Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

In-house measurement of the sulfur anomalous signal and its use for phasing

Judit É. Debreczeni, Gábor Bunkóczi, Qingjun Ma, Heiko Blaser and George M. Sheldrick*

Lehrstuhl für Strukturchemie, Georg-August
Universität, Tammannstrasse 4,
D-37077 Göttingen, Germany

Correspondence e-mail:
gsheldr@shelx.uni-ac.gwdg.de

Received 4 December 2002
Accepted 30 January 2003

Five test structures (orthorhombic and trigonal trypsin, cubic and rhombohedral insulin and thaumatin) have been solved by the SAD (single-wavelength anomalous diffraction) method using highly redundant data collected at 100 K with a CCD detector, rotating-anode generator and three-circle goniometer. The very weak anomalous scattering (primarily from sulfur) was sufficient to locate all the anomalous scatterers using the integrated direct and Patterson methods in *SHELXD*. These positions and occupancies were used without further refinement to estimate phases that were extended to native (in-house) resolution by the *sphere of influence* algorithm in *SHELXE*. The final map correlation coefficients relative to the anisotropically refined structures were in the range 0.81–0.97. The use of highly redundant medium-resolution laboratory data for sulfur-SAD phasing combined with high-resolution synchrotron native data for phase expansion and structure refinement clearly has considerable potential.

1. Introduction

The method of choice for experimental protein phase determination is currently the MAD (multiple-wavelength anomalous diffraction) technique of collecting data at two or more wavelengths close to the absorption edge of an element such as selenium (Hendrickson, 1991; Smith, 1998). Although this method can give high-quality phases, it usually involves the additional preparation of selenomethionine derivatives and always requires access to a synchrotron. The SAD (single-wavelength anomalous diffraction) method using the S atoms already present in the native protein is potentially an attractive alternative. After the pioneering work by Hendrickson & Teeter (1981) on crambin and the development of computational methods for resolving the twofold phase ambiguity inherent in this approach (Wang, 1985), little progress was made until recent developments in cryocrystallography, area detectors and software made sulfur-SAD phasing viable. The SAD approach shares with MAD the advantage that all measurements can be made on the same crystal so that lack of isomorphism does not pose problems, but has the further advantage that a single data set at a single wavelength suffices. Most sulfur-SAD applications have involved synchrotron radiation, sometimes at similar wavelengths to laboratory Cu $K\alpha$ sources (Dauter *et al.*, 1999), but usually at longer wavelengths where f'' for sulfur is larger (Liu *et al.*, 2000; Gordon *et al.*, 2001; Weiss *et al.*, 2001; Brown *et al.*, 2002; Micossi *et al.*, 2002). Orthorhombic trypsin, containing the anomalous scatterers calcium and sulfur, has been used as a test structure for in-house SAD phasing with Cu $K\alpha$ radiation by Yang & Pflugrath (2001) and very recently Lemke *et al.*

Table 1
Crystallization and cryoprotectant conditions (for references see §2.1).

	Orthorhombic trypsin	Trigonal trypsin	Cubic insulin	Rhombohedral insulin	Thaumatin
Crystallization conditions	0.1 *M* Tris–HCl pH 7.4, 0.08 *M* $(NH_4)_2SO_4$, 22% PEG 6000, 16% glycerol	0.1 *M* Tris–HCl pH 7.4, 0.08 *M* $(NH_4)_2SO_4$, 10% PEG 6000, 15% ethylene glycol	0.4 *M* Na_3PO_4/ Na_2HPO_4 pH 10.4, 0.01 *M* Na_3EDTA	0.05 *M* trisodium citrate/ citric acid pH 6.5, 14% acetone, 1.40 *M* NaCl	0.05 *M* Na_2ADA/ NaADA pH 6.8, 0.6 *M* K,Na tartrate
Cryoprotectant	Not required	Not required	30% glycerol	10% PEG 400, 5% glycerol	30% glycerol

Table 2
Data-collection statistics.

Values in parentheses are for the last resolution shell.

	Orthorhombic trypsin	Trigonal trypsin	Cubic insulin	Rhombohedral insulin	Thaumatin
Space group	$P2_12_12_1$	$P3_121$	$I2_13$	$R3$	$P4_12_12$
Unit-cell parameters (Å)	a = 53.90, b = 56.96, c = 66.06	$a = b$ = 54.74, c = 106.79	$a = b = c$ = 77.73	$a = b$ = 79.92, c = 36.45	$a = b$ = 57.90, c = 149.14
No. of reflections	1071612	1054240	853180	296346	736989
No. of unique reflections	64587	43523	19009	13823	37892
Resolution (Å)	1.19 (1.30–1.19)	1.32 (1.45–1.32)	1.30 (1.40–1.30)	1.50 (1.60–1.50)	1.55 (1.65–1.55)
Completeness (%)	98.2 (92.0)	98.5 (93.6)	98.8 (93.7)	98.8 (93.6)	99.9 (100.0)
Redundancy	16.3 (8.8)	23.9 (12.5)	44.3 (20.1)	21.2 (3.8)	19.4 (12.0)
$\langle I/\sigma(I)\rangle$	33.6 (5.2)	40.1 (5.8)	50.0 (4.7)	54.3 (5.0)	62.7 (18.5)
$R_{\rm int} = \sum\lvert I - \langle I\rangle\rvert / \sum I$	0.050 (0.407)	0.045 (0.428)	0.044 (0.706)	0.033 (0.245)	0.032 (0.154)
$R_\sigma = \sum\sigma(I)/\sum I$	0.0231	0.0161	0.0119	0.0114	0.0111
Anomalous scatterers	14 S, 1 Ca^{2+}	14 S, 1 Ca^{2+}	6 S	12 S, 2 Zn^{2+}, 3 Cl^-	17 S
$\sum\lvert I_+ - I_-\rvert/\sum(I_+ - I_-)$ (ideal data)	0.0129	0.0113	0.0126	0.0123	0.0100
Total time (d)	2.53	5.34	4.71	2.88	2.94

(2002) reported the solution of a known structure containing 19 S atoms in 455 amino acids by in-house sulfur-SAD phasing.

In this paper, we report details of the successful in-house Cu *Kα* sulfur-SAD phasing of five test structures: the orthorhombic and trigonal forms of trypsin, the cubic and rhombohedral forms of insulin and the tetragonal form of thaumatin. For cubic insulin and thaumatin the only significant anomalous scatterers present are sulfur and in all five cases the anomalous scattering is extremely weak, requiring accurate data. On the basis of the experience gained with these test structures we have recently been able to solve two unknown protein structures containing no atom heavier than sulfur by the same approach (Debreczeni, Bunkóczi *et al.*, 2003; Debreczeni, Girmann *et al.*, 2003).

2. Materials and methods

2.1. Crystallization and data collection

Bovine trypsin, bovine insulin and thaumatin were purchased from Sigma and used without further purification. Crystallization of the samples was performed by minor modifications of previously established protocols (Dodson *et al.*, 1978; Ko *et al.*, 1994; Renatus *et al.*, 1998; Schroder Leiros *et al.*, 2001; Smith *et al.*, 2001). Before measurement, crystals were first soaked in a cryoprotectant solution (where necessary) and then mounted in a loop using a cold nitrogen stream at 100 K; for the two forms of trypsin no cryoprotectant was required because the crystallization solution froze to a glass. Crystallization and cryoprotectant conditions are shown in Table 1. Data sets were collected using φ and ω scans (for thaumatin, ω-scans only) with the Bruker Nonius (2002) programs *PROTEUM* and *SMART* in fine-slice mode (0.2° steps) using a Bruker rotating-anode generator, Cu *Kα* radiation, Osmic focusing mirrors and a Bruker *SMART*6000 4K CCD detector at a distance of 4.5 or 5.0 cm except for thaumatin (12.0 cm), usually to the diffraction limit of the crystal in question. The lower completeness in the outermost shells is primarily an artefact caused by the highest resolution reflections in the corners of the square detector.

The time per frame varied from 5 to 120 s and the total data-collection time from 2.5 to 5.3 d. Most of this time was spent obtaining good-quality high-resolution data so that the phase expansion could be performed entirely using laboratory data. Had we employed our more usual strategy for unknown structures of collecting a high-resolution data set at a synchrotron and high-redundancy low-resolution Cu *Kα* laboratory data for phasing, the laboratory data collection would have taken less than 2 d in all cases.

2.2. Data processing, substructure solution, phasing and phase improvement

Data sets were indexed with the programs *PROTEUM* or *SMART* (Bruker Nonius, 2002), integrated with *SAINT* using thin-slice algorithms and scaled with *SADABS*. The program *XPREP* was used to calculate quality indicators, anomalous

Table 3
Refinement statistics.

	Orthorhombic trypsin	Trigonal trypsin	Cubic insulin	Rhombohedral insulin	Thaumatin
Resolution range (Å)	66.1–1.19	53.4–1.32	31.7–1.30	40.0–1.50	39.5–1.51
R factor [$F > 4\sigma(F)$/all data]	0.099/0.114	0.124/0.136	0.126/0.142	0.212/0.224	0.159/0.163
$R_{\rm free}$ [$F > 4\sigma(F)$/all data]	0.132/0.148	0.176/0.190	0.146/0.1695	0.261/0.276	0.193/0.199
No. of residues	223	223	51	100	207
No. of non-H atoms	1994	1934	454	808	1819
No. of solvent atoms	342	270	74	48	333
R.m.s.d. from ideal geometry					
Bond lengths (Å)	0.007	0.011	0.014	0.008	0.011
Angle distances (Å)	0.022	0.029	0.033	0.025	0.027
Ramachandran plot					
Residues in allowed region (%)	87.8	86.2	90.7	92.9	88.2
Residues in additionally allowed region (%)	12.2	13.8	9.3	7.1	11.8
Residues in unallowed region (%)	0	0	0	0	0

differences and for merging. Data statistics are listed in Table 2.

Anomalous scatterers were found using the dual-space recycling algorithm implemented in *SHELXD* (Sheldrick *et al.*, 2001; Usón & Sheldrick, 1999; Schneider & Sheldrick, 2002). Patterson seeding was used in all cases, resulting in an appreciable improvement in the success rate. Usually, 20 cycles of dual-space recycling were performed in which tangent-formula expansion alternated with a peak search and structure-factor calculation. This procedure is fast because only the largest normalized structure factors (usually $E > 1.5$) are employed; this is particularly appropriate for location of the substructure sites using anomalous differences because the E values calculated from the smaller anomalous differences are unreliable anyway; the latter represent only lower limits on the true substructure structure factors.

The *SHELXD* solution with the highest correlation coefficient was input directly into *SHELXE* (Sheldrick, 2002) for phase calculation and improvement without further refinement of the atom positions or occupancies. Although *SHELXE* was designed for fast robust high-throughput phasing, if the solvent content is high (>0.6) or the data extend to high resolution (<1.5 Å) the *sphere of influence* density modification incorporating a fuzzy solvent boundary is capable of generating high-quality maps, as illustrated by these five test structures. In this density-modification algorithm, the environment of a pixel in the electron-density map, as represented by the variance of the density on the surface of a sphere of radius 2.42 Å about the pixel in question, is used to decide whether the pixel is more likely to be in the protein or in the solvent region. Since 2.42 Å is a typical 1,3-interatomic distance in proteins, this is a simple and general way of introducing a little chemical knowledge.

2.3. Model building, refinement and phase errors

The final *SHELXE* phases were fed into *ARP*/*wARP* (Perrakis *et al.*, 1999) for main-chain tracing and side-chain docking. This initial model was completed by hand using the program *XtalView* (McRee, 1999) and refined with *SHELXL* (Sheldrick & Schneider, 1997). σ_A-weighted $2mF_o - DF_c$ and $F_o - F_c$ type maps were displayed for the identification of disordered components and the addition of water. Throughout the refinement, bond-length, bond-angle, chiral-volume and planarity restraints were applied to appropriate atoms. Anisotropic displacement parameters were refined for all non-H atoms (when justified by a significant drop in $R_{\rm free}$), with suitable rigid-bond, similarity and (for solvent waters) approximately isotropic restraints; heavier atoms (sulfur, chlorine, calcium, zinc) were refined anisotropically in all cases. Solvent atoms were added using *SHELXWAT* (Sheldrick & Schneider, 1997) and by hand. H atoms were included in the last stages of refinement. These refined models were used as reference structures for phase error and map correlation coefficient calculation by the method of Lunin & Woolfson (1993) implemented in a new pre-release version of *SHELXPRO* (Sheldrick & Schneider, 1997). Refinement parameters and statistics are presented in Table 3. The refined models were also used to calculate the ratio $\sum|I_+ - I_-|/\sum(I_+ + I_-)$ (summed over all Friedel pairs), using structure factors calculated from complex scattering factors, to estimate the strength of the anomalous scattering; since this estimate (shown in Table 2) takes disorder and different thermal displacement parameters into account, it should be more realistic than estimates based on relative scattering factors alone.

3. Results

The following protocol was employed for the solution of all five test structures by the SAD method. First, the data collection was organized so that the data resulting from φ and ω scans were of approximately equal quality, completeness and redundancy. After measurement and integration, scaling was performed separately for the two different types of scan in order to preserve their independence. The (signed) anomalous differences $F_+ - F_-$ were calculated for each scan type and the correlation coefficient between them was evaluated as a function of the resolution (Fig. 1*a*). This correlation coefficient is defined as

$$\mathrm{CC} = \frac{(N\sum\Delta_1\Delta_2 - \sum\Delta_1\sum\Delta_2)}{\{[N\sum\Delta_1^2 - (\sum\Delta_1)^2][N\sum\Delta_2^2 - (\sum\Delta_2)^2]\}^{1/2}},$$

where Δ_1 is the signed anomalous difference $F_+ - F_-$ for a Friedel pair in data set 1 *etc.* and N is the number of reflections in a resolution shell. Previous experience with MAD data processing (Schneider & Sheldrick, 2002), in which the correlation coefficient was calculated between anomalous differences at two different wavelengths, indicates that for the location of heavy atoms with *SHELXD* it is best to truncate the data to the resolution where this correlation coefficient falls below 30%. For *SHELXE* there was no need to truncate the anomalous difference data because the resolution-dependent weighting scheme serves the same purpose in a more flexible way.

After applying this test, all the raw data from the φ and ω scans were scaled together using *SADABS* so that the higher redundancy ensured better scaling and correction for absorption and other systematic errors. This final data set was then used in all the subsequent procedures. A further check on the quality of the data is the correlation coefficient between the experimental signed anomalous differences (calculated after merging the φ and ω scans) and the ideal anomalous differences derived from the structure factors calculated from the final least-squares refined structure taking anomalous dispersion into account (Fig. 1*b*). The ratio of the unsigned anomalous difference to its estimated standard deviation was also calculated as a function of the resolution, as was the mean $I/\sigma(I)$ ratio after merging Friedel opposites (Fig. 1*c*).

For substructure solution with *SHELXD*, the data were truncated to the resolution indicated by the statistical tests. The minimum distance between atoms was reset (from the usual *SHELXD* value of 3.5 Å) to 1.8 Å to search for sulfurs in disulfide bridges. Solutions were identified by high values of the correlation coefficient between $E_{\rm obs}$ and $E_{\rm calc}$ for all data and also for the weak data only (*i.e.* for the reflections not used to find the heavy atoms). In each case there was a large gap in the correlation coefficients between the correct solutions and the others. *SHELXD* assumes that all the atoms found have the same scattering factors, but also refines the occupancies, which can partially compensate for different element types and temperature factors. In most cases there was a sharp drop in occupancy after the last correct atom.

The file containing the anomalous atom positions and occupancies was read into *SHELXE* without any editing or heavy-atom refinement and two parallel *SHELXE* jobs were started for the two possible substructure enantiomorphs. The contrast and connectivity figures of merit output by *SHELXE*

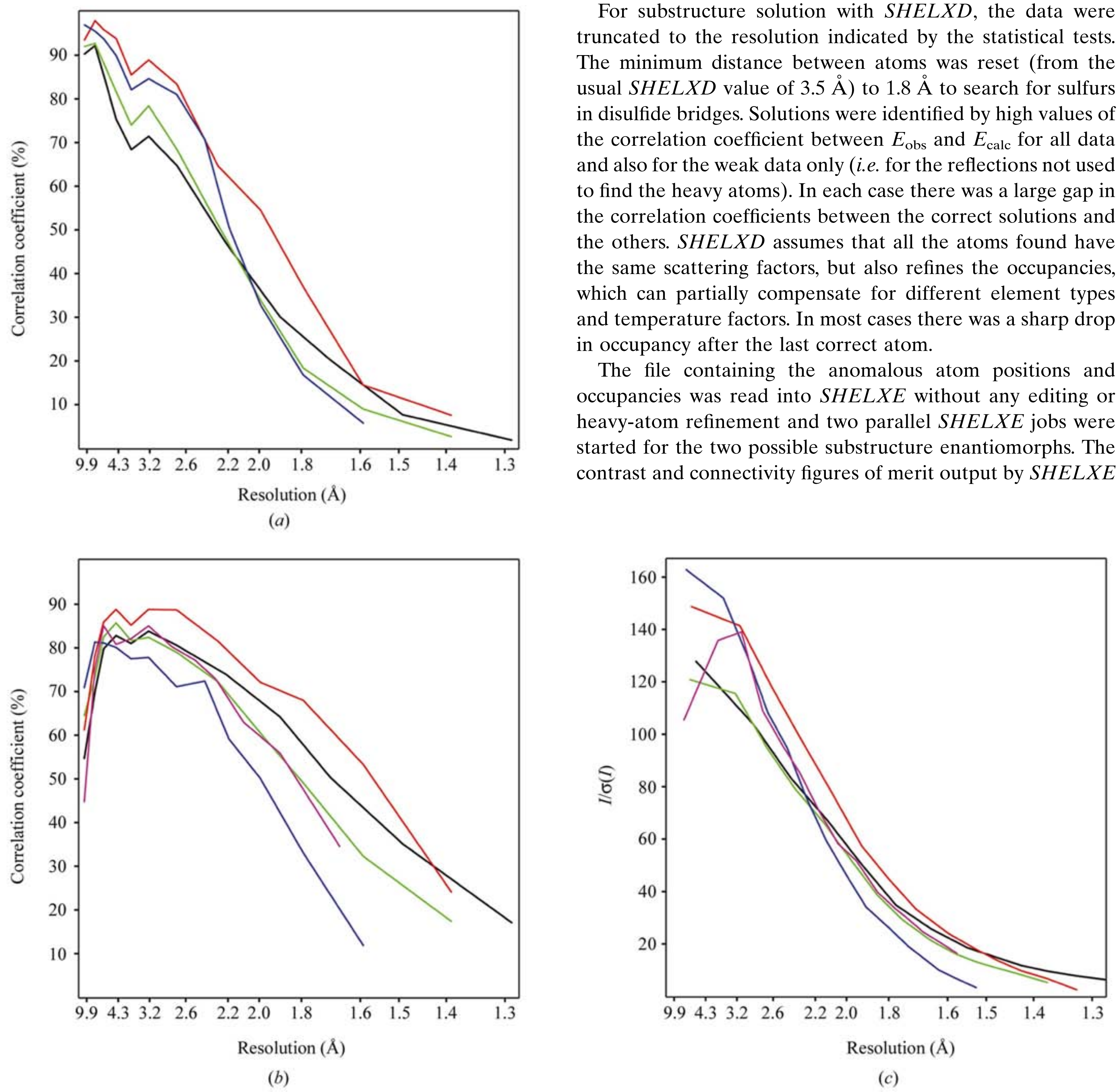

Figure 1
(*a*) Correlation coefficients (%) between the signed anomalous differences $F_+ - F_-$ for the φ and ω scan data, (*b*) correlation coefficients between the observed signed anomalous differences and those calculated from the final refined model and (*c*) mean $I/\sigma(I)$ as a function of the resolution (Å) for orthorhombic trypsin (black), trigonal trypsin (green), cubic insulin (red), rhombohedral insulin (blue) and thaumatin (purple).

Table 4
SHELXD (first four rows) and *SHELXE* phasing statistics.

Experimental phases not retained.

	Ortho- rhombic trypsin	Trigonal trypsin	Cubic insulin	Rhombo- hedral insulin	Thaumatin
Truncated to (Å)	1.70	1.80	1.60	1.90	1.80
CC	37.8	39.1	39.6	34.8	39.5
CC (weak)	22.0	22.6	24.0	14.8	23.1
Success rate (%)	76	25	28	0.9	6.3
Expanded to (Å)	1.19	1.32	1.30	1.50	1.55
No. of cycles	100	50	20	20	20
Contrast	0.48	0.43	1.17	0.63	0.72
Contrast (inverted)	0.36	0.40	0.53	0.45	0.29
Solvent content	0.25	0.33	0.50	0.40	0.45
Final mean fom	0.854	0.757	0.843	0.724	0.831
Final mean cosine	0.881	0.726	0.883	0.677	0.830
Final map CC	0.968	0.882	0.973	0.807	0.955
Final wMPE (°)	10.9	21.6	9.8	28.5	13.1

(Sheldrick, 2002) clearly identified the correct hand. The progress of the phase determination is summarized in Table 4.

3.1. Orthorhombic trypsin

For substructure solution, the anomalous difference data for the orthorhombic form of trypsin were truncated to 1.8 Å as indicated by the correlation coefficient. Since this resolution allows discrimination of sulfurs in disulfide bonds, *SHELXD* was requested to search for 15 sites. Three quarters of the trials gave the correct solution. The occupancies of the candidate atoms showed a sharp drop between the 15th and 18th site (0.314, 0.288, 0.190, 0.103). Sites 16 and 17 were confirmed by the anomalous map generated from the final *SHELXE* phases and after structure refinement they were found to be the S atoms of partially occupied sulfate anions. The final density-modified map was very similar to the map after anisotropic structure refinement, as the map correlation coefficient of 0.968 implies. Subsequent autobuilding with *wARP* resulted in a nearly complete trace (five residues were missing, but these were not difficult to add by hand).

3.2. Trigonal trypsin

For the trigonal trypsin crystal the φ scans were appreciably noisier than the ω scans, but despite this the correlation coefficient between the φ and ω scan anomalous differences led to a sensible truncation limit. All S atoms and the Ca atom were found using *SHELXD*. The enantiomorph discrimination was weaker than in the other four examples, but both the contrast (correct 0.43/inverted 0.40) and connectivity (0.90/0.84) figures of merit in *SHELXE* indicated the correct enantiomorph. The map correlation coefficient for the inverted substructure (and space group) relative to the inverted reference phases rose from about 0.12 at low resolution to 0.26 at high resolution, significantly higher than the value of zero usually found for the wrong substructure enantiomer. This indicates that although the SAD phases should lead to noise for the incorrect substructure enantiomer, the direct heavy-atom contribution is sufficient to give a partially correct but inverted map, explaining the relatively poor discrimination shown by the *SHELXE* figures of merit. Phasing and density modification with *SHELXE* starting from the correct substructure enantiomer led to a clear easily interpretable map that allowed 215 residues out of 223 to be autobuilt by *wARP*.

3.3. Cubic insulin

The relatively high proportion of sulfur and the high-symmetry space group (making it easy to achieve a high redundancy) for cubic insulin resulted in a highly significant anomalous signal as far as 1.5 Å resolution. The high redundancy also enabled some contamination caused by ice formation during an ω-scan to easily be filtered out. The six S atoms were routinely found and the contrast figure of merit in *SHELXE* distinguished particularly clearly between the two substructure enantiomorphs (Table 4). The progress of the *SHELXE* phase improvement is shown in Figs. 2, 3(*a*), 3(*b*) and 3(*c*). The density-modified map in Fig. 3(*c*) closely resembles the map from the final anisotropic refinement shown in Fig. 3(*d*) (map correlation coefficient 0.973). *ARP*/*wARP* traced the protein almost completely; only the disordered terminal atoms were missing.

3.4. Rhombohedral insulin

In rhombohedral insulin the high proportion of sulfurs is augmented by the presence of the slightly stronger anomalous scatterers zinc and chlorine that are not present in the cubic modification, but this advantage is neutralized by the lower

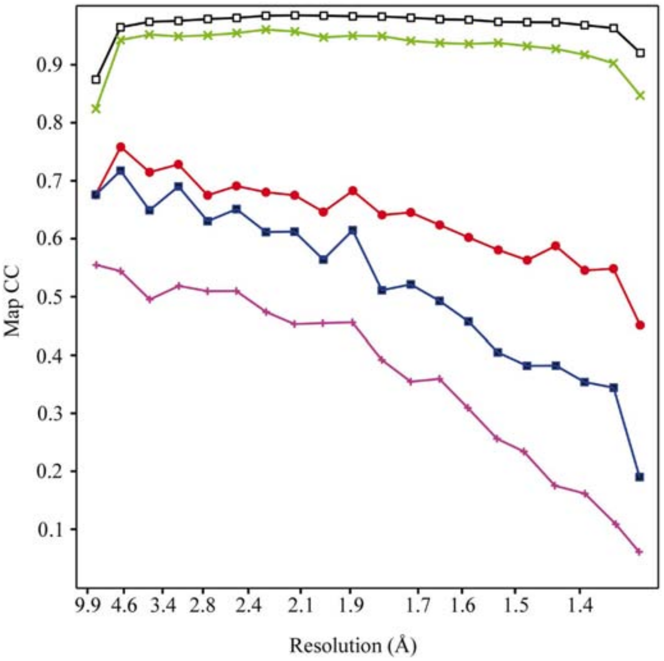

Figure 2
Map correlation coefficients for insulin during phase determination: SAD centroid phases (purple plus signs), after resolving the twofold ambiguity (blue filled squares), after including the sulfur contribution (red circles), after five cycles of density modification (green crosses) and after 20 cycles (black open squares).

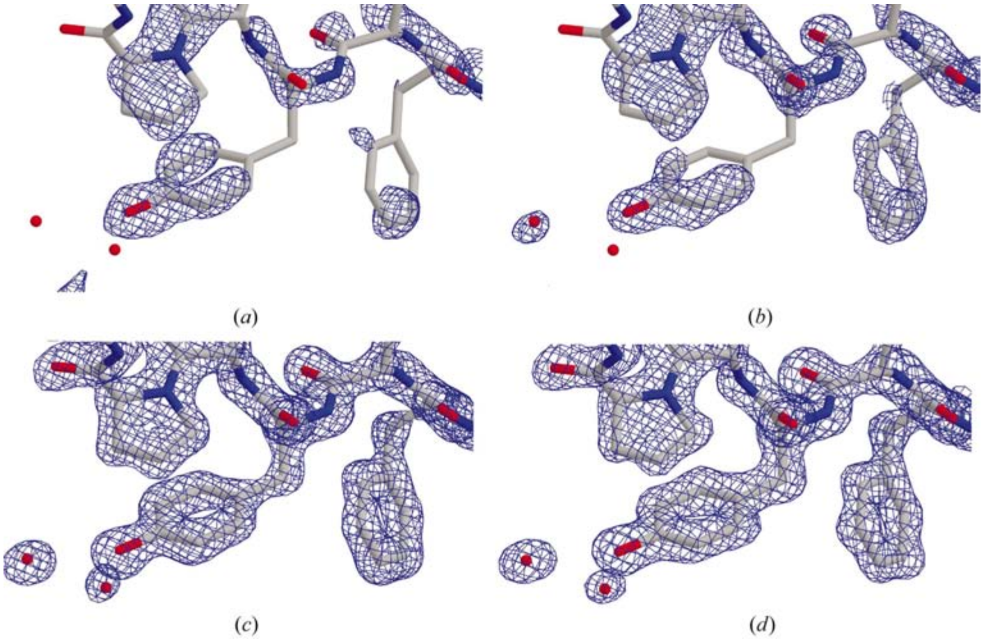

Figure 3
A region of cubic insulin showing the final refined model and the electron density at different stages: (*a*) centroid phases, (*b*) twofold ambiguity resolved and heavy-atom phases added, (*c*) after density modification, (*d*) after anisotropic least-squares structure refinement.

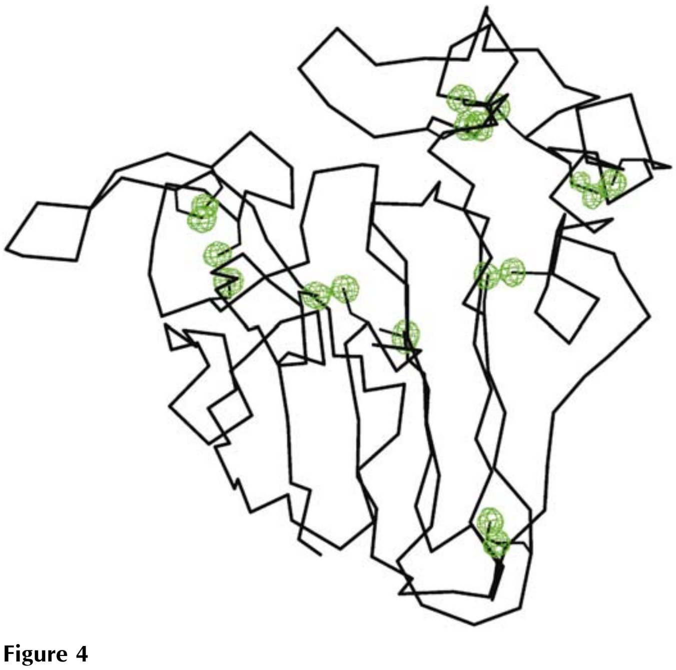

Figure 4
An anomalous Fourier map for thaumatin calculated using the observed anomalous differences and phases derived by subtracting 90° from the phases after *SHELXE* density modification, showing nicely resolved S atoms. In this test only the unresolved super-sulfur atoms were input to *SHELXE* (see discussion in §§3.5 and 4.2).

symmetry and hence the lower data redundancy. As a result of insufficient cryoprotection, all the frames exhibited weak ice rings. Despite this the data are of high quality; they were truncated to 1.9 Å as indicated by the correlation coefficient between the φ and ω scans. The special position rejection test in *SHELXD* was switched off because one of the zinc ions and one chloride ion lie on special positions. Density modification with *SHELXE* led to an interpretable map that could be autotraced with about 80% of the residues built. The high *R* factors for the refinement and the relatively modest map correlation coefficient and success in autotracing can almost certainly be attributed to the extensive disorder present in this structure.

3.5. Thaumatin

For thaumatin the 149 Å cell edge required a larger camera distance (12 cm) with a corresponding increase in the total data-collection time; the data collection was therefore restricted to ω scans that sample reciprocal space more efficiently than φ scans at higher 2θ. As a result, the correlation between φ and ω scans could not be used to establish the best resolution for truncating the data for substructure solution. Truncation to 1.8 Å led to the location of all the S atoms and density modification led to a map correlation coefficient of 0.951 relative to the anisotropically refined structure; *wARP* was able to autotrace 192 of the 207 residues.

In a separate test, the thaumatin data were truncated to 2.8 Å to test the location of 'super-sulfur' atoms (low-resolution disulfide bridges with the two S atoms fused together). Although the highest values of the correlation coefficient corresponded to the correct solution, the discrimination from the wrong solutions was poor and the values of the correlation coefficient were rather low (9.2 for all data and 5.0 for the weak data). The starting phases resulting from these positions were weak, but 250 cycles of density modification led to a slow but steady phase improvement and eventually to a map correlation coefficient almost identical to that obtained in 20 cycles from the resolved S atoms. A Fourier map calculated with the signed anomalous differences as amplitudes and phases obtained by subtracting 90° from the final density-modified phases is shown in Fig. 4; it shows the disulfide bonds with well resolved S atoms. As suggested below, this type of map may give an optimistic view of the quality of the anomalous difference data.

4. Discussion

4.1. In-house measurements

The results presented here show that it is possible to solve protein structures using the anomalous scattering from native S atoms measured on a laboratory instrument in a careful but relatively routine manner, provided that a sufficiently high real redundancy is obtained (ranging from 16 to 44 in these experiments). Real redundancy implies measurement of equivalent or identical reflections with different paths through the crystal, not just repeated measurements; this is expedited by high crystal symmetry and by the use of a three-circle (or κ) goniometer. Particular care should be taken in the cryogenic mounting of the crystal *etc.* to reduce the background scattering to a minimum and in the planning of the data collection to ensure a good spread of equivalent reflections on different frames so that the scaling program (*SADABS*) can scale and model the residual systematic errors well. These laboratory data collections required more time than synchrotron measurements, but did not suffer from variations in the incident beam intensity or crystal decomposition, both of which make accurate measurements of the very small anomalous differences more difficult. Synchrotron data collection is still invaluable for collecting the native data to the highest attainable resolution; in addition to increasing the information content of the final refined structure, the higher resolution native data lead to an appreciable improvement in the quality of the experimental density-modified maps from *SHELXE* (Sheldrick, 2002). High redundancy is equally important for SAD phasing from weak anomalous signals with synchrotron data (Dauter & Adamiak, 2001). In the experiments reported here, much of the data-collection time was spent in collecting data to as high a resolution as possible, but the data were then truncated to the limit of the anomalous signal for the substructure solution. For the purposes of sulfur-SAD phasing and subsequent refinement, a highly redundant medium-resolution laboratory data set combined with a low-redundancy high-resolution synchrotron data set could provide the best of both worlds. High resolution is not an essential requirement for Cu $K\alpha$ sulfur-SAD phasing; the recent solution (Debreczeni, Bunkóczi *et al.*, 2003) of an unknown protein from a crystal that barely diffracted to 3 Å in the laboratory and to about 2 Å on a synchrotron showed that the method is also applicable to more typical real-life cases.

4.2. The extent of the anomalous signal

It appears that the correlation coefficient between the signed anomalous differences of two statistically independent data sets (Fig. 1*a*; Schneider & Sheldrick, 2002) or between one experimental data set and the ideal data (Fig. 1*b*) give consistent and reliable measures of the strength of the anomalous signal and of its dependence on resolution. The statistically independent data sets may consist of data collected from two different crystals or two data sets collected from one crystal where the data are collected in different ways (*e.g.* φ and ω scans) so as to minimize correlations between the two. The latter procedure has the advantage that there are no problems in subsequently merging the two data sets to obtain a higher redundancy. After the assessment all the data can be scaled together (*e.g.* in the same *SADABS* run), resulting in a data set having twice the redundancy of those used for comparison. This explains why the correlation coefficients between observed and ideal data are higher than those between the φ and ω scans. After additionally merging the Friedel opposites, the mean value of $I/\sigma(I)$ was plotted as a function of resolution (Fig. 1*c*). This shows the same trends as the correlation coefficient plots; in each case cubic insulin (the red line) gives the highest values, except at very high resolution when it is overtaken by orthorhombic trypsin (black), and rhombohedral insulin (blue) gives the lowest values. Truncation to the resolution at which $\langle I/\sigma(I)\rangle$ drops to about 30 would have led to good solutions. It may be possible to reduce this limit for stronger anomalous scatterers than sulfur, but in such cases it might be better to calculate $\langle I/\sigma(I)\rangle$ without merging Friedel opposites. The figure of merit $R_\sigma = \sum\sigma(I)/\sum I$, which has been used for many years in the *SHELX* program system, is of course closely related to the reciprocal of $\langle I/\sigma(I)\rangle$. It would be reasonable to truncate the data at the resolution at which R_σ becomes less than the value of $2\sum|I_+ - I_-|/\sum(I_+ + I_-)$ calculated from the ideal data (see Table 2). Although the latter value can only be calculated after solving the structure, it is relatively constant for the five structures reported here, so a standard value could be used. The use of the ratio of the mean unsigned anomalous difference to its mean standard deviation proved to be less consistent and so was not used in this work; this may be the result of difficulties in estimating the standard deviations of the anomalous differences accurately. An alternative viable approach for laboratory data collection would be simply to continue to collect data, increasing the redundancy, until the structure can be solved.

A commonly used procedure for demonstrating the presence of anomalous signal is the calculation of a map using the experimental anomalous differences and phases obtained by subtracting 90° from the protein phase from the final refinement (Strahs & Kraut, 1968). In our experience, this gives convincing positions for the anomalous scatterers even when the anomalous signal is too weak to locate them directly (*e.g.* with *SHELXD*). This is because the phase is more important than the amplitude in a Fourier synthesis, and in this case – in contrast to the usual situation in crystallography – the phases are known but the amplitudes may be inaccurate. Although not useful for assessing the quality of the anomalous signal, this procedure is still appropriate for identifying or confirming residual anomalous scatterers, *e.g.* low-occupancy chloride ions that might otherwise have been mistaken for water (Dauter *et al.*, 1999) or the two disordered sulfate ions in the orthorhombic trypsin example presented here.

4.3. Phase determination

The location of the S atoms and possible other weak anomalous scatterers with the integrated Patterson and direct methods implemented in *SHELXD* is relatively straightforward. The resolution at which to truncate the anomalous differences is fairly critical and if this cannot be established by the above statistical tests, a good starting point for a trial-and-error approach is to take 0.5 Å lower resolution than the diffraction limit of the crystal under the data-collection conditions employed. The decision whether to search for

Table 5
Influence of the redundancy on the correlation coefficient (CC) and success rate (out of 1000 trials) of *SHELXD* with and without data scaling using *SADABS*.

Redundancy†	Redundancy‡	Best CC§ (%)	Success rate§	Best CC¶ (%)	Success rate¶
3.8	2.0	22.7	22	16.9	0
7.1	3.7	20.2	40	13.3	0
14.2	7.4	28.6	222	28.1	74
21.4	11.1	33.9	279	33.1	182
44.3	23.0	40.8	448	40.9	437

† Friedel mates treated as equivalent. ‡ Friedel mates separate. § After *SADABS* scaling. ¶ Without *SADABS*.

individual sulfurs or super-sulfur atoms affects the number of atoms to be searched for as well as the minimum distance allowed between atoms; usually, if the data are truncated to 2.1 Å or lower resolution and disulfide bonds are present it is necessary to search for super-sulfurs (S–S units scattering as a single atom). Often the number of S (or super-sulfur) atoms will be known in advance, but it is important to bear in mind the possibility of additional partially occupied chloride ions or other weak anomalous scatterers. Fortunately, the occupancy refinement in *SHELXD* handles this situation well, as shown by the location of the partially occupied sulfates in the orthorhombic trypsin structure. In the case of very weak anomalous scatterers more trials (say 1000) may be required, but often ten will suffice. The correct solution should have correlation coefficients clear of 'noise'; usually for sulfur SAD phasing the correct solutions will have correlation coefficients between 30 and 40% (all data) and between 15 and 30% (weak data), but the super-sulfur test using thaumatin shows that much lower values can still be successful. Where disulfide bridges are present and individual sulfurs are being searched for, the presence of disulfide bridges in the substructure solution is also a good confirmation that it is correct.

At the request of the referees, we have included details (Table 5) of the influence of the redundancy and the use of the scaling program *SADABS*. It will be seen that increasing the redundancy (by varying the number of scans used) has a dramatic effect on both the quality of the *SHELXD* substructure solutions (as measured by the correlation coefficient CC) and on their frequency, exactly as reported by Dauter & Adamiak (2001). For the two data sets with the smallest redundancy, no solutions were obtained unless *SADABS* was used, although scaling had less effect on the figures of merit and on the number of solutions for very high redundancy. For the data processed without *SADABS*, it was still necessary to scale scans with different exposure times *etc.* to one another using one overall scale factor per scan; the program *XPREP* was used for this. At low redundancy there was also a wider range of CC for substructures containing the six correct S atoms. In all cases, the solution with the highest CC corresponded to a correct solution except in the two cases (with lower maximum CC values) where no solution was found.

The quality of the phase refinement in *SHELXE* is dependent on the resolution of the native data and the solvent content. In all these examples it was the high resolution (1.55 Å or better) of the native data that was primarily responsible for the high quality of the maps after density modification; in other cases, including an unknown structure solved recently by the procedures described here (Debreczeni, Bunkóczi *et al.*, 2003), a high solvent content (≥60%) led to interpretable maps despite ~3 Å native data resolution (truncated to 3.7 Å for super-sulfur location).

Although these test structures were treated as unknowns and no information was introduced into the phasing process in form of partial models or sulfur positions, the *a priori* knowledge of the structures was useful in developing the procedures, including the fine tuning of the algorithms used in the currently distributed version of the *SHELXE* program.

We are grateful to Thomas R. Schneider and Isabel Usón for discussions and to the Fonds der Chemischen Industrie and the DFG (SFB416) for support. This work is dedicated to the late Robert A. Sparks who unsuccessfully tried to persuade GMS that it would be possible to determine protein structures routinely *via* the anomalous scattering of sulfur measured in the laboratory.

References

Brown, J., Esnouf, R. M., Jones, M. A., Linnell, J., Harlos, K., Hassan, A. B. & Jones, E. Y. (2002). *EMBO J.* **21**, 1054–1062.

Bruker Nonius (2002). *PROTEUM*, *SAINT*, *SADABS*, *SMART* and *XPREP* computer programs.

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.

Debreczeni, J. É., Bunkóczi, G., Girmann, B. & Sheldrick, G. M. (2003). *Acta Cryst.* D**59**, 393–395.

Debreczeni, J. É., Girmann, B., Zeeck, A. & Sheldrick, G. M. (2003). Manuscript in preparation.

Dodson, E. J., Dodson, G. G., Lewitoya, A. & Sabesan, M. (1978). *J. Mol. Biol.* **125**, 387–396.

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

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

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

Ko, T.-P., Day, J., Greenwood, A. & McPherson, A. (1994). *Acta Cryst.* D**50**, 813–825.

Lemke, C. T., Smith, G. D. & Howell, P. L. (2002). *Acta Cryst.* D**58**, 2096–2101.

Liu, Z.-J., Vysotski, E. S., Chen, C.-J., Rose, J. P., Lee, J. & Wang, B.-C. (2000). *Protein Scie.* **9**, 2085–2093.

Lunin, V. Y. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.

McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.

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

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

Renatus, M., Bode, W., Huber, R., Sturzebecher, J. & Stubbs, M. T. (1998). *J. Med. Chem.* **41**, 5445–5456.

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

Schroder Leiros, H.-K., McSweeney, S. M. & Smalås, A. O. (2001). *Acta Cryst.* D**57**, 488–497.

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

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

Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–341.

Smith, G. D., Pangborn, W. A. & Blessing, R. H. (2001). *Acta Cryst.* D**57**, 1091–1100.

Smith J. L. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 211–225. Dordrecht: Kluwer Academic Publishers.

Strahs, G. & Kraut, J. (1968) *J. Mol. Biol.* **35**, 503–512.

Usón, I. & Sheldrick, G. M. (1999). *Curr. Opin. Struct. Biol.* **9**, 643–648.

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

Weiss, M. S., Sicker, T. & Hilgenfeld, R. (2001). *Structure*, **9**, 771–777.

Yang, C. & Pflugrath, J. W. (2001). *Acta Cryst.* D**57**, 1480–1490.