Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Getting the most out of X-ray home sources

R. A. P. Nagem,[a]‡ A. L. B. Ambrosio,[a] A. L. Rojas,[a] M. V. A. S. Navarro,[a] A. M. Golubev,[b] R. C. Garratt[a] and I. Polikarpov[a]*

[a]Institute of Physics of São Carlos, University of São Paulo, Avenida Trabalhador Sãocarlense 400, CEP 13560-970, São Carlos, SP, Brazil, and [b]Petersburg Nuclear Physics Institute, Gatchina, St Petersburg, 188300, Russia

‡ Current address: Biochemistry and Immunology Department, Institute of Biological Sciences, Federal University of Minas Gerais, Avenida Antônio Carlos 6627, Caixa Postal 486, CEP 31270-901, Belo Horizonte, MG, Brazil.

Correspondence e-mail: ipolikarpov@if.sc.usp.br

Received 2 March 2005
Accepted 22 April 2005

The structures of a 14 kDa phospholipase, an 18 kDa proteinase inhibitor and a novel glycoside hydrolase with molecular weight 60 kDa were solved using the SAD technique and the effects of the amount of anomalous signal, completeness and redundancy of data on heavy-atom substructure determination, phasing and model building were analyzed. All diffraction data sets were collected on a Cu-anode X-ray home source. The structure of the phospholipase was obtained using the anomalous scattering contribution from its 16 S atoms. Three-dimensional models for the other two macromolecules were obtained using the anomalous contribution of I atoms rapidly incorporated into the crystal through the quick cryo-soaking method of derivatization. These results were used to discuss the application of sulfur- and iodine-SAD approaches in combination with X-ray home sources for high-throughput protein crystal structure solution. The estimates of the anomalous signal from S atoms in the gene products of four genomes are given and the prospects for increasing the anomalous contribution using longer wavelengths (*e.g.* from a chromium home source) and quick cryo-soaking derivatization are discussed. The possibility of rapidly preparing tangible home-source isomorphous derivatives suggests that this approach might become a valuable tool in the future of post-genomic projects.

1. Introduction

The electron-density map required to construct an atomic model of a macromolecule by means of X-ray crystallography can only be obtained if the phase and the amplitude of the structure factors are available. However, during a typical X-ray diffraction experiment only the structure-factor amplitudes can be obtained. This is the so-called 'phase problem' and has been discussed in detail in several textbooks (Blundell & Johnson, 1976; Drenth, 1999).

The first largely successful technique developed to estimate the phases and consequently to obtain an initial electron-density map of a macromolecular crystal was pioneered by Perutz and coworkers in the 1950s (Green *et al.*, 1954). This approach resulted in the elucidation of the first three-dimensional structure of a protein, haemoglobin. This technique is based on the isomorphous differences observed in the diffraction pattern owing to the incorporation of a few heavy atoms into the ordered structure of native crystals and is termed the multiple/single isomorphous replacement (MIR/SIR) method. The anomalous scattering of X-rays can also be used to yield independent phase information (Laue, 1916; Bijvoet, 1949). In practice, its use in phase determination was reported in the 1980s (Hendrickson & Teeter, 1981; Wang,

1985). The anomalous signal, which is observed as differences between Friedel-related reflections (Bijvoet differences), is usually small and depends on both the atom type and the wavelength of the incident X-rays. The combined use of such approaches in a SIRAS/MIRAS (single/multiple isomorphous replacement with anomalous scattering) experiment has also been used for phase determination of several macromolecule crystals.

More recently, with the advent of tunable synchrotron beamlines devoted to protein crystallography experiments, the use of the anomalous diffraction techniques with either single-wavelength (SAD) or a multi-wavelength (MAD; Hendrickson, 1991) data has become routine. The MAD method uses only the wavelength-dependence of f' and f'' (dispersive and anomalous components of the atomic structure factor) of the anomalously scattering atoms for phasing. Usually, three or four data sets are collected at various wavelengths in order to optimize the differences between the f' and f'' contributions. Nowadays, the MAD approach in conjunction with selenomethionine-derivatized proteins has become the method of choice for solving new protein crystal structures. Nevertheless, several recent experiments have demonstrated that even when anomalous signals are weak, the SAD approach can be used for phasing and structure determination if based on high-quality diffraction data (Dauter *et al.*, 1999, 2002; Liu *et al.*, 2000; Bond *et al.*, 2001; Dauter & Adamiak, 2001; Gordon *et al.*, 2001; Lemke *et al.*, 2002; Micossi *et al.*, 2002; Nagem *et al.*, 2003; Debreczeni, Bunkóczi, Ma *et al.*, 2003).

Now that a number of genomes have been completely elucidated and many more are being sequenced, the availability of thousands and thousands of genes has created the challenge of high-throughput structure elucidation of gene products. This in turn has stimulated an increasing interest in the development of less expensive equipment for automatic crystal growth (Stevens, 2000; Watanabe *et al.*, 2002), techniques that require little time and effort for successful phasing (Dauter *et al.*, 2000; Nagem *et al.*, 2001; Evans & Bricogne, 2002) as well as robust software for heavy-atom location (Schneider & Sheldrick, 2002; Burla *et al.*, 2003), density-modification protocols (Terwilliger, 2002; Sheldrick, 2002) and automatic electron-density interpretation (Perrakis *et al.*, 1999; Terwilliger, 2002; Ioerger & Sacchettini, 2002). Massive investment has been made in the construction, maintenance and upgrade of advanced dedicated protein crystallography beamlines on synchrotrons worldwide. Training of specialized personnel to help synchrotron users to collect diffraction data from crystals of macromolecules ranging from small proteins to viruses or ribosomes also plays an important role in the competitive world of synchrotron beam time. We therefore may conclude that within the last 50 years drastic changes have influenced the way protein crystallography is being performed.

All these technological innovations and advances developed within high-throughput structure-solution projects pose a few questions. How can laboratories with limited access to synchrotron sources still contribute to such projects? How can we make the most of an X-ray home source in such a way that it can be used not only to test crystals prior to a synchrotron data collection but also for structure solution of novel proteins?

In the present work, we describe the use of X-rays from a rotating Cu-anode home source to solve the structures of a 14 kDa sulfur-rich enzyme, an 18 kDa proteinase inhibitor and a novel 60 kDa glycosidase by the SAD method. The first protein was solved using the anomalous scattering contribution of its 16 S atoms, while the structures of the latter two macromolecules were obtained using the anomalous contribution of I atoms rapidly incorporated into the crystal through the quick cryo-soaking approach (Dauter *et al.*, 2000; Nagem *et al.*, 2001, 2003). The use of sulfur phasing, which is associated mostly with synchrotron data collection, but also with data sets collected on X-ray home sources, has been widely exploited by many crystallographers in recent years (Dauter *et al.*, 1999; Liu *et al.*, 2000; Bond *et al.*, 2001; Dauter & Adamiak, 2001; Gordon *et al.*, 2001; Lemke *et al.*, 2002; Micossi *et al.*, 2002; Debreczeni, Bunkóczi, Girmann *et al.*, 2003; Debreczeni, Girmann *et al.*, 2003; Olsen, Flensberg, Olsen, Bricogne *et al.*, 2004; Olsen, Flensberg, Olsen, Seibold *et al.*, 2004). Owing to the natural occurrence of S atoms in proteins, it has been proposed to use these anomalous scatterers for phasing despite their small anomalous contribution within the practically available X-ray range ($f'' = 0.56$ e$^-$ at the Cu $K\alpha$ wavelength or $f' = 1.14$ e$^-$ at the Cr $K\alpha$ wavelength). Alternatively, an elevated anomalous signal can be obtained by derivatization of protein crystals using the quick cryo-soaking approach with appropriate high atomic weight ions (halides or alkali metals). The quick cryo-soaking approach has mostly been used in combination with synchrotron sources (Nagem *et al.*, 2003). Here, we report our results on the application of this derivatization technique in conjunction with diffraction data collected at a Cu rotating-anode home source. The crystal structure determination of macromolecules with molecular weights of up to 60 kDa on an in-house X-ray source carry the promise that these methods can be routinely applied for protein structure solution as a complement to modern synchrotron-radiation-based techniques.

2. Materials and methods

2.1. Crystal preparation and data collection

Three different proteins were purified to homogeneity and crystallized using the hanging-drop method as described elsewhere.

The first protein is a class II Lys49-PLA$_2$ from the venom of *Agkistrodon contortrix laticinctus*. It is a sulfur-rich enzyme (14 kDa) belonging to a large well studied family of phospholipases A$_2$ (Acl-PLA$_2$; PDB code 1s8i; Ambrosio *et al.*, 2005). Well diffracting crystals were grown in hanging drops after mixing equal volumes of protein solution at 10 mg ml^{-1} and mother-liquor solution. The mother-liquor solution consisted of 2.0 *M* ammonium sulfate, 0.1 *M* Tris–HCl pH 8.5. A suitable crystal for data acquisition was obtained after transferring a single crystal from the crystallization drop to a

Table 1
Data statistics for Nat-Acl-PLA$_2$, I-Aaw-Ghy, Nat-Bba-Pin and I-Bba-Pin data sets.

Values in parentheses refer to the highest resolution shell in each data set. All data sets were collected with Cu $K\alpha$ radiation from a Rigaku Ultra-X 18 rotating-anode generator.

	Nat-Acl-PLA$_2$				I-Aaw-Ghy				Nat-Bba-Pin	I-Bba-Pin
Space group	$P4_12_12$				$P2_1$				$P2_12_12_1$	$P2_12_12_1$
Unit-cell parameters (Å, °)	$a = b = 70.57$, $c = 57.45$				$a = 49.96$, $b = 93.35$, $c = 67.83$, $\beta = 106.92$				$a = 46.70$, $b = 64.14$, $c = 59.24$	$a = 47.05$, $b = 63.12$, $c = 59.78$
Mosaicity (°)	0.45				0.9				0.5	0.5
$\Delta\varphi$ rotation (°)	1.0				1.0				0.7	1.0
Resolution range (Å)	26.63–1.61 (1.65–1.61)				26.63–1.87 (1.97–1.87)				26.43–1.87 (1.97–1.87)	25.27–2.20 (2.32–2.20)
Data-collection time (h)	21				47				15	23
Total φ rotation (°)	90	180	270	359	102	205	308	411	213	250
Total reflections	131349	263493	395738	526515	95202	180160	272594	360302	124202	88825
Unique reflections	35899	35928	35942	35959	37859	44936	45075	45205	14580	16658
Redundancy	3.6 (3.4)	7.3 (6.8)	11.0 (10.2)	14.6 (14.1)	2.5 (2.3)	4.0 (3.8)	6.0 (5.7)	8.0 (7.5)	8.5 (8.4)	5.3 (5.6)
Overall $R_{\rm sym}$ (%)	5.7 (27.5)	6.3 (30.9)	6.2 (31.4)	6.1 (31.4)	5.9 (17.9)	4.3 (12.0)	5.1 (13.5)	5.3 (13.8)	5.0 (21.0)	10.2 (36.8)
Completeness (%)	99.6 (97.2)	99.7 (97.9)	99.8 (99.3)	99.8 (99.8)	64.1 (61.5)	84.4 (72.6)	88.3 (77.9)	91.0 (81.6)	95.8 (95.8)	95.5 (95.5)
$I/\sigma(I)$	23.8 (4.0)	33.9 (5.7)	41.6 (7.0)	49.5 (8.1)	7.0 (4.2)	10.9 (6.0)	9.2 (5.4)	9.1 (5.2)	11.0 (3.6)	6.7 (2.0)

cryogenic solution containing the mother liquor and 15% ethylene glycol for a few seconds. This crystal (Nat-Acl-PLA$_2$) was then flash-cooled to 100 K in a cold nitrogen stream and used for data collection. A redundant data set was collected on a MAR345 image-plate detector using Cu $K\alpha$ radiation generated from a Rigaku Ultra-X 18 rotating-anode generator operating at 80 mA and 50 kV and focused using Osmic mirrors. Diffraction images were processed with *DENZO* (Otwinowski & Minor, 1997) and scaled with *SCALEPACK* (Otwinowski & Minor, 1997) in four different batches of total φ rotation. As a result, four different data sets of total φ rotation equal to 90, 180, 270 and 359° were obtained.

The second macromolecule is a proteinase inhibitor with 164 amino-acid residues extracted from the *Bauhinia bahiniodes* tree (Bba-Pin). Protein crystallization was achieved using 8% PEG 4000, 0.1 *M* sodium acetate trihydrate pH 4.6 as the mother-liquor solution. Two crystals were prepared for data collection: a native and an iodine derivative. The first crystal (Nat-Bba-Pin) was obtained in the same fashion as Nat-Acl-PLA$_2$. However, the second crystal (I-Bba-Pin) was prepared according to the quick cryo-soaking procedure for derivatization in which native crystals are rapidly soaked in cryogenic solutions containing a high concentration of an appropriate salt; in this case the mother liquor with 15% ethylene glycol and 0.5 *M* NaI. Both crystals were rapidly frozen in a nitrogen stream and the same technical aspects of data collection used for the first protein were applied here. Diffraction images for each data set were processed with *MOSFLM* (Leslie, 1992) and scaled with *SCALA* (Evans, 1997) in a single batch of total φ rotation.

The last protein is a novel extracellular glycoside hydrolase (537 amino acids) from the filamentous fungus *Aspergillus awamori* (Aaw-Ghy; PDB codes 1y4w and 1y9g; Nagem *et al.*, 2004). Only a single iodine derivative (I-Aaw-Ghy) was prepared for this protein using the quick cryo-soaking approach. One native crystal was transferred from the crystallization drop to a new solution drop containing the mother liquor (15% PEG 3350, 0.1 *M* sodium acetate pH 4.2), 20% glycerol and 0.5 *M* NaI. The crystal was immersed for a few minutes and then flash-cooled in the nitrogen stream. More than 400 diffraction images ($\Delta\varphi = 1.0°$) were collected from this derivative and processed with *MOSFLM*. Following the procedure adopted with the Nat-Acl-PLA$_2$ data set, intensities were scaled with *SCALA* in four different batches of total φ rotation equal to 102, 205, 308 and 411°.

Further details of data acquisition and data statistics for all crystals used in this work are shown in Table 1.

2.2. Anomalous scatterer search, phase calculation and model building

The location of S atoms in the Nat-Acl-PLA$_2$ crystal and iodide anions in the I-Bba-Pin and I-Aaw-Ghy derivatives was performed using the dual-space recycling algorithm (Miller *et al.*, 1993) implemented in the *SHELXD* program (Sheldrick *et al.*, 2001). During anomalous substructure identification *SHELXD* default parameters were used. No resolution cutoff was imposed. Anomalous differences obtained with *SHELXPRO* (Sheldrick & Schneider, 1997) were separately chosen by the *SHELXD* algorithm for 500 multi-substructure solution trials.

A total of four different runs of *SHELXD* were performed with Nat-Acl-PLA$_2$; each run corresponded to data sets with different total φ rotation (90, 180, 270 and 359°). In all four cases *SHELXD* was set to search for 16 sulfur sites. A similar approach was adopted for the location of 12 partially occupied iodide anions in the I-Aaw-Ghy data sets (102, 205, 308 and 411°). As a consequence, the effect of data-set redundancy in the location of heavy-atom sites could be studied by analyzing the outcome from each run. In the proteinase-inhibitor case, *SHELXD* was set to search for eight iodine sites in a single run, corresponding to a data set of total φ rotation equal to 250° (I-Bba-Pin).

Table 2
Detailed statistics obtained at each step from heavy-atom search to model building using Nat-Acl-PLA$_2$, I-Bba-Pin and I-Aaw-Ghy data sets.

Phase calculation using SAD or SIRAS methods and model building were carried out only with promising data sets.

Parameters	Nat-Acl-PLA$_2$				I-Bba-Pin	I-Aaw-Ghy			
Total φ rotation (°)	90	180	270	359	250	102	205	308	411
No. of heavy atoms to search	16	16	16	16	8	12	12	12	12
No. of ΔFs used in search†	2115	2127	2102	2098	697	3842	4899	5196	5418
No. of solutions/total trials	0/500	0/500	13/500	14/500	66/500	3/500	21/500	42/500	87/500
CC of best trial (%)	16.6	14.4	38.4	42.4	37.2	16.0	20.1	27.1	30.4
⟨CC⟩ of other trials (%)	13.1	10.7	12.1	12.2	17.2	8.0	10.9	11.0	13.9
No. of sites used for phasing	—	—	17	17	8	10	10	10	10
⟨FOM⟩ after phasing (lr)‡	—	—	0.53	0.54	0.52 (0.63)§	0.43	0.44	0.48	0.49
⟨FOM⟩ after phasing (hr)¶	—	—	0.42	0.45	0.39 (0.46)	0.30	0.36	0.40	0.41
⟨FOM⟩ after dm†† (lr)‡	—	—	0.92	0.92	0.90 (0.94)	0.71	0.92	0.93	0.93
⟨FOM⟩ after dm†† (hr)¶	—	—	0.94	0.94	0.86 (0.90)	0.74	0.89	0.90	0.89
Electron-density correlation‡‡	—	—	0.86	0.87	0.71 (0.85)	0.32	0.69	0.78	0.81
No. of residues built/in total	—	—	108/121	108/121	123/163 (154/163)	0/537	499/537	500/537	496/537
No. of polypeptide chains	—	—	1	1	9 (3)	0	11	10	9

† Anomalous differences used for triplet generation. ‡ lr, low-resolution range (25.0–4.0 Å). § Values in parentheses refer to phase calculation using the SIRAS method. ¶ hr, high-resolution range (25.0–2.5 Å). †† dm, density-modification protocol. ‡‡ Correlation coefficient between experimental (after density modification) and refined electron-density maps.

Data sets that exhibited potentially correct results for the anomalous scatterer searches, as indicated by the large difference between the CCs from the best solution and all other trials, were used in phase calculation. Only major anomalous scatterer sites from the best *SHELXD* solution were input into *SHARP* (de La Fortelle & Bricogne, 1997) with their respective data set for SAD phase calculation. Whenever necessary, anomalous scatterer coordinates were inverted to assign the correct enantiomorph substructure. Density modification was performed with the program *SOLOMON* (Abrahams & Leslie, 1996) using the user-friendly interface implemented in *SHARP*. The resulting electron-density maps were used for automatic model building and iterative refinement with the programs *ARP*/*wARP* (Perrakis *et al.*, 1999) and *REFMAC* (Murshudov *et al.*, 1997), respectively. In the proteinase-inhibitor case, improved experimental phases were obtained when native and iodine-derivative data sets were combined in SIRAS phase calculation with *SHARP*. As a result of this, model building was continued with two distinct maps: the first at 2.20 Å resolution with SAD-derived phases and the second at 1.87 Å resolution with extended SIRAS-derived phases.

Detailed statistics obtained for each step from heavy-atom search to model building in all cases analyzed are shown in Table 2.

2.3. Anomalous signal estimates

The anomalous signal contribution from S atoms in the Nat-Acl-PLA$_2$ crystal and iodide ions in the I-Aaw-Ghy and I-Bba-Pin derivatives was assessed by calculating the ratio $\langle \Delta F_{\rm ANO} \rangle / \langle F \rangle$ for several resolution shells in each data set. Only non-centric reflections were used in these calculations. Whenever necessary, theoretical estimation of anomalous signal was obtained using the formula $\langle \Delta F_{\rm ANO} \rangle / \langle F \rangle = 2^{1/2} N_{\rm A}^{1/2} f'' / Z_{\rm eff} N_{\rm P}^{1/2}$ suggested by Hendrickson & Teeter (1981). Here, $N_{\rm A}$ is the number of anomalous scatterers in the asymmetric unit with anomalous contribution f''. As f'' depends on the atom type and on the wavelength of the incident X-rays, the program *CROSSEC* (Cromer, 1983) was used to estimate its value for each anomalous scatterer at the corresponding wavelength. $N_{\rm P}$ is the number of non-H atoms in the asymmetric unit and $Z_{\rm eff}$ is the mean normal scattering for atoms present in proteins, which assumes the value of 6.7 e$^-$.

The anomalous contribution from S atoms in hypothetical gene-product crystals was calculated using the same formula. The amino-acid sequences of annotated genes in *Bacillus subtilis*, *Escherichia coli* K12, *Vibrio cholerae* and *Xylella fastidiosa* genomes were obtained from the National Center of Biotechnology Information (NCBI) through the FTP site ftp://ftp.ncbi.nih.gov/genbank/genomes/.

3. Results and discussion

3.1. Checking for the presence of and locating anomalous scatterers

The scattering of X-rays arising from the anomalous scatterers present in protein crystals results in breakage of Friedel's law and the emergence of non-zero Bijvoet differences ($\Delta F^{\rm ANO} = |F_{hkl} - F_{\overline{hkl}}|$). As a result, the ratio $\langle \Delta F_{\rm ANO} \rangle / \langle F \rangle$ obtained for several resolution shells can indicate not only the presence of anomalous scatterers but also the amount of anomalous signal in each data set. Moreover, Bijvoet differences can be used with direct methods to determine the complete anomalous scatterer substructure within the asymmetric unit of a crystal, as implemented in *SHELXD*.

For Nat-Acl-PLA$_2$, in which naturally occurring S atoms are the only measurable source of anomalous scattering, the total anomalous signal is lower than 2% (Fig. 1). As data-set redundancy increases, the symmetry-related reflections are measured several times and the precision of the Bijvoet differences improves. Therefore, the anomalous signal, which

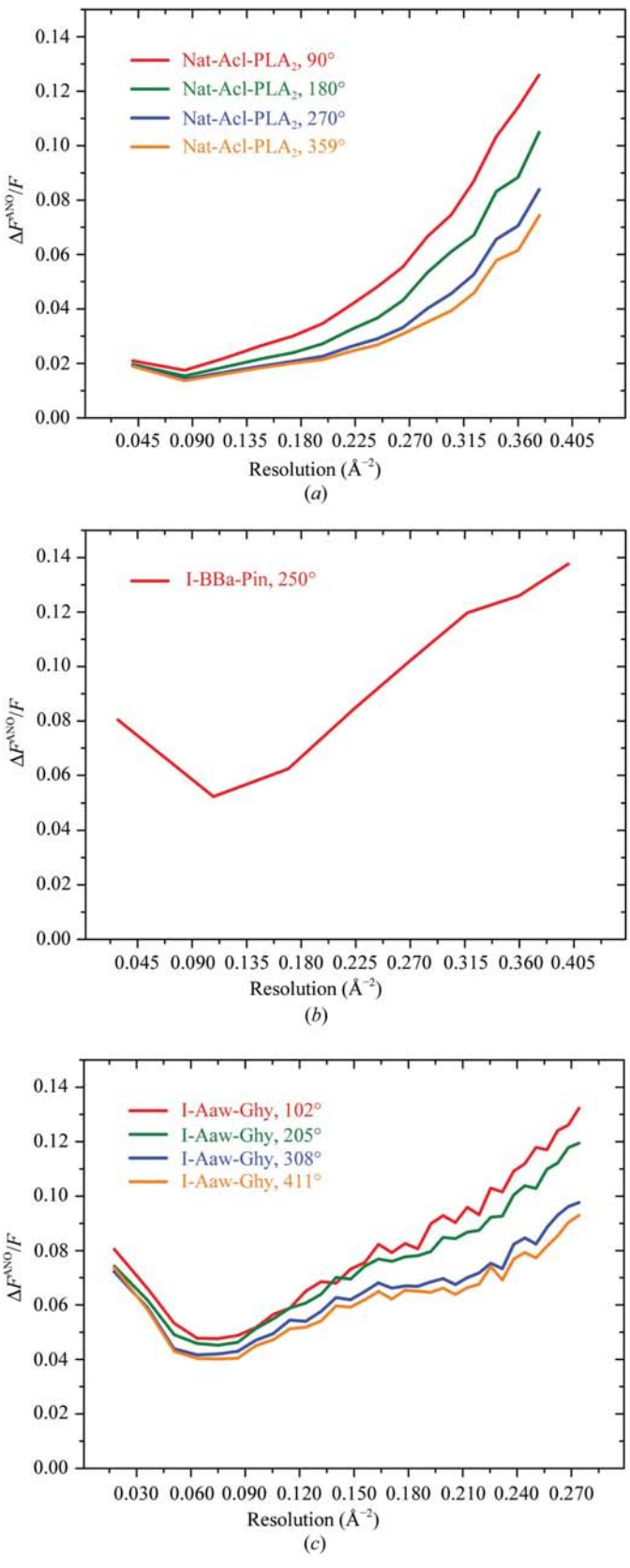

Figure 1
Anomalous signal estimates as a function of resolution for the Nat-Acl-PLA_2, I-Bba-Pin and I-Aaw-Ghy data sets. As the redundancy increases, the anomalous signal obtained for each data set can be measured with greater accuracy.

is overestimated at the beginning of data collection, approaches its theoretical value of 1.5% (corresponding to 16 fully occupied S atoms among 121 amino acids). Such redundancy-dependent behaviour has previously been observed (Dauter & Nagem, 2002) and, particularly for data sets with small anomalous signal, plays an important role during anomalous substructure identification, phasing and structure determination. This fact can be readily confirmed by analyzing the four Nat-Acl-PLA_2 subsets integrated through 90, 180, 270 and 359°. Even though all four subsets show very similar completeness and almost the same number of anomalous difference ΔF values could be obtained for each of them (Table 2), only the two most redundant subsets (270 and 359°) gave correct solutions for the anomalous scatterer substructure with *SHELXD*. Nevertheless, if appropriate resolution cutoffs were imposed, it might, in principle, be possible to locate major sulfur sites with the 180° or even 90° subsets. Surprisingly, in addition to the 16 sites expected for the 14 cysteine and two methionine residues of Acl-PLA_2, *SHELXD* was able to identify two extra sites. It was subsequently verified that the two extra sites correspond to a second conformation of one of the methionine residues and a sulfate ion which was derived from the crystallization solution.

When the Cu $K\alpha$ wavelength is used for data acquisition, S atoms display an anomalous contribution to the atomic scattering factor (f'') of only 0.56 e^-. As a result of this, the theoretical anomalous contribution from naturally occurring S atoms in both iodine derivatives tested is less than 0.5% and hence most of the anomalous contribution is provided by I atoms. The total anomalous signal of the I-Bba-Pin and I-Aaw-Ghy derivatives is plotted as a function of resolution in Fig. 1.

It is interesting to note that the I-Bba-Pin derivative, which has an asymmetric unit content slightly larger than that of Nat-Acl-PLA_2, exhibits almost a fourfold greater anomalous signal. This fact can be explained not only by the incorporation of several iodide anions into the crystallographic structure, but also by the fact that I atoms show an anomalous scattering contribution that is 14 times stronger than that of S atoms at the same wavelength (f'' = 6.8 e^- at 1.54 Å). The amount of anomalous signal present in this data set could in principle suggest that no more than two I atoms were fully incorporated into the crystallographic structure. However, in practice, eight partially occupied major iodine sites were correctly identified in the I-Bba-Pin asymmetric unit by *SHELXD* (Table 2).

In the case of I-Aaw-Ghy, identification of the iodine substructure was also straightforward, as suggested by the contrast in correlation coefficients presented in Table 2. Careful inspection of the results obtained from each of the *SHELXD* jobs (102, 205, 308 and 411°) indicates that a higher contrast in CC values is observed within the last two subsets. A combination of increasing completeness and redundancy is most likely to be the main reason for these results. On the basis of the final refined structure, it was verified that all top ten iodine sites in each *SHELXD* solution were correct sites, with the exception of the first subset (102°), where a mixture

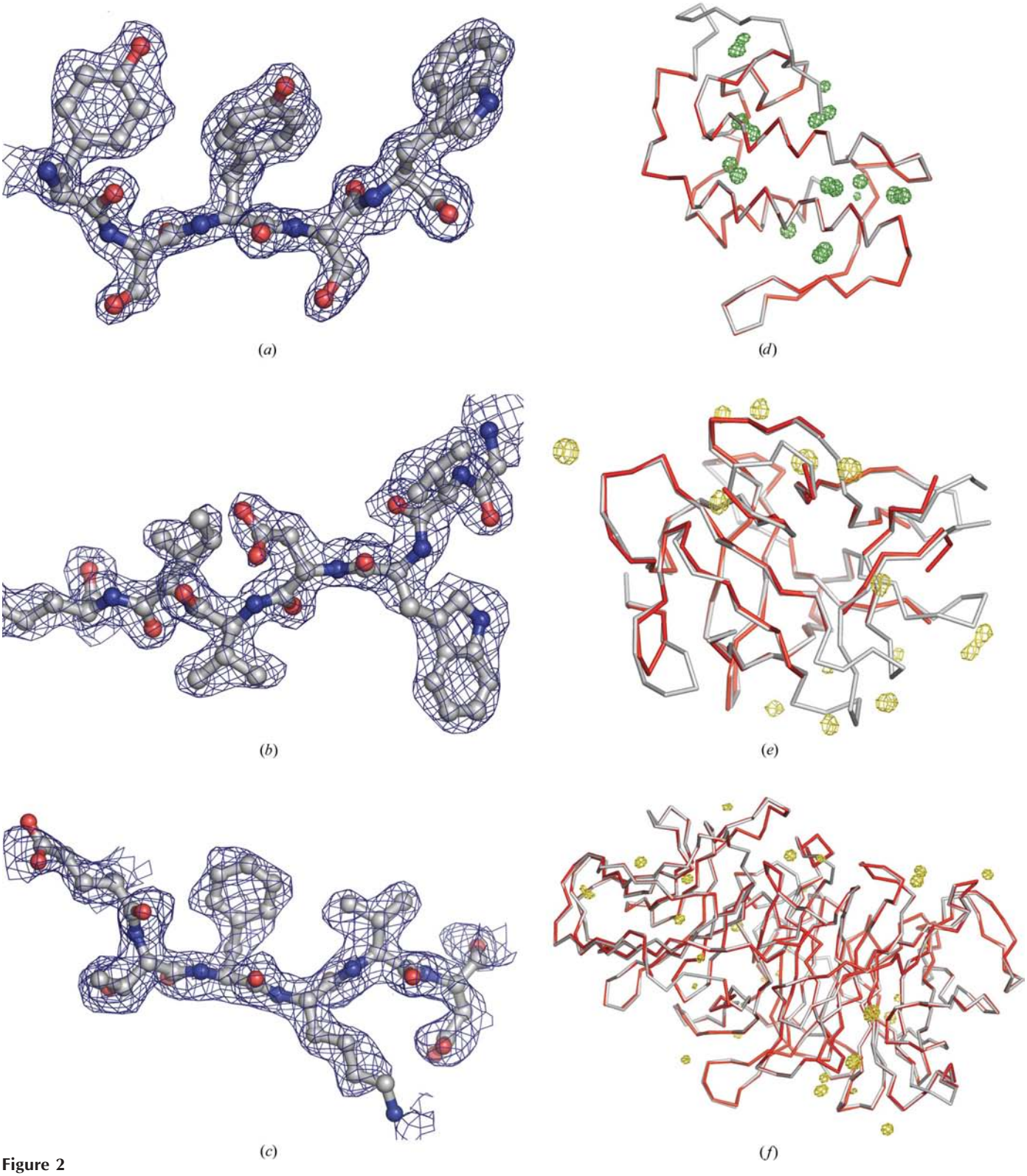

Figure 2
Automatic interpretation of experimental electron-density maps. (*a*) SAD-derived electron-density map from the Nat-Acl-PLA$_2$ most redundant data set after the density-modification protocol. (*b*) Electron-density map for Bba-Pin after density modification using SIRAS phases at 1.87 Å resolution. (*c*) Map obtained for the I-Aaw-Ghy data set after density modification using 411° of data. In all cases representative parts of refined models are shown. (*d*) C^{α} trace of Nat-Acl-PLA$_2$ hybrid model (red) and the refined model (grey) after manual construction of the C-terminal portion and positional refinement. (*e*) C^{α} trace of the Bba-Pin hybrid model (red) obtained with the experimental map at 2.20 Å resolution superimposed on the final model (grey) refined against the high-resolution data set. (*f*) C^{α} representation of the I-Aaw-Ghy hybrid model (red) obtained using 205° diffraction data and the final model (grey) refined against the most redundant data set (411°). Anomalous maps obtained with phases generated on the basis of final models are shown in green (S atoms) and gold (I atoms). All figures were prepared using the program *PyMol* (DeLano, 2002).

of correct (top six) and incorrect (last four) sites was obtained. This result indicates that even though the data set demonstrates a high anomalous signal, high completeness and also moderate to high redundancy are essential for the location of anomalous scatterers with low occupancy.

3.2. Structure determination

An initial Nat-Acl-PLA_2 hybrid model (polypeptide fragments and dummy atoms) was obtained with the program *ARP*/*wARP* using either of the last two data sets (270 and 359°) with its respective experimental electron-density map. In the two cases tested, very similar qualitative and quantitative results were obtained. However, only the 359° subset was used for subsequent steps of manual construction, side-chain docking, water picking and positional refinement, aiming towards further structural studies.

In the proteinase-inhibitor case, the resulting set of experimental phases obtained both by SAD and SIRAS methods were submitted to the *ARP*/*wARP* program in two different jobs. As a result of this, initial hybrid models of the proteinase inhibitor were built within two distinct maps: the first at 2.20 Å resolution and the second at 1.87 Å resolution. Automatic interpretation of the lower resolution map took less than 30 min, rendering a hybrid model consisting of 123 amino-acid residues distributed into nine polypeptide fragments. This partial model represents 75% of the total structure. Further automatic model building under the same initial conditions did not lead to the improvement of the hybrid model. Nevertheless, better results were obtained using the higher resolution map. When SIRAS-derived phases were used, automatic model building was extremely fast (a few minutes) and around 95% of the total structure was built within the first four building cycles.

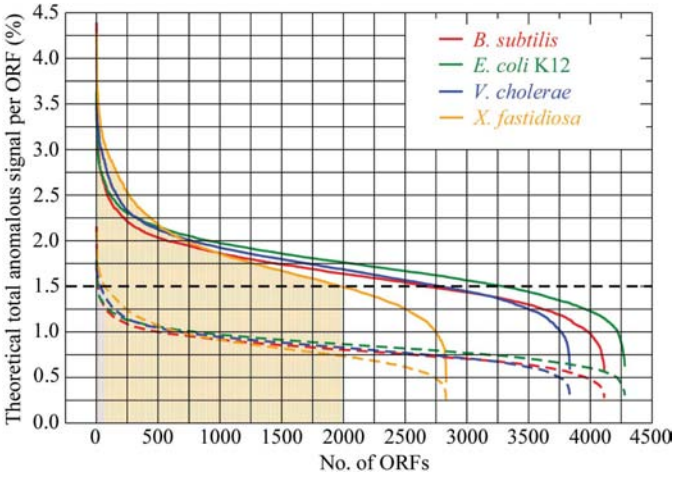

Figure 3
Theoretical anomalous signal from S atoms present in expected gene products from *B. subtilis*, *E. coli* K12, *V. cholerae* and *X. fastidiosa* genomes. Anomalous contributions calculated at the Cu $K\alpha$ (dashed lines) and Cr $K\alpha$ (filled lines) wavelengths are shown. Filled areas illustrate the number of ORFs amendable to S-SAD structure determination using Cu $K\alpha$ (in brown) and Cr $K\alpha$ (in orange).

Determination of I-Aaw-Ghy hybrid models was also performed automatically. However, it is interesting to note that even though all four data sets (102, 205, 308 and 411°) showed reasonable results during the heavy-atom search and phasing, the first data set, with the lowest completeness and redundancy, was the only one which did not render sufficient information for convergent automatic model building and refinement. In the last three cases, practically the same number of amino-acid residues were built into each experimental electron-density map into a similar number of polypeptide chains. The basic difference between the results obtained for the last three data sets was the time required for model building. While automatic model building of this 60 kDa protein structure using the most redundant data set (411°) took less than 5 min, similar results were obtained within hours when the second data set (205°) was used. The main reason for this difference is the quality of each experimental electron-density map used for model building.

In Fig. 2, representative parts of the refined models together with the experimental maps used for automatic model building are given (Figs. 2*a*, 2*b* and 2*c*). A comparison between automatically built hybrid models and refined models for all three proteins tested are also shown (Figs. 2*d*, 2*e* and 2*f*).

3.3. Genomic scale analysis

The choice of the incident X-ray wavelength is one of the major decisions required for data acquisition. The decision relies on the evaluation of several wavelength-dependent factors, such as absorption, radiation damage, photon flux, resolution limit and anomalous signal. In fact, the incident X-ray wavelength is chosen as a compromise between all these factors and the purpose of data acquisition (Polikarpov *et al.*, 1997; Teplyakov *et al.*, 1998). However, as X-ray home sources produce monochromatic radiation, optimization of these parameters by changing the wavelength is not an option and alternative procedures should be applied. In cases where anomalous signal is essential for structure determination, an increase in the anomalous scattering contribution is achieved by selecting appropriate anomalous scatterers for derivatization that match the monochromatic X-ray home-source radiation and by collecting highly complete and redundant data.

It has been suggested for a long time that S atoms could be used as the only source of anomalous signal in protein crystallography (Wang, 1985). From the biochemical point of view, this approach has two major advantages: firstly, S atoms are naturally present in proteins, which turns native crystals into derivatives, and secondly, the occupancy of sulfur sites is equal to 1 (although in some cases disordered/flexible S atoms may behave as if their occupancy was lower than 1). On the other hand, this approach has a clear disadvantage: S atoms provide a small anomalous contribution at the Cu $K\alpha$ wavelength (0.56 e^-). Nevertheless, it is known that a few proteins display a high sulfur content which increases the total anomalous contribution to a level sufficient for structure elucidation,

provided high-resolution diffraction data with adequate completeness and redundancy are available.

In this work, we demonstrate that an anomalous contribution of 1.5% is sufficient for structure elucidation of well diffracting crystals using the sulfur-SAD technique with data sets collected in-house. Similar results, sometimes with a somewhat lower or a slightly higher anomalous signal, have been obtained by other crystallographers (Ramagopal *et al.*, 2003 and references therein). However, how many proteins or gene products present in elucidated genomes display a theoretical anomalous contribution above 1.5%?

In Fig. 3, we show theoretical calculations of the anomalous contribution from S atoms at the Cu $K\alpha$ wavelength for each open-reading frame (ORF) in four different bacterial genomes. Unfortunately, the number of expected gene products with anomalous signal higher than 1.5% does not exceed 2.5% of the total number of ORFs in all four genomes evaluated in this work. Examining, for instance, all 2832 *X. fastidiosa* ORFs, approximately 97.5% of them (2760) show a theoretical anomalous contribution of less than 1.5%, while only 72 gene products display higher values. It is important to mention that our estimate does not take into account the fact that a number of proteins will have metal ions (such as Zn, Fe, Cu *etc.*) bound to them which would significantly increase their anomalous signal. In addition, considering that in some special cases an anomalous signal of around 1% has been successfully used for structure determination, more gene products could be solved by the sulfur-SAD approach, at least from the point of view of required anomalous signal. However, taking into consideration that this level of anomalous contribution requires high or even atomic resolution data, our estimates most probably represent an upper limit of proteins in the genome amenable to sulfur-SAD structure determination.

The possibility of using soft X-ray radiation obtained from specially designed chromium systems (home sources with Cr rotating anode, confocal multilayer optics designed for Cr $K\alpha$ radiation, helium beam path and large-aperture detector; Yang *et al.*, 2003) can increase the applicability of this approach. At the Cr $K\alpha$ wavelength, S atoms display twice the anomalous signal ($f'' = 1.14$ e$^-$) than at the Cu $K\alpha$ wavelength. Surprisingly, however, the number of gene products with total anomalous signal superior than 1.5% is increased by a factor that can reach two orders of magnitude (Fig. 3). Under these conditions, around 1982 *X. fastidiosa* ORFs would naturally display an anomalous signal higher than 1.5%. Our purely theoretical analysis may give an overoptimistic estimate of the benefits of soft X-ray radiation. Nevertheless, the results allow one to consider the possibility of phasing and determining the structure of a considerable number of proteins using the S-SAD approach with copper and/or chromium rotating-anode home sources.

3.4. Obtaining more from rotating anodes

Apart from the potential success of the sulfur-SAD approach, it is clear that a considerable number of macromolecules will still require preparation of derivatives, either because they do not display a high anomalous signal or because in-house data quality is not good enough for phase determination. Preparation of isomorphous derivatives through the quick cryo-soaking approach using halides seems to be an appropriate option in such cases. The introduction of negatively charged anomalous scatterers (such as iodide) into the crystal structure is achieved during a short soak (30–300 s) of the native crystal in a cryogenic solution containing in addition a high concentration of an appropriate salt. The procedure is performed during preparation of the crystal for data collection. It is cheap and not time-consuming.

The use of iodine compounds (such as sodium iodide and potassium iodide) for fast derivatization has some advantages in comparison with traditional transition-metal reagents (potassium tetrachloroplatinum, potassium dicyanoaurate and mercury acetate). Although iodide anions interact with water molecules through hydrogen bonds, they do not form strong complexes or bind covalently to certain chemical functions of the protein. As a result, some protein crystals, including Bba-Pin and Aaw-Ghy, can undergo derivatization soaks at high concentration of such reagents without damaging their crystalline order. Additionally, as diffusion through the crystal channels takes place very rapidly, a great number of sites around the protein surface are filled with iodide anions (Fig. 2). Most frequently, these ions replace ordered water molecules and their occupancy results from the competition between them. Since I atoms possess a strong anomalous signal at the copper $K\alpha$ wavelength, comparable to that of platinum (6.9 e$^-$), gold (7.3 e$^-$) and mercury (7.7 e$^-$), the contribution of all partially occupied sites can dramatically increase the total anomalous signal of the crystal, as observed in Fig. 1.

In application to structural genomics, the total anomalous signal of both iodine derivatives obtained in the present study is much higher (above 5%) than the theoretical anomalous signal from S atoms in all gene products analyzed, which makes it instrumental for structure determination of protein crystals that diffract to medium and low resolution. For example, as observed for the I-Bba-Pin data set, a strong anomalous signal facilitated the search for anomalous scatterers and phase determination.

In principle, the use of an in-house chromium radiation source ($\lambda = 2.29$ Å) would further increase the strength of anomalous scattering of I atoms. At this wavelength, they display an f'' of 12.8 e$^-$! However, special attention must be taken during data acquisition and data processing in order to avoid radiation damage and to include absorption corrections from the mounting loop and the solvent. This is especially true for quick cryo-soaked crystals, where a high concentration of heavy atoms is found in the non-crystalline phase.

The quick cryo-soaking approach can also be used with positively charged anomalous scatterers (Nagem *et al.*, 2001) such as alkali metals and lanthanides. Indeed, during the course of this work, the structure of a 2×56 kDa enzyme was determined in our laboratory using such an approach (Mantovani *et al.*, manuscript in preparation). This provides additional flexibility in the process of derivatization and

increases the chances of success in phase determination, since anions and cations naturally occupy different interaction sites on protein surfaces and display different affinities which depend on the protein surface-charge distribution and the pH of the cryo-derivatization solution. Additionally, different heavy-ion sites permit the combined use of derivatives in MIR(AS) phasing, even when none of them are strong enough to produce interpretable electron-density maps *via* the SAD or SIR(AS) methods.

In this work, a simple consecutive data-acquisition strategy has been applied. It has been suggested, however, that synchrotron data involving anomalous signal should be collected in such a way that Friedel mates are measured close in time (Hendrickson & Ogata, 1997). This approach takes into account the fact that small anomalous differences cannot be determined precisely if errors from radiation damage accumulate in measured intensities. This strategy can be applied to home-source data acquisition, particularly when the crystals belong to a low-symmetry space group or suffer from considerable radiation damage.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

Ambrosio, A. L. B., Nonato, M. C., Selistre-de-Araujo, H. S., Arni, R., Ward, R. J., Ownby, C. L., de Souza, D. H. F. & Garratt, R. C. (2005). *J. Biol. Chem.* **280**, 7326–7335.

Bijvoet, J. M. (1949). *Nature (London)*, **173**, 888–891.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. New York: Academic Press.

Bond, C. S., Shaw, M. P., Alphey, M. S. & Hunter, W. N. (2001). *Acta Cryst.* D**57**, 755–758.

Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2003). *Acta Cryst.* D**59**, 662–669.

Cromer, D. T. (1983). *J. Appl. Cryst.* **16**, 437–438.

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. J. (2002). *Acta Cryst.* D**58**, 494–506.

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

Dauter, Z. & Nagem, R. A. P. (2002). *Z. Kristallogr.* **217**, 694–702.

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

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

Debreczeni, J. É., Girmann, B., Zeeck, A., Krätzner, R. & Sheldrick, G. M. (2003). *Acta Cryst.* D**59**, 2125–2132.

DeLano, W. L. (2002). *The PyMol Molecular Graphics System.* Delano Scientific, San Carlos, CA, USA.

Drenth, J. (1999). *Principles of Protein X-ray Crystallography*, 2nd ed. Heidelberg: Springer.

Evans, G. & Bricogne, G. (2002). *Acta Cryst.* D**58**, 976–991.

Evans, P. R. (1997). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **33**, 22–24.

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

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. R. Soc. London Ser. A*, **225**, 287–307.

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

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

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

Ioerger, T. R. & Sacchettini, J. C. (2002). *Acta Cryst.* D**58**, 2043–2054.

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

Laue, M. von (1916). *Ann. Phys.* **50**, 433–446.

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

Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.

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

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

Miller, R., De Titta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science*, **259**, 1430–1433.

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

Nagem, R. A. P., Dauter, Z. & Polikarpov, I. (2001). *Acta Cryst.* D**57**, 996–1002.

Nagem, R. A. P., Polikarpov, I. & Dauter, Z. (2003). *Methods Enzymol.* **374**, 120–137.

Nagem, R. A. P., Rojas, A. L., Golubev, A. M., Korneeva, O. S., Eneyskaya, E. V., Kulminskaya, A. A., Neustroev, K. N. & Polikarpov, I. (2004). *J. Mol. Biol.* **344**, 471–480.

Olsen, J. G., Flensburg, C., Olsen, O., Bricogne, G. & Henriksen, A. (2004). *Acta Cryst.* D**60**, 250–255.

Olsen, J. G., Flensburg, C., Olsen, O., Seibold, M., Bricogne, G. & Henriksen, A. (2004). *Acta Cryst.* D**60**, 618.

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

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

Polikarpov, I., Teplyakov, A. & Oliva, G. (1997). *Acta Cryst.* D**53**, 734–737.

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

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

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

Stevens, R. C. (2000). *Curr. Opin. Struct. Biol.* **10**, 558–563.

Teplyakov, A., Oliva, G. & Polikarpov, I. (1998). *Acta Cryst.* D**54**, 610–614.

Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1937–1940.

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

Watanabe, N., Murai, H. & Tanaka, I. (2002). *Acta Cryst.* D**58**, 1527–1530.

Yang, C., Pflugrath, J. W., Courville, D. A., Stence, C. N. & Ferrara, J. D. (2003). *Acta Cryst.* D**59**, 1943–1957.