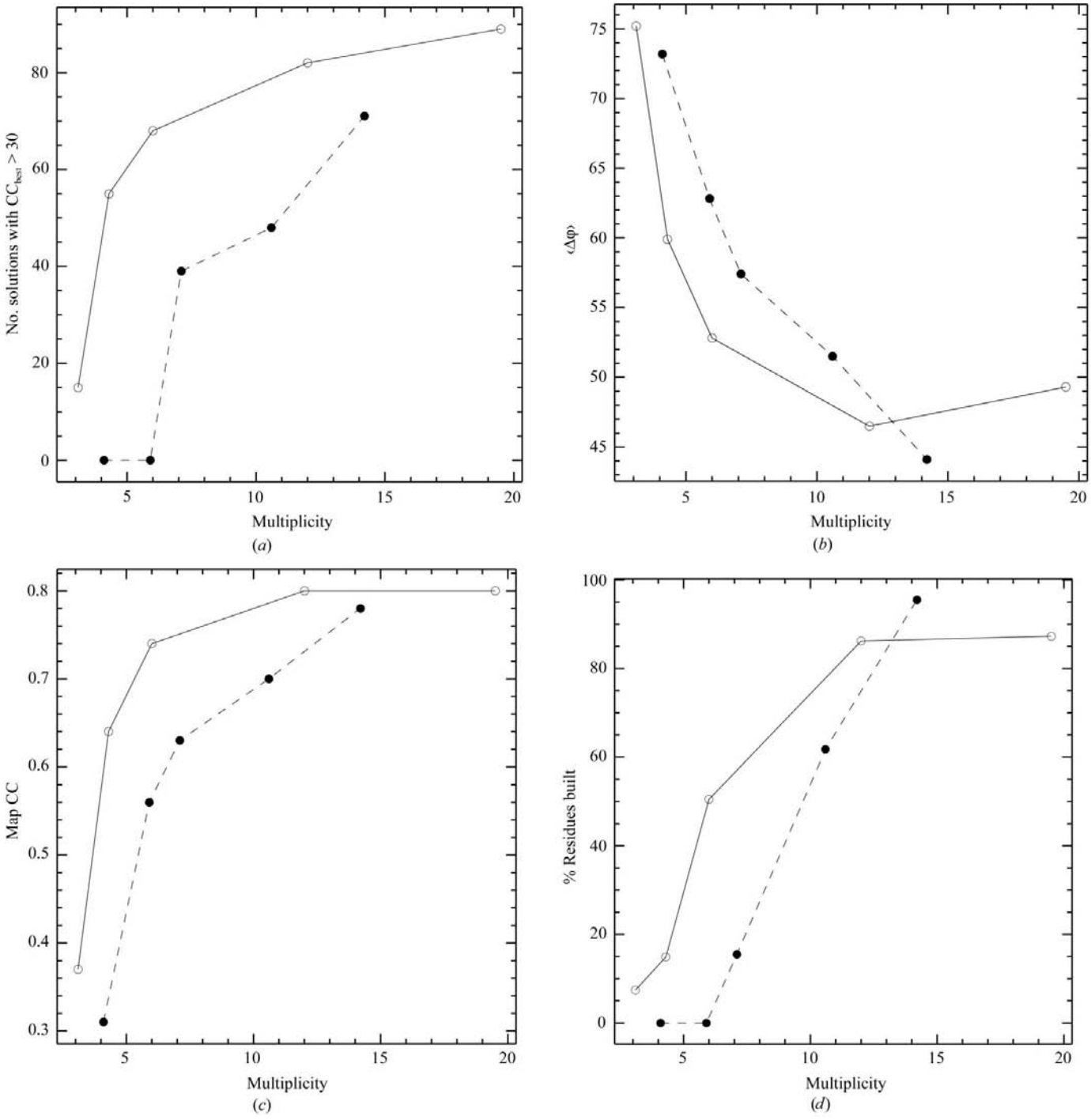

Figure 3
Plots *versus* multiplicity for both SAP (open circles, solid lines) and FdVI (solid circles, broken lines) of (*a*) the number of substructure solutions with CC_{best} greater than 30, (*b*) the correlation of experimentally phased electron-density maps with those calculated from final models, (*c*) mean phase differences, $\langle\Delta\varphi\rangle$, and (*d*) the percentage of residues for which the main-chain atoms could be built automatically by *ARP*/*wARP*.

using minimal data sets for which the overall completeness is around 90%, the anomalous completeness is around 80% and the multiplicity is between 3 and 4. However, in Fig. 3(*a*) we plot as a function of multiplicity the number of substructure solutions found by *SHELXD* with CC_{best} greater than 30. A clear change in gradient is visible when data sets have a multiplicity of about 6 (anomalous completeness >90%). This indicates that it is desirable to initially collect rather limited amounts of data. Thus, an automatic data-collection strategy should aim at producing an initial data set of multiplicity 6 as rapidly as possible and these data should be used to determine the positions of the anomalous scatterers. Initial trials could realistically commence as soon as the multiplicity is greater than 3. The determination of the correct enantiomorph could begin at this point using either *SHELXE* (run with both hands examined) or by the automatic procedures implemented in *SOLVE*. Only if these steps are successful would further data be automatically collected. Samples (or data sets) failing this test may require manual examination.

With the anomalous scatterer substructure accurately determined, the question then arises as to how much data one really needs for successful structure determination and for automatic phasing and model-building protocols to perfom efficiently. In Figs. 3(*b*), 3(*c*) and 3(*d*) we plot as a function of the multiplicity of the data sets $\langle\Delta\varphi\rangle$, MCC and the percentage of residues for which the main-chain atoms can be built automatically by *ARP*/*wARP*. As can be seen, $\langle\Delta\varphi\rangle$ and MCC decrease and increase, respectively, as the multiplicity of the data used in the process increases. However, both for FdVI and (particularly) SAP, at a certain point the rates of change in $\langle\Delta\varphi\rangle$ and MCC appear to be flattening off and, more importantly, the number of residues automatically built using the *ARP*/*wARP* procedure does not increase dramatically. Indeed, between SAP180 and SAP300 $\langle\Delta\varphi\rangle$ actually increases, which may be an indication of the onset of radiation damage. There appears therefore to be a point in the data collections where the 'law of diminishing returns' starts to apply and further data collection is not necessary (or indeed becomes counterproductive). It is clear that this point (a multiplicity of around 12 for these test cases) will not be the same for all systems and a truly automatic process for high-throughput SAD structure determination would need to identify this point properly in order that beam time is not wasted collecting unproductive or even deleterious data.

How to determine the optimal data set for structure determination is not obvious, as indicators such as $\langle\Delta\varphi\rangle$ and MCC used above would not be available during *de novo* structure determinations. One possibility would be to monitor the quality of experimentally phased electron-density maps at various stages of the data collection. Estimation of electron-density map quality is already an integral part of the *SOLVE* automatic phasing program (Terwilliger & Berendzen, 1999*b*) and *SHELXE*. Tables 2 and 4 give the correlation of r.m.s. density in neighbouring boxes ($CC_{r.m.s.}$) as calculated by *SOLVE* for the experimental electron-density maps without solvent-flattening derived at each stage of the data collections for both SAP and FdVI. As can be seen, $CC_{r.m.s.}$ generally increases as $\langle\Delta\varphi\rangle$ decreases. It could thus be an excellent monitor of improvement (or otherwise) of electron-density map quality in any automatic system for SAD structure determination and it would seem feasible that this or some other similar indicator could be used to determine whether more data are required for successful structure determination. The use of $CC_{r.m.s.}$ would be particularly powerful when used in conjunction with the information from density-modification techniques.

A plausible strategy for automatic data collection and structure determination in routine cases would thus be to first collect a minimal data set for substructure solution (multiplicity >3) and then, while continuing with the data collection, to attempt substructure solution *via* direct or other methods. If this appeared unpromising when the multiplicity was 6 or greater, based on analysis of peak heights and CC_{best}, data collection would be terminated and the sample stored. Otherwise, further data would be collected and electron-density maps calculated at suitable points. During this process there would be no need for interruption of data collection. Heavy-atom substructures can be redetermined as more data becomes available; this would provide a useful check for consistency. Once no improvement in the experimentally phased electron-density maps is seen, data collection could be halted, density modification carried out and automatic model building attempted. If automatic model building cannot produce a model with connectivity index greater than 0.90, it could well be worthwhile remounting the crystal and collecting further (appropriate) data. One concern during the accumulation of sufficient data will be the onset of radiation damage. A robust way of testing this would be the (artificial) separation of data into low-multiplicity data sets. Merging statistics between these data sets should be consistent with the data quality of the individual data sets and deviations from this would indicate that data collection should cease.

If interruption of data collection was required, as might be the case for example for systems with very large cells, for very complex heavy-atom substructures or if the computation time is long compared with that for data collection, the development of automated sample-changing systems provides an excellent means of recovering potentially unproductive beam time as crystals could be stored (and other systems used for data collection) until such time as the automatic structure-determination procedure was ready to recommence.

5. Conclusions

In this paper, it is shown that intensity data collected on fixed-wavelength synchrotron beamlines has the potential to prove extremely useful in the automatic *de novo* phasing of very large numbers of macromolecular structures using SAD. At the relatively short wavelengths normally available (that are usually rather remote from the absorption edges of interest), the anomalous signals are rather small but may be usefully exploited. As has already been remarked by others in slightly different contexts (see, for example, Debreczeni *et al.*, 2003; Dauter & Adamiak, 2001), increased data multiplicity is

crucial to the success of phasing undertaken in this manner. However, it is clear that a number of indicators (for solution of heavy-atom substructure and for evaluation of the quality of non-solvent-flattened electron-density maps) are available which, when monitored at pertinent points during experiments, should help to limit the amount of data required and which provide a basis for the development of software tools to monitor the progress of the experiment. The simplification of the data-collection protocol offered by this technique means that automation of the structure-determination process should be more straightforward. The remote SAD method could therefore play an important part in the high-throughput completely automatic procedures currently being planned for structural genomics initiatives.

The authors would like to thank Professor Edward Hough for providing the crystals of SAP and Dr Raimond Ravelli for discussions and critical reading of the manuscript.

References

Abola, E., Kuhn, P., Earnest, T. & Stevens, R. C. (2000). *Nature Struct. Biol.* **7**(*Suppl.*), 973–977.

Armengaud, J., Sainz, G., Jouanneau, Y. & Sieker, L. C. (2001). *Acta Cryst.* D**57**, 301–303.

Arzt, S. *et al.* (2005). In the press.

Brodersen, D. E., de La Fortelle, E., Vonrhein, C., Bricogne, G., Nyborg, J. & Kjeldgaard, M. (2000). *Acta Cryst.* D**56**, 431–441.

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

Cianci, M., Rizkallah, P. J., Olczak, A., Raftery, J., Chayen, N. E., Zagalsky, P. F. & Helliwell, J. R. (2001). *Acta Cryst.* D**57**, 1219–1229.

Dauter, Z. & Adamiak, D. A. (2001). *Acta Cryst.* D**57**, 990–995.

Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.

Dauter, Z., Dauter, M. & Dodson, E. (2002). *Acta Cryst.* D**58**, 494–506.

Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Acta Cryst.* D**56**, 232–237.

Dauter, Z., Li, M. & Wlodawer, A. (2001). *Acta Cryst.* D**57**, 239–249.

De Backer, M., McSweeney, S., Rasmussen, H. B., Riize, B. W., Lindley, P. & Hough, E. (2002). *J. Mol. Biol.* **318**, 1265–1274.

Debreczeni, J. E., Bunkoczi, G., Ma, Q., Blaser, H. & Sheldrick, G. M. (2003). *Acta Cryst.* D**59**, 688–696.

Delft, F. von, Inoue, T., Saldanha, S. A., Ottenhof, H. H., Schmitzberger, F., Birch, L. M., Dhanaraj, V., Witty, M., Smith, A. G., Blundell, T. L. & Abel, C. (2003). *Structure*, **11**, 985–996.

Diederichs, K. & Karplus, P. A. (1997). *Nature Struct. Biol.* **4**, 269–275.

Gordon, E. J., Leonard, G. A., McSweeney, S. & Zagalski, P. F. (2001). *Acta Cryst.* D**57**, 1230–1237.

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

Harvey, I., Hao, Q., Duke, E. M., Ingledew, W. J. & Hasnain, S. S. (1998). *Acta Cryst.* D**54**, 629–635.

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

Hendrickson, W. A. (1991). *Science*, **254**, 51–58.

Hendrickson, W. A. (1999). *J. Synchrotron Rad.* **6**, 845–851.

Hendrickson, W. A. & Ogata, C. M. (1997). *Methods Enzymol.* **276**, 494–523.

Hendrickson, W. A., Smith, J. L. & Sheriff, S. (1985). *Methods Enzymol.* **115**, 41–55.

Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.

Howell, P. L. & Smith, G. D. (1992). *J. Appl. Cryst.* **25**, 81–86.

Korolev, S., Dementieva, I., Sanishvili, R., Minor, W., Otwinowski, Z. & Joachimiak, A. (2001). *Acta Cryst.* D**57**, 1008–1012.

Langs, D. A., Blessing, R. H. & Guo, D. (1999). *Acta Cryst.* A**55**, 755–760.

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

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Micossi, E., Hunter, W. N. & Leonard, G. A. (2002). *Acta Cryst.* D**58**, 21–28.

Nilsen, I. W., Overbo, K. & Olsen, R. L. (2001). *Comput. Biochem. Physiol. B*, **129**, 853–861.

Olsen, R., Overbo, K. & Myrnes, B. (1991). *Comput. Biochem. Physiol. B*, **99**, 755–761.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

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

Ramagopal, U. A., Dauter, M. & Dauter, Z. (2003*a*). *Acta Cryst.* D**59**, 868–875.

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

Ravelli, R. G. B. & McSweeney, S. M. (2000). *Structure Fold. Des.* **8**, 315–328.

Rice, L. M., Earnest, T. N. & Brunger, A. T. (2000). *Acta Cryst.* D**56**, 1413–1420.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.

Sheldrick, G. M. (2002). *Z. Kristallogr.* **217**, 644–650.

Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R. & Usón, I. (2001). *International Tables for Crystallography*, Vol. *F*, edited by E. Arnold & M. Rossmann, pp. 333–351. Dordrecht: Kluwer Academic Publishers.

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

Terwilliger, T. C. & Berendzen, J. (1999*a*). *Acta Cryst.* D**55**, 849–861.

Terwilliger, T. C. & Berendzen, J. (1999*b*). *Acta Cryst.* D**55**, 1872–1877.

Walsh, M. A., Dementieva, I., Evans, G., Sanishvili, R. & Joachimiak, A. (1999). *Acta Cryst.* D**55**, 1168–1173.

Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–111.

Weiss, M. S., Sicker, T., Djinovic Carugo, K. & Hilgenfeld, R. (2001). *Acta Cryst.* D**57**, 689–695.