Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the routine use of soft X-rays in macromolecular crystallography. Part V. Molecular replacement and anomalous scattering

Johan Unge,‡ Christoph Mueller-Dieckmann,§ Santosh Panjikar, Paul A. Tucker, Victor S. Lamzin and Manfred S. Weiss*¶

EMBL Hamburg Outstation, c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany

‡ Present address: Lund University, MAX-lab, PO Box 118, 221 00 Lund, Sweden.
§ Present address: ESRF, 6 Rue Jules Horowitz, BP 220, F-38043 Grenoble CEDEX, France.
¶ Present address: Helmholtz-Zentrum Berlin, Macromolecular Crystallography (HZB-MX), Albert-Einstein-Strasse 15, D-12489 Berlin, Germany.

Correspondence e-mail: msweiss@helmholtz-berlin.de

Received 25 January 2011
Accepted 24 June 2011

Currently, about two thirds of all new macromolecular structures are determined by molecular replacement. In general the method works reliably, but it reaches its limits when the search model differs too much from the target structure in terms of coordinate deviations or completeness. Since anomalously scattering substructures are better conserved than the overall structure, these substructures and the corresponding anomalous intensity differences can be utilized to enhance the performance of molecular-replacement approaches. It is demonstrated that the combined and concomitant use of structure-factor amplitudes and anomalous differences constitutes a promising approach to push the limits of molecular replacement and to make more structures amenable to structure solution by this technique.

1. Introduction

The past decade has seen an increasing number of experiments in macromolecular crystallography performed at wavelengths longer than 1.54 Å (corresponding to Cu $K\alpha$ radiation). The main motivation for collecting diffraction data at these wavelengths is the possibility of phase determination based on the natively present S and P atoms in biomolecules. Although the sulfur and phosphorus K absorption edges are not easily accessible experimentally, it has been demonstrated a number of times that accurately measured anomalous difference data at wavelengths of around 2.0 Å or less can be used for phase determination in a single-wavelength anomalous diffraction (SAD) approach. The first such experiment was carried out by Hendrickson & Teeter (1981), who solved the structure of the small protein crambin (46 amino acids, six S atoms) based solely on anomalous difference data collected using a Cu $K\alpha$ source. 18 years later, Dauter and coworkers demonstrated that the structure of hen egg-white lysozyme (129 amino acids, ten S atoms) could be successfully solved based on accurately measured anomalous differences collected at λ = 1.54 Å at a synchrotron source (Dauter *et al.*, 1999). In the latter case, the structure solution was aided by the presence of seven additional chloride ions bound to the surface of the protein. From this point onwards, a number of test cases as well as a significant number of new structures have been determined by the sulfur-SAD (S-SAD) approach based on diffraction data collected both at home and at synchrotron sources.

Owing to the spectral properties of synchrotron radiation, wavelengths longer than Cu $K\alpha$ are relatively easily available

at many synchrotron beamlines (Djinovic Carugo *et al.*, 2005). Despite this fact, the use of longer wavelengths to enhance the anomalous scattering from light atoms has only been investigated systematically in the last decade (Weiss, Sicker, Djinovic Carugo *et al.*, 2001; Weiss, Sicker & Hilgenfeld, 2001; Mueller-Dieckmann *et al.*, 2004, 2005). From the experimental point of view, the increased anomalous scattering at longer wavelengths appears to be counteracted by the increased level of noise in the processed data, for instance resulting from increased absorption of the X-rays by the sample. Therefore, Mueller-Dieckmann *et al.* (2005) concluded from a study involving 74 diffraction data sets that for current state-of-the-art synchrotron beamlines the optimum wavelength to measure these small differences is around 2.0 Å. Probably the first *de novo* structure solved using S-SAD at the longer wavelength of $\lambda = 1.74$ Å was that of the photoprotein obelin (Liu *et al.*, 2000). Since then, a few dozen other structures have been solved by S-SAD based on data collected at longer wavelengths.

It has been claimed for some time that the use of longer X-ray wavelengths could become a useful general tool in macromolecular crystallography. For instance, the potential of a chromium anode ($\lambda = 2.29$ Å for Cr $K\alpha$ radiation) was discussed more than 50 years ago by Blow (1958), but apart from a few scattered experiments (*e.g.* Anderson *et al.*, 1996; Kwiatkowski *et al.*, 2000) it never really caught on until about 8–10 years ago, when the first experiments utilizing longer wavelengths were carried out at synchrotrons. In response to this trend, Rigaku/MSC brought a rotating anode made of chromium (Yang *et al.*, 2003) onto the market, making such experiments possible in home laboratories as well (Nan *et al.*, 2009).

A further benefit of having available a diffraction data set collected at longer wavelengths is that the anomalously scattering substructure can be unequivocally determined after the structure has been refined (Einspahr *et al.*, 1985; Weiss, Sicker, Djinovic Carugo *et al.*, 2001; Weiss *et al.*, 2002; Kuettner *et al.*, 2002; Ferreira *et al.*, 2004; Sekar *et al.*, 2004; Mueller-Dieckmann *et al.*, 2007; Raaf *et al.*, 2008).

Following its inception in the 1960s (Rossmann & Blow, 1962), molecular replacement (MR) has become a widely used method for successful structure determination in cases where a homologous structure is already known. Many programs have become available which employ different approaches and algorithms for MR, including various Patterson techniques (*e.g.* Rossmann & Blow, 1962; Huber, 1965; DeLano & Brünger, 1995), structure-factor correlation (Navaza, 1987) and the use of statistical targets (Bricogne, 1992, 1997; Read, 2001; McCoy *et al.*, 2007). More recently, several automatic structure-determination systems have been assembled which use MR in informed pipeline approaches (Panjikar *et al.*, 2005, 2009; Keegan & Winn, 2007; Long *et al.*, 2008).

Years ago, it was suggested that the number of protein folds was limited (Chothia, 1992), and several estimates have been made as to how many folds naturally occurring proteins may be sorted into (see, for example, Liu *et al.*, 2004 and references therein). Estimates range from 400 to 10 000. Although it is not clear whether all theoretically possible folds can be found in nature, and although the definition of a fold itself is qualitative rather than quantitative, the number of new folds deposited in the PDB (*e.g.* as defined by the SCOP classification; Murzin *et al.*, 1995) has been decreasing on a yearly basis, in stark contrast to the ever-increasing number of structures deposited each year (Berman *et al.*, 2000). An analysis of the PDB showed that as many as two thirds of the structures solved in the year 2006 were solved using MR techniques (Long *et al.*, 2008). As a consequence of the high probability of new proteins or their constituents being similar to those already known, together with the many advances in software over the years, the use of MR to solve new structures is steadily increasing.

MR works reliably and well when the search model and the target structure are rather similar. However, the less similar the search model and the target structure are, the lower the probability that MR can successfully be applied. Differences may result from, for example, sequence discrepancies or differences in functional state or simply because the search model comprises only part of the target structure. A high degree of noncrystallographic symmetry (NCS) also decreases the direct correspondence between the search model and the crystal asymmetric unit, blurring the MR evaluation function. Owing to the multiple local minimum nature of the rotation function (RF) and the translation function (TF), pinpointing the correct MR solution among a large number of possible candidate solutions is not always straightforward.

The use of the anomalous dispersion in combination with MR structure solution is sometimes also referred to as MRSAD (molecular replacement with single-wavelength anomalous diffraction). Its use has so far been limited to comparing the anomalous substructure (obtained from the MR solution directly or by the usual SAD or MAD phasing techniques) with the MR solution in order to assess whether the anomalous substructure is consistent with the MR solution (Schuermann & Tanner, 2003; Madauss *et al.*, 2004). However, this approach is tedious and is limited to cases in which a manageable list of MR candidate solutions is available, which may then be analysed one by one. Another approach is to use the anomalous substructure and the anomalous differences together to generate an electron-density map which is unbiased by the MR search model (Baker *et al.*, 1995). In this way, the possibility of combining two independent phase sets often leads to more effective structure determination. Recently, this idea has been automated and implemented in the *Auto-Rickshaw* structure-determination pipeline (Panjikar *et al.*, 2009).

This manuscript demonstrates, in a test-of-concept approach, the beneficial use of the anomalous differences together with the structure-factor amplitudes in MR. We would like to suggest its general applicability wherever an anomalously scattering substructure is available and whenever appropriate measures have been taken to carefully record the anomalous differences. It appears that this method has the potential to expand the current limits of the technique, making more structures amenable to determination by MR.

Table 1
Summary of the 23 diffraction data sets and the corresponding refined structures used in this study.

The structures were collected at a wavelength of 2.0 Å on beamline X12 at EMBL Hamburg, DESY, Germany (Mueller-Dieckmann *et al.*, 2007).

Protein	Space group	Resolution limits (Å)	No. of molecules in ASU	No. of protein atoms in ASU	No. of anomalously scattering atoms in ASU†	R_{anom} (%)	$R_{p.i.m.}$ (%)	PDB code
Apoferritin	$F432$	99–2.00	1	1364	16	2.9	n.d.	2g4h
Concanavalin A	$I222$	99–2.40	1	1809	8	2.2	2.1	2g4h
Glucose isomerase	$I222$	99–1.85	1	3049	11	3.6	4.2	2g4j
hARH3	$P2_12_12_1$	99–1.82	1	2607	19	1.9	1.8	2g4k
HEL-45	$P4_32_12$	99–1.84	1	1000	17	2.0	1.0	2g4p
HEL-80	$P4_32_12$	99–1.84	1	1001	16	1.9	1.1	2g4q
HNL	$C222_1$	99–1.84	1	2118	16	1.9	2.1	2g4l
Insulin	$I2_13$	99–1.80	1	411	6	2.2	1.1	2g4m
α-Lactalbumin	$P2_12_12$	99–2.30	6	5856	50	3.7	3.3	2g4n
LeuB	$P2_12_12_1$	99–2.00	4	10038	32	1.9	2.1	2g4o
MogA	$P2_1$	99–1.92	1	3180	7	3.0	3.3	2g4r
NBR1 PB1	$P6_322$	99–2.15	1	691	5	1.3	1.1	2g4s
PPE-Ca	$P2_12_12_1$	99–1.84	1	1845	13	1.8	1.3	2g4u
PPE-Na	$P2_12_12_1$	99–2.15	1	1831	11	1.5	1.3	2g4t
Proteinase K	$P4_32_12$	99–2.14	1	2031	15	1.3	1.1	2g4v
RNAse A ($C2$)	$C2$	99–1.84	2	1902	25	3.3	3.4	2g4w
RNAse A ($P3_221$)	$P3_221$	99–1.95	1	951	18	3.3	3.0	2g4x
Thaumatin	$P4_12_12$	99–1.98	1	1557	17	1.9	1.7	2g4y
Thermolysin	$P6_122$	99–1.98	1	2437	16	1.8	1.1	2g4z
Titin (A168-A169)	$I222$	99–2.20	1	1532	6	1.9	1.7	2ill
Trypsin ($P1$)	$P1$	99–1.84	1	1553	10	2.4	2.2	2g51
Trypsin ($P2_1$)	$P2_1$	99–1.84	1	1551	9	2.2	1.8	2g52
Trypsin ($P3_121$)	$P3_121$	99–1.82	1	1626	18	1.9	1.3	2g54

† Reported are those atoms which have been found and described in Mueller-Dieckmann *et al.* (2007).

2. Materials and methods

2.1. The diffraction data sets and refined structures

The 23 data sets and refined structures were taken from a previously published study on the detection of anomalously scattering substructures (Mueller-Dieckmann *et al.*, 2007). In order to maximize the anomalous signal, each of the data sets had been collected at a wavelength of 2.0 Å on EMBL beamline X12 (DESY, Hamburg, Germany). In all cases, 360° of data were collected, resulting in completeness values of over 99% for 17 of the 23 data sets and of 92–99% for another five. Only the trypsin ($P1$) data set exhibited a lower overall completeness of 88%. The Bjivoet redundancy values ranged from 1.5 for trypsin ($P1$) to 35.5 for apoferritin (see also Table 1 of Mueller-Dieckmann *et al.*, 2007). For each of the 23 cases the corresponding structure has been determined, refined and deposited in the PDB together with the structure-factor amplitudes and the anomalous differences. The data sets and their main characteristics are listed in Table 1.

2.2. Preparation of the search models

The refined protein structures, and the corresponding anomalous substructures, were used to generate the rotation searches for reference. The anomalous substructures included protein S atoms as well as bound ligands with a measurable anomalous signal: calcium, cadmium, chlorine, *S*,*S*-(2-hydroxyethyl)thiocysteine (abbreviated HEC or MEC), manganese, potassium, sulfate and zinc. For creation of the search models, the coordinates were altered by applying random errors using *MOLEMAN* (Kleywegt, 1996). Search models with 1.0, 1.5 and 2.0 Å coordinate displacements for all non-H atoms were generated in this way. Exactly the same coordinate shifts, in terms of both absolute value and direction, were applied to the anomalously scattering substructures, which included all atoms with a detectable anomalous signal. In order to simulate a real-case scenario using the protein test case PPE-Ca, related structures were identified using *BLAST* (Altschul *et al.*, 1990) and used as templates. From the multiple hits obtained, a subset of structures with different degrees of sequence similarity was retrieved and three structures were selected as potential models: salmon pancreatic elastase (PDB entry 1elt; Berglund *et al.*, 1995) with 67% sequence identity, bovine chymotrypsinogen C (PDB entry 1pyt, chain *D*; Gomis-Rüth *et al.*, 1995) with 54% sequence identity and human β-tryptase (PDB entry 1a0l; Pereira *et al.*, 1998) with 38% sequence identity. Although the structures contained different sets of anomalous scatterers from particular crystallization conditions, for consistency none of the nonprotein anomalous scatterers found in the model structure files were used in the anomalous rotation search.

2.3. Calculation of the rotation function

Rotation functions (RFs) were computed using *AMoRe* v.6.0 (Navaza, 2001). For each data set, two fast cross-RFs were calculated and output as three-dimensional maps in CCP4 format (Winn *et al.*, 2011). Firstly, a cross-RF using all atoms of the search model and the observed structure-factor amplitudes F_P was computed (RF_{prot}), and then a cross-RF using only the anomalously scattering atoms of the search model and the anomalous differences Δ_{anom} (RF_{anom}) (Fig. 1). In all cases, the low-resolution limit was set to 99 Å and the high-resolution limit to 2.5 Å. The *AMoRe* SAMPLE SCALE

parameter was adjusted between 4 and 8 for the protein calculation and 4 and 30 for the substructure calculation. The sphere radius and step angles were 25 Å and 3.0°, respectively. These *AMoRe* input parameters proved to be appropriate and yielded suitable RFs for all calculations using both the structure-factor amplitudes and the anomalous differences.

2.4. Map combination

The histograms of the RF maps from *AMoRe* obeyed or closely followed a Gaussian distribution (data not shown). Hence, no normalization was performed. However, in order to be able to compare the two RF maps, they were calculated on the same grid and with the same dimensions. For analysis, the two RFs were combined according to (1) using values for the weight of the anomalous RF, $w_{\rm anom}$, from 0.0 to 1.0 in steps of 0.1. The peaks occurring in the combined RF ($\rm RF_{comb}$) were then extracted. The minimum angle difference allowed between two peaks was set to 10°.

$$\mathrm{RF_{comb}}(\alpha\beta\gamma) = (1 - w_{\rm anom})\mathrm{RF_{prot}}(\alpha\beta\gamma) + w_{\rm anom}\mathrm{RF_{anom}}(\alpha\beta\gamma). \quad (1)$$

As the RF maps from *AMoRe* often contain streaky features close to the origin, care was taken that none of the angles was closer than 10–15° (depending on the resolution) to the origin. This was achieved by trying two random orientations of the search model for each rotation search. In the case where both orientations yielded a result, the better of the two was then used in subsequent analysis. All calculations were performed using a Perl script, which was written to automatically perform the tests, to combine the RF maps and to evaluate the results.

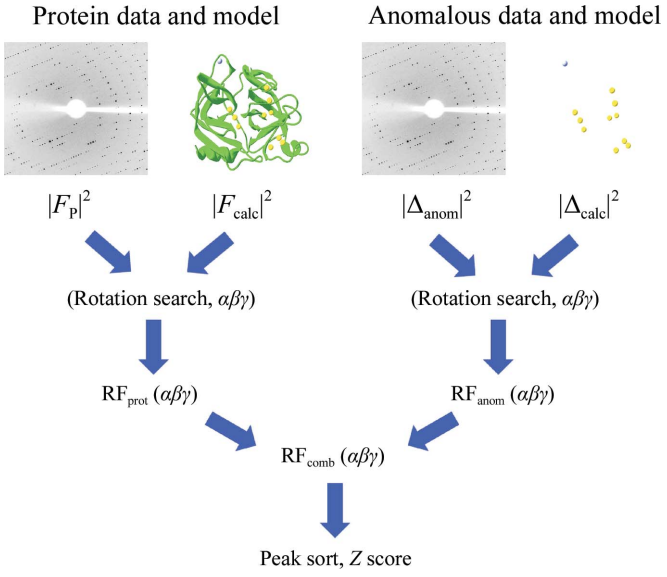

Figure 1
Schematic illustration of the approach proposed here. The combined rotation function $\rm RF_{comb}$ is calculated from the rotation functions based on the real structure factors ($\rm RF_{prot}$) and on the anomalous differences ($\rm RF_{anom}$). $\rm RF_{prot}$ is calculated from the observed structure-factor amplitudes and the calculated structure-factor amplitudes of the corresponding protein model. The anomalous rotation function $\rm RF_{anom}$ is calculated based on the experimental anomalous differences and the corresponding anomalous substructure of the model.

2.5. Scoring

Having a refined model available for each data set, the correct rotation in the RF map was identified in an initial run with no displacements applied to the coordinates. The Eulerian angles of the correct solution were then used as a reference in the evaluation. The highest-ranking peaks were identified for evaluation. For each observed false peak the minimal rotation required to overlap with the correct peak was calculated and peaks closer than a given minimum value were discarded. The combined RFs were then normalized and their Z scores [as defined in (2)] were used to rank peaks in the lists.

$$Z_{\mathrm{RF}(\alpha\beta\gamma)} = [\mathrm{RF}(\alpha\beta\gamma) - \langle \mathrm{RF}(\alpha\beta\gamma)\rangle]/\sigma_{\rm RF} \quad (2)$$

In order to find the optimum weights for combining the two RFs, the maximum Z score and the highest ranking position of the correct peak in the sorted list were taken into account. For the cases in which the optimum weight determination was ambiguous based on the two criteria employed, the weight was chosen such that it would yield the most reasonable combination of a good peak position and the highest possible Z score. Most often, as the weight was shifted to a slightly higher value than the optimum Z-score value the correct solution was listed in a higher position. However, the opposite correlation was not observed (data not shown).

3. Results

3.1. The diffraction data sets and refined structure

Based on common standards such as resolution, merging statistics, completeness and redundancy values, all 23 diffraction data sets used here are of good quality and exhibit a significant anomalous signal, which is manifested in the values of $R_{\rm anom}$ and the ratio of $R_{\rm anom}$ and $R_{\rm p.i.m.}$ (Weiss, 2001) (Table 1). With the exception of apoferritin and thermolysin, no heavy atoms were included in the crystal structure and the anomalous signal thus stems solely from the lighter atoms: cysteine and methionine S atoms and buffer ions such as potassium, calcium, chloride and sulfate ions.

3.2. The search models

The search models for MR were generated by applying random coordinate displacements using the program *MOLEMAN* (Kleywegt, 1996). The coordinate shifts introduced were the maximum shifts, *i.e.* for a displacement of 1.0 Å each atom was given a new position at random but exactly 1.0 Å away from its original position. The evenly distributed and equally large coordinate differences generated in this way between corresponding atoms in the template and target structure are not to be expected in reality. Two sources of errors may be anticipated. Firstly, the extent to which a protein is structurally conserved varies considerably along the peptide chain, as do the positional coordinate differences between any two related proteins. While the hydrophobic protein cores are more likely to be conserved, loop regions may exhibit larger discrepancies. Secondly, since native S

Table 2
Results of the rotation searches in RF_{prot} and RF_{anom} using the refined structures with no coordinate errors applied.

The performance is evaluated by the position of the correct solution in the sorted output RF list and the corresponding Z score.

Protein	Space group	Peak pos. RF_{prot}	Peak pos. RF_{anom}	Z score RF_{prot}	Z score RF_{anom}
Apoferritin	$F432$	1	18	5.04	3.57
Concanavalin A	$I222$	1	9	13.76	3.82
Glucose isomerase	$I222$	1	1	16.31	6.58
hARH3	$P2_12_12_1$	1	1	16.32	6.20
HEL-45	$P4_32_12$	1	33	8.52	3.23
HEL-80	$P4_32_12$	1	1	9.41	4.87
HNL	$C222_1$	1	4141	16.47	2.11
Insulin	$I2_13$	1	1	4.28	5.19
α-Lactalbumin	$P2_12_12$	1	22145	6.56	1.29
LeuB	$P2_12_12_1$	1	2053	14.02	2.28
MogA	$P2_1$	1	18605	10.09	1.72
NBR1 PB1	$P6_322$	1	6143	5.90	1.29
PPE-Ca	$P2_12_12_1$	1	23	18.30	4.05
PPE-Na	$P2_12_12_1$	1	7	19.26	4.09
Proteinase K	$P4_32_12$	1	1	14.08	5.22
RNAse A ($C2$)	$C2$	1	6560	19.46	2.21
RNAse A ($P3_221$)	$P3_221$	1	7	8.20	3.76
Thaumatin	$P4_12_12$	1	1	11.23	5.17
Thermolysin	$P6_122$	1	4239	6.95	1.77
Titin (A168-A169)	$I222$	1	1	12.23	4.76
Trypsin ($P1$)	$P1$	1	38	29.98	4.10
Trypsin ($P2_1$)	$P2_1$	1	287	20.77	3.18
Trypsin ($P3_121$)	$P3_121$	1	1	13.59	4.91
All examples, mean		**1**	**2796**	**13.08**	**3.71**

atoms tend to be located in the hydrophobic protein cores, the anomalous substructure is expected to be somewhat more conserved than the overall protein structure. Nevertheless, in our experiments the artificial anomalous substructures exhibit exactly the same coordinate shifts as the artificial protein models used here and the coordinate 'errors' are exactly the same in the corresponding Patterson functions. As a result, the optimum weight for the anomalous Patterson function is expected to be underestimated. In order to avoid any bias, evaluations of the rotation searches were performed using statistical properties and relative measures based on normalized rotation functions. For the three examples containing more than one molecule in the asymmetric unit the search model consisted of all molecules in one rotation search in order to avoid the problems associated with multi-copy MR.

3.3. RF calculation

After initial tests using different molecular-replacement programs, the program *AMoRe* (Navaza, 2001) was chosen for the following purely practical reasons: (i) it handles different inputs relatively well and produces easy-to-interpret results, which made it suitable for automation, (ii) it is relatively fast and, most importantly, (iii) it is able to calculate RFs based on an anomalous substructure model only, which typically only consists of a few atoms, against the anomalous differences. However, the fact that *AMoRe* was used here is by no means meant to pose a restriction on the generality of the proposed approach. It can safely be assumed that similar results will be obtained with any of the other MR programs available. The *AMoRe* input parameters used were not systematically optimized, since an exhaustive search of the best input-data combination is typically not performed in the early stages of a real molecular-replacement trial. The diffraction data used in this investigation are generally of good quality and extend to high resolution. In order to make the results comparable between the different cases, a high-resolution limit of 2.5 Å was used for all rotation-function calculations, which suggests that the results reported here are also applicable to more typical data with limitations in the diffraction power.

3.4. Map combination

The best results were obtained using a simple weighted summation of the two RF maps for calculation of a combined RF as described in §2, which then was used for further analysis. Since the optimal proportion of RF_{anom} to RF_{prot} was not known *a priori*, the calculations were carried out for all values of w_{anom} between 0 and 1 in 0.1 steps for each data set.

3.5. Comparing RF_{prot}, RF_{anom} and RF_{comb}

In order to assess the benefits of the combined rotation function RF_{comb} over RF_{prot} and RF_{anom}, all three rotation functions were calculated and analysed. In the case of RF_{prot} calculated using the search model with no coordinate displacements applied, the correct solutions could be identified unambiguously in all 23 examples, indicating both good-quality models and data (Table 2). The results obtained from the anomalous differences were much less obvious and the correct solutions were only found in the top position for seven examples. Furthermore, for seven of the 23 examples the correct solution could not be found among the highest 1000 peaks. This demonstrates that the conventional rotation function RF_{prot} is greatly superior in an MR experiment to the anomalous rotation function RF_{anom} alone. As soon as random coordinate errors are applied, the correct solutions move down the list of candidate solutions in both RF_{prot} and RF_{anom} and their Z scores are decreased (Table 3). Obviously, as the coordinate errors become larger the observed effect is also more pronounced (Tables 3*a*, 3*b* and 3*c*). Not unexpectedly, the tendency here is the same for the real and the anomalous rotation searches. The mean Z scores for the correct orientation in RF_{prot} are 13.1 with no displacements applied and 3.1 with 2.0 Å random displacements applied. The corresponding values using RF_{anom} are 3.7 and 2.3, respectively. With displacements of 2.0 Å applied, the identification of the correct rotation is in most cases not straightforward. The mean rank of the correct solution is 1337 in RF_{prot} and the correct rotation is only found at the top rank in RF_{prot} for one example. In only one other case is the correct solution among the top ten RF_{prot} solutions. However, the chance of finding a solution using the same coordinate errors increases when RF_{comb} instead of RF_{prot} is used. For 15 of the 23 examples the rank of the correct solution has improved, in some individual cases dramatically (Table 3*c*). Despite the seemingly much lower efficiency of RF_{anom} in filtering out the correct solution compared with RF_{prot}, it does provide a significant improvement to RF_{comb}. This improvement is accompanied by the

Table 3
Peak position in the rotation-function (RF) peak list and standard score (Z-score) values for the conventional and combined rotation functions (denoted RF_{prot} and RF_{comb}, respectively).

The fraction of the anomalous rotation function RF_{anom} in the calculation of RF_{comb} is given by the weight w_{anom}.

(a) Random coordinate displacement = 1.0 Å. The weight w_{anom} given is that resulting in the highest value for Z score(RF_{comb})/Z score(RF_{prot}).

Protein	Space group	Peak pos. RF_{prot}	Peak pos. RF_{comb}	Z score RF_{prot}	Z score RF_{comb}	w_{anom}
Apoferritin	$F432$	4	1	3.44	3.66	0.3
Concanavalin A	$I222$	1	1	10.51	10.51	0
Glucose isomerase	$I222$	1	1	9.47	9.57	0.3
hARH3	$P2_12_12_1$	1	1	9.61	10.04	0.3
HEL-45	$P4_32_12$	1	1	5.18	5.49	0.2
HEL-80	$P4_32_12$	1	1	4.32	5.76	0.6
HNL	$C222_1$	1	1	8.67	8.56	0.1
Insulin	$I2_13$	200	1	3.49	3.75	0.6
α-Lactalbumin	$P2_12_12$	20	19	4.37	4.46	0.3
LeuB	$P2_12_12_1$	1	1	8.67	8.67	0
MogA	$P2_1$	1	1	6.87	6.77	0.3
NBR1 PB1	$P6_322$	313	304	2.59	2.64	0.2
PPE-Ca	$P2_12_12_1$	1	1	10.50	10.51	0.1
PPE-Na	$P2_12_12_1$	1	1	10.07	10.15	0.2
Proteinase K	$P4_32_12$	1	1	7.34	7.86	0.3
RNAse A ($C2$)	$C2$	1	1	10.16	10.16	0
RNAse A ($P3_221$)	$P3_221$	692	171	2.77	2.95	0.4
Thaumatin	$P4_12_12$	1	1	7.00	7.66	0.4
Thermolysin	$P6_122$	1	1	4.32	4.27	0.1
Titin (A168-A169)	$I222$	1	1	7.94	7.97	0.1
Trypsin ($P1$)	$P1$	1	1	14.76	14.78	0.2
Trypsin ($P2_1$)	$P2_1$	1	1	11.28	11.23	0.2
Trypsin ($P3_121$)	$P3_121$	1	1	6.63	7.11	0.4
All examples, mean		**54**	**22**	**7.39**	**7.59**	**0.24**

(b) As in (a), but with random coordinate displacement = 1.5 Å.

Protein	Space group	Peak pos. RF_{prot}	Peak pos. RF_{comb}	Z score RF_{prot}	Z score RF_{comb}	w_{anom}
Apoferritin	$F432$	1097	128	2.35	2.83	1.0
Concanavalin A	$I222$	1	1	6.57	6.57	0.0
Glucose isomerase	$I222$	1	1	5.32	5.52	0.3
hARH3	$P2_12_12_1$	5	1	4.31	5.01	0.4
HEL-45	$P4_32_12$	480	1	2.58	3.64	0.6
HEL-80	$P4_32_12$	67	2	2.91	3.53	0.5
HNL	$C222_1$	1	1	4.35	4.35	0
Insulin	$I2_13$	302	138	3.13	3.40	0.3
α-Lactalbumin	$P2_12_12$	1543	1492	2.38	2.40	0.2
LeuB	$P2_12_12_1$	1	1	5.47	5.47	0
MogA	$P2_1$	1	1	4.64	4.64	0.3
NBR1 PB1	$P6_322$	1234	55	2.10	3.34	0.7
PPE-Ca	$P2_12_12_1$	1	1	4.96	5.32	0.3
PPE-Na	$P2_12_12_1$	1	1	4.27	4.29	0.1
Proteinase K	$P4_32_12$	68	64	3.30	3.45	0.2
RNAse A ($C2$)	$C2$	1	1	5.10	5.27	0.2
RNAse A ($P3_221$)	$P3_221$	231	214	3.02	3.05	0.1
Thaumatin	$P4_12_12$	1	1	4.47	4.59	0.3
Thermolysin	$P6_122$	74	46	3.17	3.05	0.2
Titin (A168-A169)	$I222$	1	1	5.44	5.44	0
Trypsin ($P1$)	$P1$	1	1	6.08	6.23	0.3
Trypsin ($P2_1$)	$P2_1$	1	1	6.31	6.31	0
Trypsin ($P3_121$)	$P3_121$	92	5	3.36	4.05	0.4
All examples, mean		**226**	**94**	**4.16**	**4.42**	**0.28**

observed increased Z scores (Table 3). The mean position for the correct solutions for all 23 cases is 411 using RF_{comb}, which corresponds to an increase of the Z score from 3.07 for RF_{prot} to 3.38. These examples clearly demonstrate that although the anomalous data are substantially weaker than the real

Table 3 (continued)

(c) As in (a), but with random coordinate displacement = 2.0 Å.

Protein	Space group	Peak pos. RF_{prot}	Peak pos. RF_{comb}	Z score RF_{prot}	Z score RF_{comb}	w_{anom}
Apoferritin	$F432$	1518	2	2.19	2.49	1.0
Concanavalin A	$I222$	1	1	5.28	5.28	0.0
Glucose isomerase	$I222$	80	22	3.38	3.80	0.5
hARH3	$P2_12_12_1$	992	415	2.70	2.96	0.4
HEL-45	$P4_32_12$	1011	72	2.42	3.46	0.5
HEL-80	$P4_32_12$	219	24	2.69	3.16	0.4
HNL	$C222_1$	834	528	2.66	2.74	0.4
Insulin	$I2_13$	548	548	2.92	2.92	0
α-Lactalbumin	$P2_12_12$	16863	1253	1.41	2.63	0.9
LeuB	$P2_12_12_1$	4	4	4.25	4.25	0
MogA	$P2_1$	73	71	3.54	3.53	0.1
NBR1 PB1	$P6_322$	651	20	2.33	3.33	0.6
PPE-Ca	$P2_12_12_1$	134	64	3.27	3.42	0.2
PPE-Na	$P2_12_12_1$	29	28	3.70	3.72	0.1
Proteinase K	$P4_32_12$	81	81	3.24	3.24	0
RNAse A ($C2$)	$C2$	164	123	3.51	3.92	0.6
RNAse A ($P3_221$)	$P3_221$	36	36	3.67	3.67	0
Thaumatin	$P4_12_12$	35	35	3.41	3.41	0
Thermolysin	$P6_122$	156	14	2.92	3.21	0.6
Titin (A168-A169)	$I222$	910	250	2.63	3.06	0.8
Trypsin ($P1$)	$P1$	5576	5576	2.48	2.48	0
Trypsin ($P2_1$)	$P2_1$	249	249	3.41	3.41	0
Trypsin ($P3_121$)	$P3_121$	583	27	2.70	3.72	0.4
All examples, mean		**1337**	**411**	**3.07**	**3.38**	**0.33**

amplitudes F_P and although they are more susceptible to several sources of errors, their inclusion in the molecular replacement can make a very important contribution. Furthermore, the nature of anomalous data is close to being independent of the real amplitudes and therefore, as will be shown, the use of anomalous data in a rotation search is, for some examples, of pivotal importance to successfully finding the crystal rotation where the search model differs from the target structure. Three examples presented here demonstrate the usage of the combined RF to facilitate a rotation search, as well as of the decisions involved in the process.

3.6. Examples

3.6.1. HEL-45. The model of HEL-45 includes 129 amino acids in one molecule constituting the asymmetric unit. The original data, extending to 1.84 Å resolution, are of very good quality, with an $R_{p.i.m.}$ of 1.0% and an R_{anom} of 2.0%. The R_{anom}-to-$R_{p.i.m.}$ ratio thus indicates a significant anomalous signal in the data (Table 1). Using data between 99 and 2.5 Å resolution, the standard rotation search using the structure-factor amplitudes (RF_{prot}) and the refined structure as the search model show a clear peak for the correct rotation (Z score = 8.52; Table 2). Using anomalous data only, the correct solution is ranked at position 33 with a Z score of 3.23. The application of random coordinate shifts lowers the chances of finding the correct solution. While a displacement of 1.0 Å in all positions still allows the correct solution to be found at position 1 in RF_{prot} (Table 3), with 1.5 Å displacements the correct solution is found at position 480 and with 2.0 Å shifts it is found at position 1011. When the real and the anomalous rotation functions are combined in RF_{comb} using the empirically determined best weight for the combination, the

situation improves in all cases. Interestingly, the improvements are larger for the larger coordinate shifts. With the 1.0 Å coordinate displacements, the Z score is slightly higher for the correct solution in RF_{comb} than in RF_{prot} (5.49 compared with 5.18). When the coordinate displacements are 1.5 and 2.0 Å, the Z scores in RF_{comb} are increased by more than 40% compared with those in RF_{prot}. In practice, this means that at an artificial error level of 1.5 Å the combined rotation function still returns the correct solution in the top position, which presents a dramatic improvement compared with using the standard rotation function. With the largest coordinate displacements of 2.0 Å the rank of the solution improves from position 1011 to position 72, which still demonstrates the beneficial effect of including the anomalous data. As the purpose of this study is to assess whether a combination of the real and the anomalous RFs can give the crystallographer a tool to assist in molecular-replacement structure determination, this example is encouraging: not only does the increase in the Z score using the combined RF instead of the conventional RF suggest an increased probability of finding the correct rotation, the use of RF_{comb} also elevates the correct peak in the 1.5 Å displacements case to the top rank from a much lower position.

3.6.2. Trypsin ($P3_121$). The model of trypsin includes 224 amino-acid residues and 18 anomalously scattering atoms in the asymmetric unit. With R_{anom} and $R_{p.i.m.}$ values of 1.9% and 1.3%, respectively, their ratio of approximately 1.5 also indicates a significant anomalous signal in the data (Table 1). In both RF_{prot} and RF_{anom} the correct solution appears at the top position, with Z scores of 13.59 and 4.91, respectively (Table 2). Calculating RF_{prot} after applying random atomic displacements of 1.0 Å does not change the correct peak position (Table 3). The RF combination with the optimum anomalous weight w_{anom} of 0.4 only increases the Z score by approximately 7%. However, using the model in which 1.5 Å coordinate displacements have been applied the correct solution in RF_{prot} appears at position 92 with a Z score of 3.4 (Table 3*b*). In this situation, it would be very difficult and time-consuming to identify this peak as the correct solution. In the combined rotation function, however, using the optimal optimum anomalous weight of 0.4, the correct position appears at position 5 with a 21% increase in Z score. Any follow-up technique such as translation function (TF), rigid-body refinement, atomic positional refinement and manual inspection of the electron-density maps phased from the potential solution would very likely identify this peak as the correct one. When 2.0 Å coordinate displacements are applied, the ranking of the correct solution is dramatically improved from position 583 to 27 using the combined RF compared with the standard RF. This solution would most likely not be successfully identified, but the 38% increase in Z score is significant and lends further support to the idea of using the RF_{comb} as a standard protocol in MR.

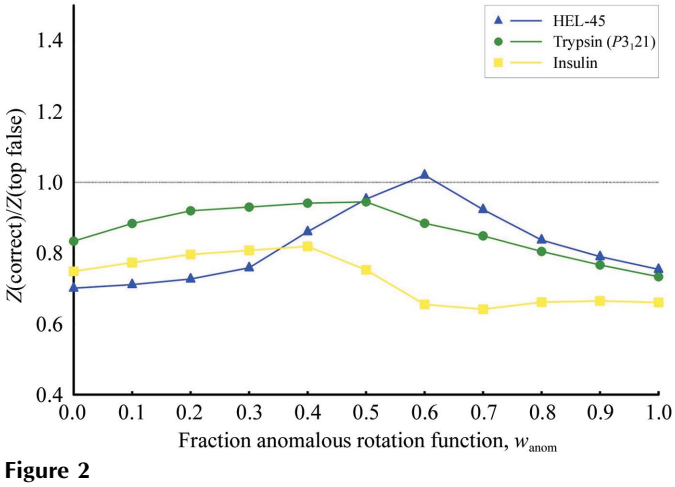

Figure 2
Effect of the weight w_{anom} on the combined rotation function RF_{comb}. The ratio of the Z score of the correct peak to the Z score of the highest-scoring false peak is plotted as a function of the relative contribution (defined by the weight w_{anom}) of the anomalous RF_{anom} in the combined RF_{comb}. A ratio of ≥ 1 indicates that the correct peak was found in the top position in the sorted list of RF peaks. Shown are the results for the three examples discussed in §3.6: HEL-45, trypsin ($P3_121$) and insulin. The respective coordinate sets with 1.5 Å random coordinate displacements were used as search models.

3.6.3. Insulin. The insulin structure contains 51 amino-acid residues and was refined to 1.80 Å resolution. The anomalous substructure consists of six atoms, which results in one of the highest R_{anom}-to-$R_{p.i.m.}$ ratios of all 23 examples of 2.0. Finding the correct rotation using the refined structure is easy (Table 2) in both the conventional and anomalous rotation search. With 1.0 Å coordinate displacements the correct orientation is ranked at position 200 using RF_{prot} only and is at position 548 when 2.0 Å shifts are applied. Use of RF_{comb}, on the other hand, still finds the correct solution at the top position at 1.0 Å displacement, which again clearly demonstrates the usefulness of RF_{comb}. With larger displacements the solution is not found in the first 100 candidates. This would probably not allow the correct solution to be found in either case, although the positive influence of RF_{anom} in RF_{comb} is still clearly discernible.

3.7. The optimum anomalous weight

For any automatic implementation of the linear combination of real and anomalous RF maps, the most appropriate weight for the anomalous RF maps (w_{anom}) must be chosen. This weight influences how large the improvement will be upon combination of the two RF maps. The optimum anomalous weights for the examples presented here have been determined by systematically varying w_{anom} from 0 to 1.0 in steps of 0.1. For the three examples described in §3.6 the effect of w_{anom} on RF_{comb} is illustrated in Fig. 2. The figure clearly shows that a certain contribution of RF_{anom} increases the usefulness of RF_{comb} in the sense that either the contrast between the correct and the highest incorrect peak increases or the correct peak moves closer to the top of the peak list. However, from this figure and the numbers presented in Table 3 it is also evident that the optimum value of w_{anom} varies between different cases. In order to find a way to predict w_{anom} from properties that could be known prior to an MR trial, the dependencies between w_{anom} and several

data-quality-indicating quantities that possibly affect the optimum w_{anom} were investigated. These property measures included R_{anom}, $R_{p.i.m.}$, data redundancy, $I/\sigma(I)$, the number of unique reflections, the size of the protein (or the number of non-H protein atoms), the number of S atoms in the protein and the total number of anomalously scattering atoms per protein molecule, as well as several combinations of these.

The largest correlation found was that between w_{anom} and the ratio $R_{anom}/R_{p.i.m.}$: CC(w_{anom}, $R_{anom}/R_{p.i.m.}$) = 0.36. The weight w_{anom} also correlates with $I/\sigma(I)$ with a correlation coefficient CC[w_{anom}, $I/\sigma(I)$] of 0.26. In principle, these quantities could be used to calculate a w_{anom} directly from the derived statistics of the collected data, but one has to keep in mind that $I/\sigma(I)$ and the $R_{anom}/R_{p.i.m.}$ ratio are not independent of each other. The correlation coefficient for these two entities, CC[$I/\sigma(I)$, $R_{anom}/R_{p.i.m.}$], for the data presented here is 0.76. An additional observation is that the mean optimum weight increases as the applied positional displacements increase. The mean optimum weight is 0.24 using 1.0 Å displacements and increases to 0.28 and 0.33 for 1.5 and 2.0 Å displacements, respectively. With a mean optimum weight of 0.15 as determined by the Z score when no coordinate shifts are applied, the weight is very well modelled as a linear dependency on the applied coordinate displacements. These correlations suggested an empirically determined linear expression for the estimation of w_{anom},

$$w_{anom}(\text{est}) \simeq 0.1R_{anom}/R_{p.i.m.} + 0.1E, \qquad (3)$$

where E is the model error, which is the applied r.m.s. error for the test case. For all the examples discussed in this study, the correlation of the observed w_{anom} and the estimated w_{anom} is 0.37. While $R_{anom}/R_{p.i.m}$ can be derived directly from the data-processing statistics, the equivalent of E in a real-case scenario would be the structural similarity in terms of coordinate r.m.s.d. between the two structures. Clearly, this cannot be known *a priori*; however, it has been known for a long time that the coordinate difference between any two related structures is linked to the sequence similarity between them (Chothia & Lesk, 1986) and has been analysed into detail of structural elements (Williams & Lovell, 2009). The $w_{anom, est.}$ can thus be estimated by taking into account both data-quality measures and similarity between the protein model and the target structure, both of which are typically known before a molecular-replacement attempt. The absolute value of w_{anom} may also be dependent on other factors not explored in this paper, such as for instance the overall strength of the anomalous signal in cases when stronger anomalous scatterers, *e.g.* selenomethionine or heavy-atom derivatives, are present. Nevertheless, the results presented here suggest a way to predict w_{anom} from data-quality indicators and model features that, importantly, can be known prior to an MR trial. The predicted value can then serve as a useful initial value for w_{anom} despite the large variability seen in w_{anom} within this study.

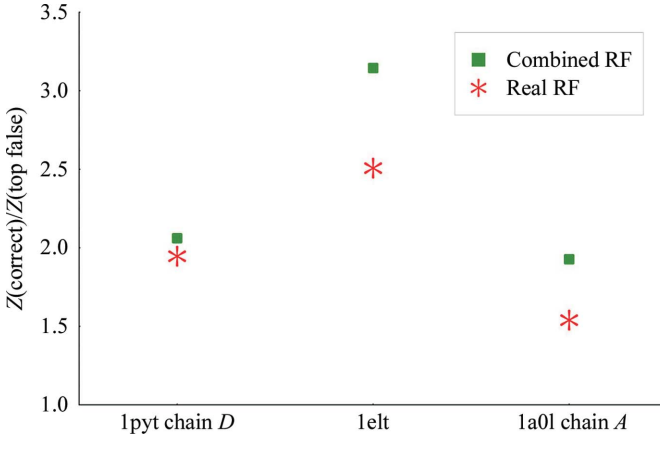

Figure 3
Improvement of RF_{comb} over RF_{prot}. The scores of the correct solutions in the sorted RF peak list are plotted for the example PPE-Ca as discussed in §3.8. The related structures salmon pancreatic elastase (PDB entry 1elt), bovine chymotrypsinogen C (PDB entry 1pyt, chain *D*) and human β-tryptase (PDB entry 1a0l, chain *A*) were used as search models. The sequence identities for these structures are 67, 54 and 38% respectively.

3.8. Simulating a real case using PPE-Ca

To verify these results in a more realistic context, structures related to PPE-Ca were identified using a *BLAST* search against the PDB. Three examples, which represent different degrees of sequence identity, were selected: salmon pancreatic elastase (67% sequence identity for 237 aligned residues), bovine chymotrypsinogen C (54% sequence identity for 235 aligned residues) and human β-tryptase (38% sequence identity or 250 aligned residues). Only salmon pancreatic elastase shares all eight cysteine and two methionine residues with PPE-Ca; the other two contain one Met residue fewer. The results, shown in Fig. 3, verify that the combined RF_{comb} is indeed superior to the conventional RF_{prot}. The optimum value of w_{anom} is between 0.3 and 0.4 for the three cases, which means that it varies much less than when all test cases are considered. It should be noted that the structure of PPE-Ca contains a Ca atom with a peak almost twice as high as the highest sulfur peak, which does not have any correspondence in the related structures (the Ca atom in salmon pancreatic elastase was not taken into account in this example). It therefore seems highly probable that the combined rotation function is perfectly valid also in cases where the anomalous substructures are not identical.

4. Discussion

4.1. Evaluation of the proposed approach

The proposed and described method is aimed at either identifying the correct solution as the top-ranked candidate with a higher degree of confidence compared with the conventional approach or minimally at ranking the correct solution into a sufficiently small subset such that it can be identified rapidly by a second filtering step. The second option is justified by the availability of more time-consuming methods that are used to analyse the correctness of a candidate MR solution. These mostly assume a full MR solution (including

the TF) and include subsequent rigid-body and atom-positional refinement as well as the analysis of the packing function together with the protein rigidity in these areas. In our tests, the combination of the conventional and anomalous RFs scores the correct peak higher in almost all cases (Table 3). Assuming that a set of ten potential rotation candidates would still be feasible for further analysis, the number of structures which can be determined by MR would certainly increase. It should be added, however, that in some of the other examples the method still does not raise the RF solution sufficiently to guarantee a full MR solution.

In this study, artificially disturbed models have been used to calculate the rotation function by introducing distributed positional shifts into the coordinates of refined structures. It may be argued that these models are not realistic. This may affect the absolute values of the rotation functions, as well as the derived optimum weights. However, the tendencies revealed here will not be greatly influenced. The tendencies are furthermore evaluated using statistical Z scores, whereby the statistical properties of the rotation functions can be neglected as long as the different distributions are similar. The same is true for the analysis of the optimum weight (see below). Only the absolute value of the weight may be less well predicted and somewhat underestimated in these examples. A few examples were chosen to verify this hypothesis. The plots of the rank of the correctly found solutions as a function of the anomalous weight using artificial errors were compared with the corresponding plots using related structures from the PDB. As a clear tendency could not be seen, the effect on the optimum weight of nonrealistic coordinate errors appears to be small.

The 23 examples included in this study have deliberately been chosen so that none of them contained atoms traditionally associated with anomalous dispersion experiments such as selenomethionine or typical heavy-atom compounds. Only one example (apoferritin) contains cadmium ions, which produce a somewhat stronger anomalous signal than those observed for the examples containing light atoms only.

With the main anomalous contribution arising from light atoms only, it is clear that the anomalous differences in the observed structure factors are substantially smaller than the observed structure factors. In spite of this, the anomalous rotation function constitutes an important contribution in several examples. If examples were used containing atoms exhibiting stronger anomalous dispersion, the anomalous signal-to-noise ratio would obviously increase, with an expected increase in the anomalous contribution of the combined rotation function. This study shows that the anomalous contribution for any crystal with well diffracting properties containing native S atoms or ions from the crystallization cocktails has the potential to be of pivotal value in a molecular-replacement approach.

4.2. Applicability

The experience from this study, considering the quality of the anomalous data used in these examples, is that the conventional rotation function is superior to the anomalous rotation function but that a combined rotation function is superior to both. Studying the properties of the optimum weight in the linear combination gives us some interesting clues as to how the combined map should be interpreted. As the data show, the optimum anomalous weight varies in the 23 examples presented. The reason for these variations may include geometrical considerations, as the shape of the protein and the anomalous models probably do influence the success rate of sorting out the correct rotation. Also, the quality of the data, especially the anomalous data, is expected to be somewhat correlated with the behaviour of the rotation search. Although a high signal-to-noise ratio for the anomalous data would be expected to yield a more reliable anomalous RF_{anom}, and thus a higher optimum anomalous weight, this correlation could only be partly discerned in our tests (see §3). The reason for this is not clear. It is possible that the anomalous data quality needs to be substantially improved before it can have a clearly detectable effect on the RF_{anom}. As experimental techniques for the collection of anomalous data are constantly improving, a future study could address this issue. In cases where there is a complete model that resembles the target structure, structure determination by MR is in most cases straightforward. As the agreement between the target structure and the model decreases, both the conventional and the anomalous RFs become more difficult to interpret. It also must be expected that the RF_{anom} will rapidly become more difficult to interpret because the anomalous differences are measured much less accurately. Fortunately, RF_{prot} and RF_{anom} may be regarded as containing independent information. The anomalous data are extracted from $|F_P|$ differences, and the anomalously scattering model is usually only a small subset of the complete model. As a consequence, the combination of the two, RF_{comb}, may still work in spite of the relatively lower chance of finding the correct orientation in the RF_{anom} alone in some cases. Other parameters that one would intuitively anticipate to influence the optimal ratio between RF_{prot} and RF_{anom} are the number of anomalous scatterers, as well as their anomalous scattering power, and the resolution, both of which potentially increase the reliability of the anomalous data. Tests using several resolution limits showed that the resolution is indeed an important variable, with an increased resolution generally yielding better performances of the combined rotation function (data not shown). A rather weak correlation could also be found between w_{anom} and the number of anomalous scatterers. We may thus conclude from the analyzed examples that the combined map is more useful when the initial model is far from the correct one. A possible interpretation of this observation is that as long as the model is good the better data ($|F_P|$) should be used. When the model is poorer, additional ($|\Delta_{anom}|$) data may be needed. This obviously gives another clue of how to estimate the optimal anomalous weight: the worse the model, the more weight should be given to RF_{anom}, up to some empirically determined value. The experiments presented here lend some support to this strategy, as the mean optimum anomalous weight using 2.0 Å random-error models (0.33) is larger than

the weight using 1.0 Å error models (0.24). Although $w_{\rm anom}$ is somewhat correlated with the model and data quality, these relations are apparently not strong enough to allow a proper estimate of $w_{\rm anom}$, as can be seen from the large variability of $w_{\rm anom}$. If a better method of estimating $w_{\rm anom}$ *a priori* could be found, the concomitant use of the two RFs would benefit substantially.

5. Summary and conclusions

While anomalous data have been used extensively in *de novo* macromolecular structure determinations in SAD-, MAD-, SIRAS- or MIRAS-type approaches, their use in MR has so far been limited. However, anomalous data present information orthogonal to the reflection intensities. In this study, we demonstrate that in selected cases a structure-determination strategy using MR is facilitated by the inclusion of anomalous differences. With the increased experience in the collection of anomalous data, as well as the ongoing improvements to detectors and other parts of the experimental setup, the measurement of small anomalous differences will contribute important complementary information in an increasing number of projects. We conclude that anomalous data may be of help in MR projects where the evolutionary distance between the target structure and the model is large.

We would like to acknowledge the support of this work by the EC 6th Framework Programme 'Life Sciences, Genomics and Biotechnology for Health' (Integrated Research Project BIOXHIT, Contract No. LHSG-CT-2003-503420) and by the Deutsche Forschungsgemeinschaft (DFG grant WE2520/2 to MSW). We would also like to thank Olle Terenius (Swedish University of Agricultural Sciences, Uppsala, Sweden) for useful discussions.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.
Anderson, D. H., Weiss, M. S. & Eisenberg, D. (1996). *Acta Cryst.* D**52**, 469–480.
Baker, E. N., Anderson, B. F., Dobbs, A. J. & Dodson, E. J. (1995). *Acta Cryst.* D**51**, 282–289.
Berglund, G. I., Willassen, N. P., Hordvik, A. & Smalås, A. O. (1995). *Acta Cryst.* D**51**, 925–937.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Blow, D. M. (1958). *Proc. R. Soc. Lond. A*, **247**, 302–336.
Bricogne, G. (1992). *Proceedings of the CCP4 Study Weekend. Molecular Replacement*, edited by W. Wolf, E. J. Dodson & S. Gover, pp. 62–75. Warrington: Daresbury Laboratory.
Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.
Chothia, C. (1992). *Nature (London)*, **357**, 543–544.
Chothia, C. & Lesk, A. M. (1986). *EMBO J.* **5**, 823–826.
Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.
DeLano, W. L. & Brünger, A. T. (1995). *Acta Cryst.* D**51**, 740–748.
Djinović Carugo, K., Helliwell, J. R., Stuhrmann, H. & Weiss, M. S. (2005). *J. Synchrotron Rad.* **12**, 410–419.
Einspahr, H., Suguna, K., Suddath, F. L., Ellis, G., Helliwell, J. R. & Papiz, M. Z. (1985). *Acta Cryst.* B**41**, 336–341.
Ferreira, K. N., Iverson, T. M., Maghlaoui, K., Barber, J. & Iwata, S. (2004). *Science*, **303**, 1831–1838.
Gomis-Rüth, F. X., Gómez, M., Bode, W., Huber, R. & Avilés, F. X. (1995). *EMBO J.* **14**, 4387–4394.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
Huber, R. (1965). *Acta Cryst.* **19**, 353–356.
Keegan, R. M. & Winn, M. D. (2007). *Acta Cryst.* D**63**, 447–457.
Kleywegt, G. J. (1996). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **32**, 32–36.
Kuettner, E. B., Hilgenfeld, R. & Weiss, M. S. (2002). *J. Biol. Chem.* **277**, 46402–46407.
Kwiatkowski, W., Noel, J. P. & Choe, S. (2000). *J. Appl. Cryst.* **33**, 876–881.
Liu, X., Fan, K. & Wang, W. (2004). *Proteins*, **54**, 491–499.
Liu, Z.-J., Vysotski, E. S., Chen, C.-J., Rose, J. P., Lee, J. & Wang, B.-C. (2000). *Protein Sci.* **9**, 2085–2093.
Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *Acta Cryst.* D**64**, 125–132.
Madauss, K., Juzumiene, D., Waitt, G., Williams, J. & Williams, S. (2004). *Endocr. Res.* **30**, 775–785.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
Mueller-Dieckmann, C., Panjikar, S., Schmidt, A., Mueller, S., Kuper, J., Geerlof, A., Wilmanns, M., Singh, R. K., Tucker, P. A. & Weiss, M. S. (2007). *Acta Cryst.* D**63**, 366–380.
Mueller-Dieckmann, C., Panjikar, S., Tucker, P. A. & Weiss, M. S. (2005). *Acta Cryst.* D**61**, 1263–1272.
Mueller-Dieckmann, C., Polentarutti, M., Djinovic Carugo, K., Panjikar, S., Tucker, P. A. & Weiss, M. S. (2004). *Acta Cryst.* D**60**, 28–38.
Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.
Nan, J., Zhou, Y., Yang, C., Brostromer, E., Kristensen, O. & Su, X.-D. (2009). *Acta Cryst.* D**65**, 440–448.
Navaza, J. (1987). *Acta Cryst.* A**43**, 645–653.
Navaza, J. (2001). *Acta Cryst.* D**57**, 1367–1372.
Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst.* D**61**, 449–457.
Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2009). *Acta Cryst.* D**65**, 1089–1097.
Pereira, P. J., Bergner, A., Macedo-Ribeiro, S., Huber, R., Matschiner, G., Fritz, H., Sommerhoff, C. P. & Bode, W. (1998). *Nature (London)*, **392**, 306–311.
Raaf, J., Issinger, O. G. & Niefind, K. (2008). *Mol. Cell. Biochem.* **316**, 15–23.
Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.
Schuermann, J. P. & Tanner, J. J. (2003). *Acta Cryst.* D**59**, 1731–1736.
Sekar, K., Rajakannan, V., Velmurugan, D., Yamane, T., Thirumurugan, R., Dauter, M. & Dauter, Z. (2004). *Acta Cryst.* D**60**, 1586–1590.
Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.
Weiss, M. S., Panjikar, S., Nowak, E. & Tucker, P. A. (2002). *Acta Cryst.* D**58**, 1407–1412.
Weiss, M. S., Sicker, T., Djinovic-Carugo, K. & Hilgenfeld, R. (2001). *Acta Cryst.* D**57**, 689–695.
Weiss, M. S., Sicker, T. & Hilgenfeld, R. (2001). *Structure*, **9**, 771–777.
Williams, S. G. & Lovell, S. C. (2009). *Mol. Biol. Evol.* **26**, 1055–1065.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Yang, C., Pflugrath, J. W., Courville, D. A., Stence, C. N. & Ferrara, J. D. (2003). *Acta Cryst.* D**59**, 1943–1957.