Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

A multivariate likelihood SIRAS function for phasing and model refinement

Pavol Skubák,[a]* Garib Murshudov[b] and Navraj S. Pannu[a]*

[a]Biophysical Structural Chemistry, Leiden Institute of Chemistry, Gorlaeus Laboratories, Leiden University, PO Box 9502, 2300 RA Leiden, The Netherlands, and [b]York Structural Biology Laboratory, Chemistry Department, University of York, Heslington, York, England

Correspondence e-mail: p.skubak@chem.leidenuniv.nl, raj@chem.leidenuniv.nl

Received 18 February 2009
Accepted 16 July 2009

A likelihood function based on the multivariate probability distribution of all observed structure-factor amplitudes from a single isomorphous replacement with anomalous scattering experiment has been derived and implemented for use in substructure refinement and phasing as well as macromolecular model refinement. Efficient calculation of a multidimensional integration required for function evaluation has been achieved by approximations based on the function's properties. The use of the function in both phasing and protein model building with iterative refinement was essential for successful automated model building in the test cases presented.

1. Introduction

Despite the dramatic increase in the number of macromolecular structures in the Protein Data Bank (PDB; Berman *et al.*, 2000) phased by molecular replacement, about 20% of recently determined crystal structures have still been solved by experimental phasing methods (Long *et al.*, 2008) phased by molecular replacement. Experimental phase information can also serve as an additional source of information in model building and refinement, especially at lower resolutions when the observation-to-parameter ratio is very low (DeLaBarre & Brunger, 2006). Recent developments in heavy-atom soaking (Boggon & Shapiro, 2000; Wernimont *et al.*, 2000; Dauter *et al.*, 2000, 2001; Szczepanowski *et al.*, 2005; Beck *et al.*, 2008) have increased the success rate and extended the application of both single-wavelength anomalous diffraction (SAD) and single isomorphous replacement with anomalous scattering (SIRAS). These techniques are often applied to large molecular complexes or flexible molecules, which typically provide fragile crystals and weak diffraction data with a small signal-to-noise ratio. Improved statistical techniques that exploit all information simultaneously are needed to optimally extract information from such data. Furthermore, the enhanced exploitation of SIRAS data from a native and a soaked crystal may lead to solutions which elude SAD data collected from a soaked crystal alone.

The SIRAS experiment involves data collected from a native crystal and Friedel mates from a derivative crystal containing a 'heavy atom' [for a review of SIRAS and experimental phasing, see Taylor (2003) and references therein],

$$\begin{aligned} |F_1| &= |F_{\rm o}^{N}| \\ |F_2| &= |F_{\rm o}^{D+}| \\ |F_3| &= |F_{\rm o}^{D-}|, \end{aligned} \qquad (1)$$

with the corresponding model structure factors

$$\mathbf{F}_4 = \mathbf{F}_c^N$$
$$\mathbf{F}_5 = \mathbf{F}_c^{D+}$$
$$\mathbf{F}_6 = (\mathbf{F}_c^{D-})^* \quad (2)$$

[for simplicity, both the D (derivative) superscript and the complex-conjugate sign will be omitted]. Currently, the best approach for SIRAS substructure phasing and refinement neglects the correlation between the isomorphic and anomalous sources of information. Indeed, the (univariate) likelihood-based SIRAS function used in *SHARP* (de La Fortelle & Bricogne, 1997) and *BP*3 (Pannu *et al.*, 2003) assumes the independence of a Gaussian term involving anomalous differences (North, 1965; Matthews, 1966) and a Rice function modelling the isomorphic data for acentric reflections,

$$\begin{aligned}
&P(|F_o^N|, |F_o^+|, |F_o^-|; \mathbf{F}_c^+, \mathbf{F}_c^-) \\
&= \int_0^\infty \int_0^{2\pi} |F| P_{\rm prior}(|F|, \alpha) \frac{2|F_o^N|}{V_N} \exp\left(-\frac{|F_o^N|^2 + |F|^2}{V_N}\right) \\
&\quad \times I_0\left(\frac{2|F_o^N||F|}{V_N}\right) \frac{2|F_o^D|}{V_D} \exp\left(-\frac{|F_o^D|^2 + |F_c|^2}{V_D}\right) I_0\left(\frac{2|F_o^D||F_c|}{V_D}\right) \\
&\quad \times \frac{1}{(2\pi V_a)^{1/2}} \exp\left[-\frac{(\Delta_{\rm obs} - \Delta_{\rm calc})^2}{2V_a}\right] {\rm d}\alpha\, {\rm d}|F|, \quad (3)
\end{aligned}$$

where the 'true' native amplitude $|F|$ and phase α are integrated out, $P_{\rm prior}$ describes the prior knowledge about $\mathbf{F}$ (if any), F_o^D is the average of $|F_o^+|$, $|F_o^-|$ and $|F_c|$ is the average of $|F_c^+|$, $|F_c^-|$, the calculated structure factors determined from the 'true' native structure factor and the calculated heavy-atom structure factors. $\Delta_{\rm obs}$ is the Bijvoet difference of the observed Friedel pairs $|F_o^+|$ and $|F_o^-|$, V_N, V_D and V_a are variances and $\Delta_{\rm calc} = |F_c^+| - |F_c^-|$ is the calculated Bijvoet difference.

Previously, we have shown that substructure phasing and refinement using a multivariate likelihood function that directly considers the correlation between Friedel pairs in a SAD experiment provides better results than the same Gaussian-based term using anomalous differences (Pannu & Read, 2004; Ness *et al.*, 2004). Thus, a multivariate function which directly accounts for all correlations between structure factors in a SIRAS experiment should allow the extraction of more information from low signal-to-noise data.

Currently, a function that simultaneously and directly exploits native and derivative data from a SIRAS experiment has not been implemented in macromolecular refinement. The best available approach for considering phase information from a SIRAS experiment in model refinement is with the 'MLHL' target function, a univariate likelihood function which incorporates experimental phase information *via* Hendrickson–Lattman coefficients (Pannu *et al.*, 1998). However, this indirect use of experimental phases suffers from shortcomings such as the assumption of independence of experimental phase information from the model, an inability for simultaneous refinement of (a perhaps updated) substructure and protein models and a dependency on the accuracy and reliability of the phasing program used to generate the Hendrickson–Lattman coefficients (Skubák *et al.*, 2004). A multivariate single anomalous diffraction (SAD) function has been shown to overcome these shortcomings and to extend the resolution and phase-quality limits needed for successful automated model building with iterative refinement against the SAD data set (Skubák *et al.*, 2005). In this paper, a multivariate likelihood function for macromolecular refinement against SIRAS experimental data is presented.

2. Method

The probability distribution of three structure-factor amplitudes for a reflection [as specified by (1)] given $N - 3$ model structure factors is derived in Appendix A. In our current implementation, three models specified by (2) are used, leading to the following probability distribution:

$$\begin{aligned}
&P(|F_1|, |F_2|, |F_3|; |F_4|, \alpha_4, |F_5|, \alpha_5, |F_6|, \alpha_6) \\
&= \frac{2|F_1||F_2||F_3| \det(C_3)}{\pi^2 \det(C_6)} \exp\Bigg[-\sum_{i=1}^{3} |F_i|^2 a_{ii} \\
&\quad - \sum_{i=4}^{6} \Bigg(|F_i|^2 (a_{ii} - c_{ii}) + \sum_{j=i+1}^{6} \{2|F_i||F_j|[(a_{ij} - c_{ij}) \cos(\alpha_j - \alpha_i) \\
&\quad - (b_{ij} - d_{ij}) \sin(\alpha_j - \alpha_i)]\} \Bigg) \Bigg] \\
&\quad \times \int_0^{2\pi} \int_0^{2\pi} \exp\Bigg(-\sum_{j=2}^{3} \sum_{i=j+1}^{6} \{2|F_j||F_i|[a_{ji} \cos(\alpha_i - \alpha_j) \\
&\quad - b_{ji} \sin(\alpha_i - \alpha_j)]\} \Bigg) I_0[2|F_1|\xi(\alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)]\, {\rm d}\alpha_2\, {\rm d}\alpha_3, \quad (4)
\end{aligned}$$

where

$$\begin{aligned}
\xi(\alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = \Bigg(&\sum_{i=2}^{6} \Bigg\{ |F_i|^2 (a_{1i}^2 + b_{1i}^2) \\
&+ \sum_{j=i+1}^{6} 2|F_i||F_j|[(a_{1i}a_{1j} + b_{1i}b_{1j}) \cos(\alpha_j - \alpha_i) \\
&+ (a_{1j}b_{1i} - a_{1i}b_{1j}) \sin(\alpha_j - \alpha_i)] \Bigg\} \Bigg)^{1/2}. \quad (5)
\end{aligned}$$

C_6 is the covariance matrix of the complex Gaussian distribution $P(\mathbf{F}_1, \ldots, \mathbf{F}_6)$, with the real and imaginary components of its inverse denoted as a_{jk} and b_{jk}, respectively. Similarly, C_3 is the covariance matrix of the Gaussian distribution $P(\mathbf{F}_4, \mathbf{F}_5, \mathbf{F}_6)$, with the real and imaginary components of its inverse denoted as c_{ij} and d_{ij}, respectively.

We define the SIRAS function as the sum over all reflections of the minus logarithm of the derived probability distribution (4). Evaluation of the function requires a three-dimensional integration over the unknown observed phases, one of which is solved analytically (Appendix A). Since we were not able to find a usable analytical solution to the remaining two integrals, a two-dimensional numerical integration was used for evaluation of the remaining two integrals. However, a SIRAS function employing an integral evaluated by the Gaussian method, as a two-dimensional extension from an accurate implementation of the one-dimensional SAD numerical integration, required up to 1000×1000 nodes to achieve an acceptable precision and stable refinement. Clearly, more advanced approximations were needed to speed up the

SIRAS function evaluation and to achieve a speed comparable with the currently used functions. The solution adopted is based on analysis of the specific properties of the integral.

In Appendix *C*, we show that the three-dimensional SIRAS integral $I(\alpha_1, \alpha_2, \alpha_3)$ (before analytical integration) is dependent on nine real-valued parameters (denoted as ennead $\boldsymbol{\varepsilon}$): six w parameters (analogous to vector amplitudes) and three φ parameters (analogous to vector angles). If the values of the w parameters are small, the surface of the function $I(\alpha_1, \alpha_2, \alpha_3)$ (for a given $\boldsymbol{\varepsilon}$) is flat and a small number of sampling points over the whole integration range is sufficient for reasonable precision of the numerical integration of *I*. However, higher values of the w parameters generally give rise to a sharp and high peak and very dense integration sampling would be required to sample over the whole integration area. Therefore, the position of the peak of the integrand in three-dimensional space $(\alpha_1, \alpha_2, \alpha_3)$ is important for numerical integration of *I*.

The statement in Appendix *D* provides a partial localization of the position of the maximum of *I*: the maximum is close to a certain plane for the majority of reflections in a typical SIRAS experiment. The statement defines the plane and provides the maximal distance of the maximum from the plane. This information can be used to limit the large number of sampling points needed for the numerical integration for large w parameters. A transformation of the coordinate system α_1, α_2, α_3 can be performed such that one of the new coordinate axes is perpendicular to the plane. Sampling of this variable over a short range covering the peak within the maximal distance given by the statement then provides an efficient method for the numerical integration. The transformation and the complete algorithm for the SIRAS function evaluation are specified in Appendix *E*.

The SIRAS function was implemented according to this algorithm in the refinement program *REFMAC*5 (Murshudov *et al.*, 1997). Validation of the implementation of the function evaluation in terms of precision and actual number of nodes used has been performed on several SIRAS data sets, showing that a relative precision of the order of 10^{-5} is achieved by the use of an average of 100–150 Gaussian integration nodes per reflection.

The SIRAS function has been implemented in *REFMAC*5 (v.5.6) for substructure refinement and phasing and also for protein refinement with the direct use of SIRAS phase information. The performance of the 'multivariate' phasing function has been compared with the currently used univariate function as implemented in the program *BP*3 (v.1.01), denoted below by 'univariate'. The function for protein model refinement has been compared against the 'Rice' likelihood function lacking prior phase information (Bricogne & Irwin, 1996; Murshudov *et al.*, 1997; Pannu & Read, 1996) denoted below as Rice, and the likelihood function encoding prior phase information with Hendrickson–Lattman coefficients, denoted below as MLHL, both implemented in *REFMAC*5 (v.5.6) in the context of automated model building with iterative refinement by *ARP*/*wARP* (v.7.0; Perrakis *et al.*, 1999).

For the three SIRAS test cases described below, the *CRANK* suite (v.1.2.1; Ness *et al.*, 2004) from *CCP*4 (v.6.10; Collaborative Computational Project, Number 4, 1994) was used for automatic structure solution starting with the SIRAS data and the protein sequence. *CRANK* uses the programs *SHELXD* (Sheldrick, 2008) or *CRUNCH*2 (de Graaff *et al.*, 2001) for substructure detection, *SHELXE* (Sheldrick, 2008) for hand determination, *BP*3 or *REFMAC*5 for substructure refinement and phasing, *SOLOMON* (Abrahams & Leslie, 1996) for density modification and *ARP*/*wARP* for automated model building with iterative refinement by *REFMAC*5. *EMMA* from the *CCTBX* toolbox (Grosse-Kunstleve *et al.*, 2002) was used to transform all substructure sites to the same origin as the final published model. This simplified the calculation of map correlations with the final map in *SFTOOLS* (Bart Hazes, unpublished work). Unless otherwise stated, the default *CRANK* parameters were used in all runs.

The Hendrickson–Lattman coefficients required for the MLHL function were derived using the phasing program in a given pipeline (either *BP*3 or *REFMAC*5). MLHL with Hendrickson–Lattman coefficients from the density-modification programs *SOLOMON* and *DM* (Cowtan, 1999) was also tested, but produced poorer results. When the SIRAS function was used for protein refinement, the refined substructures from *BP*3 or *REFMAC*5 were input into *ARP*/*wARP*. For all likelihood functions and test cases, 200 cycles (four times the default) of automated model building with iterative structure refinement were performed to allow convergence of the model-building process. The native data were used for model building and refinement in all the test cases (except for the SIRAS function, which used all observations). The resulting models were compared with the final refined structure by a compare-protein script (S. Ness & P. Skubak, unpublished work) from the *CRANK* suite, which provides the number of 'correctly built' residues. A residue is regarded as correctly built if its C^{α} atom lies within 1 Å of a C^{α} position from the final model (Badger, 2003).

All the data sets used for the tests described below were acquired from the PDB. Since the PDB stores reflection data in mmCIF format, conversion to MTZ format was required. The conversion was performed in several steps: firstly, the downloaded mmCIF file was manually analyzed to separate, by hand, the multiple data sets (native, derivative plus and derivative minus) into multiple mmCIF files. The separated mmCIF files were converted using either the *CIF*2*MTZ* utility from the *CCP*4 suite or *SF-CONVERT* (http://sw-tools.pdb.org/apps/SF-CONVERT) depending on how the anomalous data were represented in the mmCIF file. Finally, the separated MTZ data sets were merged into a single MTZ file using *SFTOOLS* from *CCP*4.

3. Results

3.1. DNA-packaging protein Gp17

The initial phases for bacteriophage T4 gp17 ATPase domain mutant complexed with ATP (PDB code 2o0h; Sun *et al.*, 2007) were determined from a native data set and a selenomethionine derivative containing eight Se atoms collected

Table 1
The effect of the f'' used for phasing on the resulting map correlation after phasing for the 2o0h data set.

	f''							
Map correlation	3	4	6	8	10	12	14	16
Univariate	0.333	0.355	0.379	0.392	0.396	0.397	0.395	0.385
Multivariate	0.405	0.412	0.414	0.413	0.409	0.412	0.414	0.413

at the selenium absorption edge. Although the theoretical value of the anomalous scattering coefficient f'' for an Se atom at its peak is close to 4, a value of 12 was used for the structure determination by the authors (Sun, personal communication). Because of this large discrepancy, we investigated the effects of f'' on the phasing by running a series of phasing jobs starting from the same substructure and varying f'' for both functions. As Table 1 documents, the multivariate function provides almost equally good results in the whole f'' range from 3 to 16, while the univariate-function results deteriorate with an f'' lower than 6. Since the value of 12 suggested by the original authors provided close to optimal results for both functions, it was used in all the pipelines (Tables 2 and 3).

The structure contains a single monomer with 357 residues in the asymmetric unit, a majority of which were correctly built by *CRANK* using the multivariate SIRAS function for both phasing and protein refinement. Only small fragments of the structure were built using any other combination of the functions for phasing and protein refinement (Table 3). Fig. 1 demonstrates the differences in the performance of different refinement target functions in a refinement-only pass from a model built after five *ARP*/*wARP* rebuilding cycles.

Density modification was an essential step in the structure-solution process, probably owing to the phase extension of the phases from the 3.29 Å selenium derivative to the 1.88 Å native data. *CRANK* options for automated optimization of the solvent content and a higher number of *SOLOMON* cycles were used in all runs to enable effective phase extension.

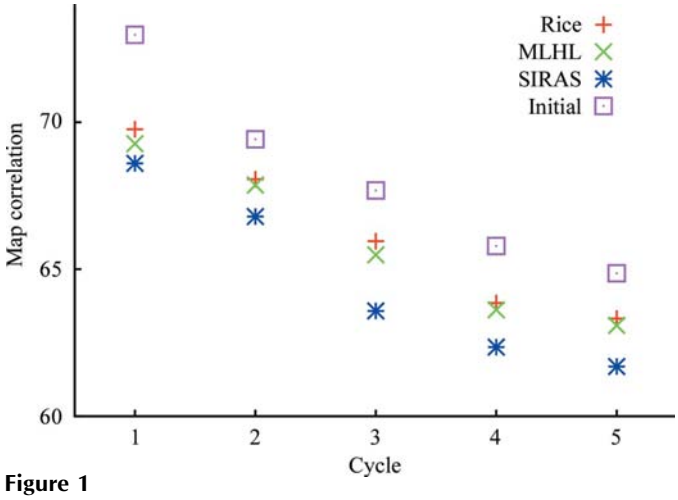

Figure 1
The phase error for the Gp17 structure after a refinement-only pass. A model built in the first five *ARP*/*wARP* rebuilding cycles of a pipeline with multivariate phasing and refinement was inputted for refinement with Rice, MLHL and SIRAS targets.

Table 2
The map correlations after phasing and density modification (DM).

	2o0h		2b78		2b79	
Map correlation after	Phasing	DM	Phasing	DM	Phasing	DM
Univariate	0.397	0.366	0.368	0.488	0.457	0.718
Multivariate	0.412	0.429	0.374	0.535	0.486	0.739

Table 3
The number of residues built using various target functions in phasing and model building.

		Correctly built residues		
Phasing function	Refinement function	2o0h	2b78	2b79
Univariate	Rice	24	51	139
Univariate	MLHL	34	96	197
Univariate	Multivariate SIRAS	53	345	235
Multivariate	Rice	68	106	190
Multivariate	MLHL	128	341	201
Multivariate	Multivariate SIRAS	310	353	238

3.2. SMU.776

A single Hg atom provided sufficient signal to automatically build a majority of the 385 residues of a putative SAM-dependent methyltransferase (SMU.776; PDB code 2b78; J. Nan, K. T. Wang & X.-D. Su, unpublished work) from the experimental phases, helped by the relatively good resolution of both the native (1.8 Å) and derivative (1.94 Å) data. The use of the multivariate function in either phasing or protein model refinement was essential to obtain the almost completely traceable density maps (Fig. 2; Table 3). However, there is little discrimination between the performance of the functions after phasing; a significant difference only appeared after density modification using the different probability distributions from phasing.

The model could not be built without the use of experimental phase information in either an indirect (MLHL function) or direct way (SIRAS function). The indirect use of the phase information was sufficient to build a model of similar quality to that built using the SIRAS function provided that the starting map was obtained using the multivariate function in substructure phasing (Table 3).

3.3. SMU.440

Similarly to SMU.776, the structure of the SMU.440 protein (PDB code 2b79; J. Nan, X. Y. Zhang, X. Y. Liu & X.-D. Su, unpublished work) from *Streptococcus mutans* was determined by the Joint Center of Structural Genomics (JSCG: http://www.jcsg.org). The maps after phasing and density modification were of significantly higher quality than in the previous two test cases and approximately half of the structure was built immediately in the first *ARP*/*wARP* cycle. However, tracing of the remaining residues was more difficult owing to poor electron density in some regions. The use of prior phase information during model building improved the problematic map regions and better models were subsequently built. The derivative data were of slightly better quality than the native data (the former were obtained to approximately 2.38 Å resolution and the latter to 2.35 Å with a similar signal-to-

noise ratio), possibly improving the SIRAS function refinement compared with the native data-based refinement of the Rice and MLHL functions. The $R_{\rm free} - R$ difference during the first model-building cycles (Fig. 3) suggested that a great deal of the improvement could be attributed to decreased overfitting by the direct use of experimental phase information. Although the maps after phasing and density modification differed slightly in their quality depending on the function used in phasing, these differences did not play a significant role in the building of the model.

4. Discussion and conclusions

The previous automated structure-solution results can be used as an additional validation of the implementation of the SIRAS function in *REFMAC*5. More importantly, they show that the use of the function can provide significant improvements over the currently used functions in difficult cases. The improvements of phasing by the multivariate function over phasing by the univariate function lead to significant differences in the Gp17 and SMU.776 models built and a significantly smaller sensitivity of the multivariate function to f'' values has been observed. However, it is not clear how much these results are influenced by other differences in the programs used for the comparison. An implementation of the multivariate and univariate SIRAS function in the same program could provide better discrimination.

The SIRAS function model refinement does not suffer from this problem since all the functions tested have been implemented in the same program. In the three test cases above the use of experimental SIRAS phase information was essential to build the structures, with the direct incorporation of the information in the multivariate SIRAS function providing better results than the indirect MLHL function.

For all three test cases we could not automatically build any of the structures using SAD data from the 'derivative' data sets alone. Thus, collecting 'native' unsoaked data and optimally using this additional information could be the difference between a successful and unsuccessful structure solution.

Model building with the SIRAS function was approximately 1.5–1.7 times slower compared with the Rice target in the tests above, which is satisfactory given the computational complexity differences between the two targets. The speed could be further improved by decreasing the current high precision of the function evaluations. Furthermore, the evaluation of the function for a given set of reflections is an 'embarrassingly parallel' problem; thus the speed of a parallelized SIRAS function evaluation on a currently standard quad-core processor could be close to that of the SAD function evaluation.

The results of the SIRAS function implementation are also promising with respect to the direct use of prior phase infor-

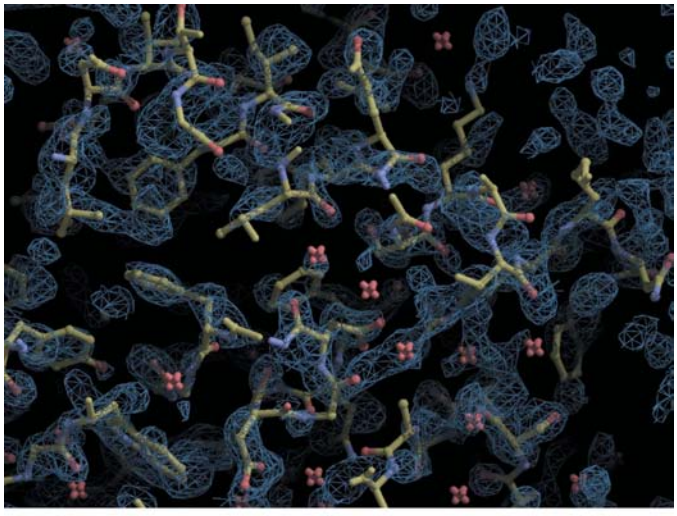

Figure 2
Density maps (contoured at 1.5σ) of an SMU.776 region at the end of *ARP*/*wARP* building superimposed on the final deposited model using (*a*) univariate and (*b*) multivariate targets in phasing and model building.

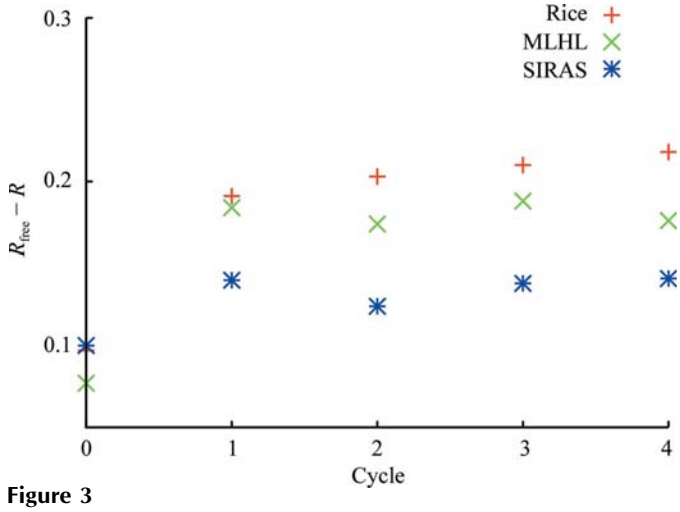

Figure 3
The $R_{\rm free} - R$ difference in the first five *ARP*/*wARP* macrocycles of the 2b79 pipelines with the univariate phasing function. Since the number of residues built was similar (approximately 50%) for all target functions in the first cycles of the model building (and the same holds for the numbers of model parameters and restraints), the comparison of the difference for different target functions can be used as an estimator of relative overfitting. The final value of the ratio in each refinement block is reported in order to compare values close to the convergence of refinement.

mation in the MAD experiment: according to a preliminary analysis, a modified partial localization of the maximum could also be applied to the four-dimensional two-wavelength MAD integration problem. Since the MAD experiment is a popular method for solving the phase problem in protein X-ray crystallography, a proper implementation of the MAD function with the direct use of prior phase information and modelling all the correlations is a challenge for the future.

APPENDIX *A*
Derivation of the required distribution

Since the conditional probability distribution of the three observed structure-factor amplitudes given three model structure factors can be derived in analogy to the derivation of the SAD function, the derivation of the SIRAS function will be somewhat compressed here (for more details, see Skubák *et al.*, 2004). Using the central limit theorem, the starting point for the derivation will be the multivariate complex Gaussian probability distribution of structure factors (see, for example, Pannu *et al.*, 2003). $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{F}_3$ will represent the 'observed' structure factors from a SIRAS experiment and $\mathbf{F}_4, \mathbf{F}_5, \ldots, \mathbf{F}_N$ will represent the 'model' structure factors. The amplitude of a structure factor $\mathbf{F}_i$ will be denoted by $|F_i|$ and its phase by α_i.

$$P(\mathbf{F}_1, \mathbf{F}_2, \ldots \mathbf{F}_N) = \frac{1}{\pi^N \det(\mathrm{C}_N)} \exp\left(- \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{F}_i^* z_{ij} \mathbf{F}_j \right). \quad (6)$$

C_N is the Hermitian covariance matrix of the N-dimensional Gaussian probability distribution and z_{ij} denotes the ijth element of the inverse matrix of C_N. After separately summing over the diagonal and off-diagonal terms, transformation to polar coordinates and simplification, we obtain

$$\begin{aligned} &P(|F_1|, \alpha_1, |F_2|, \alpha_2, \ldots, |F_N|, \alpha_N) \\ &= \frac{\prod_{i=1}^{N} |F_i|}{\pi^N \det(\mathrm{C}_N)} \exp\Bigg[- \sum_{i=1}^{N} \Bigg(|F_i|^2 a_{ii} \\ &\quad + \sum_{j=i+1}^{N} \{2|F_i||F_j|[a_{ij}\cos(\alpha_j - \alpha_i) - b_{ij}\sin(\alpha_j - \alpha_i)]\} \Bigg) \Bigg]. \end{aligned} \quad (7)$$

In the above equation, a_{ij} and b_{ij} represent the real and imaginary components of the inverse covariance matrix. The unknown phase angles α_1, α_2, α_3 need to be integrated out.

$$\begin{aligned} &P(|F_1|, |F_2|, |F_3|, |F_4|, \alpha_4, \ldots, |F_N|, \alpha_N) \\ &= \frac{\prod_{i=1}^{N} |F_i|}{\pi^N \det(\mathrm{C}_N)} \exp\Bigg[- \sum_{i=1}^{3} |F_i|^2 a_{ii} - \sum_{i=4}^{N} \Bigg(|F_i|^2 a_{ii} \\ &\quad + \sum_{j=i+1}^{N} \{2|F_i||F_j|[a_{ij}\cos(\alpha_j - \alpha_i) - b_{ij}\sin(\alpha_j - \alpha_i)]\} \Bigg) \Bigg] \\ &\quad \times \int_0^{2\pi} \int_0^{2\pi} \exp\Bigg(- \sum_{j=2}^{3} \sum_{i=j+1}^{N} \{2|F_j||F_i|[a_{ji}\cos(\alpha_i - \alpha_j) - b_{ji}\sin(\alpha_i - \alpha_j)]\} \Bigg) \\ &\quad \times \int_0^{2\pi} \exp\Bigg(- \sum_{i=2}^{N} \{2|F_1||F_i|[a_{1i}\cos(\alpha_i - \alpha_1) \\ &\quad - b_{1i}\sin(\alpha_i - \alpha_1)]\} \Bigg) \mathrm{d}\alpha_1 \, \mathrm{d}\alpha_2 \, \mathrm{d}\alpha_3. \end{aligned} \quad (8)$$

The inner integral can be solved analytically:

$$\begin{aligned} &\int_0^{2\pi} \exp\Bigg(- \sum_{i=2}^{N} \{2|F_1||F_i|[a_{1i}\cos(\alpha_i - \alpha_1) - b_{1i}\sin(\alpha_i - \alpha_1)]\} \Bigg) \mathrm{d}\alpha_1 \\ &\quad = 2\pi I_0 \Bigg(4|F_1|^2 \Bigg\{ \Bigg[\sum_{i=2}^{N} |F_i|(a_{1i}\cos\alpha_i - b_{1i}\sin\alpha_i) \Bigg]^2 \\ &\quad + \Bigg[\sum_{i=2}^{N} |F_i|(a_{1i}\sin\alpha_i + b_{1i}\cos\alpha_i) \Bigg]^2 \Bigg\} \Bigg)^{1/2}, \end{aligned} \quad (9)$$

where $I_0(\mathbf{x})$ is the zero-order modified Bessel function of the first kind.

From the definition of conditional probability, the required probability distribution can be obtained as follows,

$$\begin{aligned} &P(|F_1|, |F_2|, |F_3|; |F_4|, \alpha_4, \ldots, |F_N|, \alpha_N) \\ &\quad = \frac{P(|F_1|, |F_2|, |F_3|, |F_4|, \alpha_4, \ldots, |F_N|, \alpha_N)}{P(|F_4|, \alpha_4, \ldots, |F_N|, \alpha_N)}. \end{aligned} \quad (10)$$

$P(|F_1|, |F_2|, |F_3|, |F_4|, \alpha_4, \ldots, |F_N|, \alpha_N)$ is given by (8) and (9) and $P(|F_4|, \alpha_4, \ldots, |F_N|, \alpha_N)$ can be obtained by (7), denoting the corresponding covariance matrix by C_{N-3} and the ijth element of its inverse by $c_{ij} + id_{ij}$. Thus, the required distribution is

$$\begin{aligned} &P(|F_1|, |F_2|, |F_3|; |F_4|, \alpha_4, \ldots, |F_N|, \alpha_N) \\ &= \frac{2|F_1| \ldots |F_3| \det(\mathrm{C}_{N-3})}{\pi^2 \det(\mathrm{C}_N)} \exp\Bigg[- \sum_{i=1}^{3} |F_i|^2 a_{ii} \\ &\quad - \sum_{i=4}^{N} \Bigg(|F_i|^2 (a_{ii} - c_{ii}) + \sum_{j=i+1}^{N} \{2|F_i||F_j|[(a_{ij} - c_{ij})\cos(\alpha_j - \alpha_i) \\ &\quad - (b_{ij} - d_{ij})\sin(\alpha_j - \alpha_i)]\} \Bigg) \Bigg] \\ &\quad \times \int_0^{2\pi} \int_0^{2\pi} \exp\Bigg(- \sum_{j=2}^{3} \sum_{i=j+1}^{N} \{2|F_j||F_i|[a_{ji}\cos(\alpha_i - \alpha_j) \\ &\quad - b_{ji}\sin(\alpha_i - \alpha_j)]\} \Bigg) I_0[2|F_1|\xi(\alpha_2, \alpha_3)] \, \mathrm{d}\alpha_2 \, \mathrm{d}\alpha_3, \end{aligned} \quad (11)$$

where

$$\begin{aligned} \xi(\alpha_2, \alpha_3, \ldots, \alpha_N) = \Bigg(\sum_{i=2}^{N} \Bigg\{ &|F_i|^2 (a_{1i}^2 + b_{1i}^2) \\ &+ \sum_{j=i+1}^{N} 2|F_i||F_j|[(a_{1i}a_{1j} + b_{1i}b_{1j})\cos(\alpha_j - \alpha_i) \\ &+ (a_{1j}b_{1i} - a_{1i}b_{1j})\sin(\alpha_j - \alpha_i)] \Bigg\} \Bigg)^{1/2}. \end{aligned} \quad (12)$$

APPENDIX *B*
The covariance matrix

In order to take into account all the correlations between the three observations and three models, a 6 × 6 covariance matrix must be constructed:

$$C_6 = \begin{bmatrix} \Sigma_N+2(\sigma_o^N)^2 & D_2\Sigma_N & D_2\Sigma_N & D_1\Sigma_P & D_3\Sigma_P & D_3\Sigma_P \\ D_2\Sigma_N & \Sigma_{N2}+2(\sigma_o^+)^2 & \Sigma'_{N2} & D_3\Sigma_P & D_1\Sigma_{P2} & D_1\Sigma'_{P2} \\ D_2\Sigma_N & \Sigma_{N2} & \Sigma_{N2}+2(\sigma_o^-)^2 & D_3\Sigma_P & D_1\Sigma'_{P2} & D_1\Sigma_{P2} \\ D_1\Sigma_P & D_3\Sigma_P & D_3\Sigma_P & \Sigma_P & D_2\Sigma_P & D_2\Sigma_P \\ D_3\Sigma_P & D_1\Sigma_{P2} & D_1\Sigma'_{P2} & D_2\Sigma_P & \Sigma_{P2} & \Sigma'_{P2} \\ D_3\Sigma_P & D_1\Sigma'_{P2} & D_1\Sigma_{P2} & D_2\Sigma_P & \Sigma'_{P2} & \Sigma_{P2} \end{bmatrix}, \tag{13}$$

with the model part covariance matrix C_3 being the right bottom 3×3 submatrix of (13),

$$C_3 = \begin{pmatrix} \Sigma_P & D_2\Sigma_P & D_2\Sigma_P \\ D_2\Sigma_P & \Sigma_{P2} & \Sigma'_{P2} \\ D_2\Sigma_P & \Sigma'_{P2} & \Sigma_{P2} \end{pmatrix}, \tag{14}$$

where D is a refinable Luzzati (1952) error parameter which absorbs the errors in both model phases and amplitudes: the D_1 parameter accounts for the errors between the observed and calculated phases and amplitudes, the D_2 error parameter accounts for the errors between the native and derivative structure factors caused by non-isomorphism and D_3 accounts for the combination of these errors. In general, the covariance term Σ'_{N2} is complex; however, the imaginary term is small compared with the real term for a large number of reflections and is thus omitted. Furthermore, the real part of this term is a function of the difference between 'observed' phases which are unknown and is approximated by the difference between the model phases. The following covariance-matrix terms arise,

$$\begin{aligned} \Sigma_N &= \langle |F_o^N|^2 \rangle \\ \Sigma_P &= \langle |F_c^N|^2 \rangle \\ \Sigma_{N2} &= \frac{\langle |F_o^+|^2 + |F_o^-|^2 \rangle}{2} \\ \Sigma'_{N2} &= \langle |F_o^+||F_o^-| \cos(\alpha_c^+ - \alpha_c^-) \rangle \\ \Sigma_{P2} &= \frac{\langle |F_c^+|^2 + |F_c^-|^2 \rangle}{2} \\ \Sigma'_{P2} &= \langle |F_c^+||F_c^-| \cos(\alpha_c^+ - \alpha_c^-) \rangle. \end{aligned} \tag{15}$$

APPENDIX C
Properties of the three-dimensional SIRAS integral

Let us consider the integral and its properties before the analytical integration is performed. From (8), after disregarding the anomalous terms (see Appendix B), the integral is as follows:

$$I \equiv \int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \exp\left[\sum_{i=1}^{3}\sum_{j=i+1}^{N} 2|F_i||F_j|a_{ij}\cos(\alpha_i - \alpha_j)\right] d\alpha_1\, d\alpha_2\, d\alpha_3. \tag{16}$$

For simplicity, let us define the w_{ij} term as

$$w_{ij} \equiv 2|F_i||F_j|a_{ij}, \tag{17}$$

then after expanding the integrand by using the trigonometric relations and rearranging the terms we obtain

$$\begin{aligned} I &= \int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \exp\left[\sum_{i=1}^{3}\sum_{j=i+1}^{N} w_{ij}\cos(\alpha_i - \alpha_j)\right] d\alpha_1\, d\alpha_2\, d\alpha_3 \\ &= \int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \exp\Big\{ - w_{12}\cos(\alpha_1 - \alpha_2) - w_{13}\cos(\alpha_1 - \alpha_3) \\ &\quad - w_{23}\cos(\alpha_2 - \alpha_3) - \sum_{i=4}^{N}[w_{1i}\cos(\alpha_i - \alpha_1) \\ &\quad + w_{2i}\cos(\alpha_i - \alpha_2) + w_{3i}\cos(\alpha_i - \alpha_3)]\Big\} d\alpha_1\, d\alpha_2\, d\alpha_3 \\ &= \int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \exp\Big[- w_{12}\cos(\alpha_1 - \alpha_2) - w_{13}\cos(\alpha_1 - \alpha_3) \\ &\quad - w_{23}\cos(\alpha_2 - \alpha_3) - \cos(\alpha_1)\sum_{i=4}^{N} w_{1i}\cos(\alpha_i) \\ &\quad - \sin(\alpha_1)\sum_{i=4}^{N} w_{1i}\sin(\alpha_i) - \cos(\alpha_2)\sum_{i=4}^{N} w_{2i}\cos(\alpha_i) \\ &\quad - \sin(\alpha_2)\sum_{i=4}^{N} w_{2i}\sin(\alpha_i) - \cos(\alpha_3)\sum_{i=4}^{N} w_{3i}\cos(\alpha_i) \\ &\quad - \sin(\alpha_3)\sum_{i=4}^{N} w_{3i}\sin(\alpha_i)\Big] d\alpha_1\, d\alpha_2\, d\alpha_3. \end{aligned} \tag{18}$$

If we now define vectors $\mathbf{W}_i$, $i = 1, 2, 3$, by

$$\mathbf{W}_i = (W_i^c, W_i^s) \equiv \left[\sum_{j=4}^{N} w_{ij}\cos(\alpha_j), \sum_{j=4}^{N} w_{ij}\sin(\alpha_j)\right] \tag{19}$$

and denote their modulus and polar angle by W_i and φ_i, respectively, then

$$\begin{aligned} I = \int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \exp\Big[& - w_{12}\cos(\alpha_1 - \alpha_2) - w_{13}\cos(\alpha_1 - \alpha_3) \\ & - w_{23}\cos(\alpha_2 - \alpha_3) - \cos(\alpha_1)W_1^c - \sin(\alpha_1)W_1^s - \cos(\alpha_2)W_2^c \\ & - \sin(\alpha_2)W_2^s - \cos(\alpha_3)W_3^c - \sin(\alpha_3)W_3^s\Big] d\alpha_1\, d\alpha_2\, d\alpha_3 \end{aligned} \tag{20}$$

and we obtain the simplified form of the integral with the integrand consisting of only six terms:

$$\begin{aligned} I = \int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \exp\Big[& - w_{12}\cos(\alpha_1 - \alpha_2) - w_{13}\cos(\alpha_1 - \alpha_3) \\ & - w_{23}\cos(\alpha_2 - \alpha_3) - W_1\cos(\alpha_1 - \varphi_1) - W_2\cos(\alpha_2 - \varphi_2) \\ & - W_3\cos(\alpha_3 - \varphi_3)\Big] d\alpha_1\, d\alpha_2\, d\alpha_3. \end{aligned} \tag{21}$$

Thus, the integral depends on nine real-number parameters: the w parameters W_1, W_2, W_3, w_{12}, w_{13}, w_{23} and phases φ_1, φ_2, φ_3, so we can look at it as a function of nine real variables $I = I(W_1, W_2, W_3, w_{12}, w_{13}, w_{23}, \varphi_1, \varphi_2, \varphi_3)$ (in the following, the set of nine variables will be denoted as an ennead). We can now reduce the range of the definition of this function.

Let the ennead $\boldsymbol{\varepsilon} \equiv (W_1, W_2, W_3, w_{12}, w_{13}, w_{23}, \varphi_1, \varphi_2, \varphi_3)$ be I-equivalent to ennead $(W'_1, W'_2, W'_3, w'_{12}, w'_{13}, w'_{23}, \varphi'_1, \varphi'_2, \varphi'_3)$ if $I(W_1, W_2, W_3, w_{12}, w_{13}, w_{23}, \varphi_1, \varphi_2, \varphi_3)$ is the same as $I(W'_1, W'_2, W'_3, w'_{12}, w'_{13}, w'_{23}, \varphi'_1, \varphi'_2, \varphi'_3)$.

Furthermore, define w_{ab} as w-least in $\boldsymbol{\varepsilon}$ if $|w_{ab}| \leq |w_{12}|$, $|w_{ab}| \leq |w_{13}|$ and $|w_{ab}| \leq |w_{23}|$.

The following statement holds.

Statement. For any ennead $\boldsymbol{\varepsilon} \equiv (W_1, W_2, W_3, w_{12}, w_{13}, w_{23}, \varphi_1, \varphi_2, \varphi_3)$ an ennead $\boldsymbol{\varepsilon}' \equiv (W_1', W_2', W_3', w_{12}', w_{13}', w_{23}', \varphi_1', \varphi_2', 0)$ exists which is I-equivalent with $\boldsymbol{\varepsilon}$ and for which all W_1', W_2', W_3' are nonpositive and all $w_{12}', w_{13}', w_{23}'$ up to the w-least in $\boldsymbol{\varepsilon}'$ are nonpositive.

Proof. We will construct $\boldsymbol{\varepsilon}'$ in two steps. At first, let us construct $\boldsymbol{\varepsilon}'' \equiv (W_1'', W_2'', W_3'', w_{12}'', w_{13}'', w_{23}'', \varphi_1'', \varphi_2'', \varphi_3'')$ I-equivalent with $\boldsymbol{\varepsilon}$ for which all $w_{12}'', w_{13}'', w_{23}''$ up to the w-least in $\boldsymbol{\varepsilon}''$ are nonpositive. Four distinct cases can occur:

(i) All w_{12}, w_{13}, w_{23} are nonpositive. Then, trivially, $\boldsymbol{\varepsilon}'' = \boldsymbol{\varepsilon}$.

(ii) Exactly one of w_{12}, w_{13}, w_{23} is positive. If w-least in $\boldsymbol{\varepsilon}$ is positive, $\boldsymbol{\varepsilon}'' = \boldsymbol{\varepsilon}$. Let us assume that the only positive parameter is not w-least in $\boldsymbol{\varepsilon}$. Because of the formal symmetry of I with regards to indices 1, 2, 3, we can freely choose w_{12} to be positive and w_{13} to be (nonpositive) w-least in $\boldsymbol{\varepsilon}$ without the loss of generality (the proof would be symbolically the same for any other permutation of positive and w-least variables). We perform the following linear transformation of the integral from $(\alpha_1, \alpha_2, \alpha_3)$ to $(\alpha_1', \alpha_2', \alpha_3')$,

$$\begin{aligned}\alpha_1' &= \alpha_1 - \pi \\ \alpha_2' &= \alpha_2 \\ \alpha_3' &= \alpha_3\end{aligned} \tag{22}$$

$$\begin{aligned} I = &\int_0^{2\pi}\int_0^{2\pi}\int_{-\pi}^{\pi} \exp\Big[- w_{12}\cos(\alpha_1' + \pi - \alpha_2') \\ &- \ w_{13}\cos(\alpha_1' + \pi - \alpha_3') - w_{23}\cos(\alpha_2' - \alpha_3') \\ &- W_1\cos(\alpha_1' + \pi - \varphi_1) - W_2\cos(\alpha_2' - \varphi_2) \\ &- W_3\cos(\alpha_3' - \varphi_3)\Big]\,\mathrm{d}\alpha_1'\mathrm{d}\alpha_2'\mathrm{d}\alpha_3' \\ = &\int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \exp\Big[w_{12}\cos(\alpha_1' - \alpha_2') + w_{13}\cos(\alpha_1' - \alpha_3') \\ &- w_{23}\cos(\alpha_2' - \alpha_3') - W_1\cos(\alpha_1' + \pi - \varphi_1) \\ &- W_2\cos(\alpha_2' - \varphi_2) - W_3\cos(\alpha_3' - \varphi_3)\Big]\,\mathrm{d}\alpha_1'\,\mathrm{d}\alpha_2'\,\mathrm{d}\alpha_3'. \end{aligned} \tag{23}$$

If we now set $w_{12}'' \equiv -w_{12}$, $w_{13}'' \equiv -w_{13}$, $\varphi_1'' \equiv \varphi_1 - \pi$ then

$$\begin{aligned} I = &\int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \exp\Big[- w_{12}''\cos(\alpha_1' - \alpha_2') - w_{13}''\cos(\alpha_1' - \alpha_3') \\ &- \ w_{23}\cos(\alpha_2' - \alpha_3') - W_1\cos(\alpha_1' - \varphi_1'') - W_2\cos(\alpha_2' - \varphi_2) \\ &- W_3\cos(\alpha_3' - \varphi_3)\Big]\,\mathrm{d}\alpha_1'\,\mathrm{d}\alpha_2'\,\mathrm{d}\alpha_3'. \end{aligned} \tag{24}$$

Thus, $\boldsymbol{\varepsilon}'' = (W_1'', W_2'', W_3'', w_{12}'', w_{13}'', w_{23}'', \varphi_1'', \varphi_2'', \varphi_3'') = (W_1, W_2, W_3, -w_{12}, -w_{13}, w_{23}, \varphi_1 - \pi, \varphi_2, \varphi_3)$ is I-equivalent with $\boldsymbol{\varepsilon}$, w_{13}'' is the w-least in $\boldsymbol{\varepsilon}''$ and both w_{12}'', w_{23}'' are nonpositive.

(iii) Exactly two of w_{12}, w_{13}, w_{23} are positive. Again, we can freely choose w_{12} and w_{13} to be positive and the proof would be symbolically the same for any other choice. The linear transformation (22) shows that $\boldsymbol{\varepsilon}'' = (W_1, W_2, W_3, -w_{12}, -w_{13}, w_{23}, \varphi_1 - \pi, \varphi_2, \varphi_3)$ is I-equivalent with $\boldsymbol{\varepsilon}$ and all $w_{12}'', w_{13}'', w_{23}''$ are nonpositive.

(iv) All w_{12}, w_{13}, w_{23} are positive. If we choose w_{23} to be w-least in $\boldsymbol{\varepsilon}$, then again the transformation (22) ensures that $\boldsymbol{\varepsilon}'' = (W_1, W_2, W_3, -w_{12}, -w_{13}, w_{23}, \varphi_1 - \pi, \varphi_2, \varphi_3)$ is I-equivalent with $\boldsymbol{\varepsilon}$, w_{23}'' is w-least in $\boldsymbol{\varepsilon}'$ and w_{12}'', w_{13}'' are nonpositive.

Now $\boldsymbol{\varepsilon}$ will be constructed from $\boldsymbol{\varepsilon}''$. The transformation

$$\begin{aligned}\alpha_1' &= \alpha_1 - \varphi_3'' \\ \alpha_2' &= \alpha_2 - \varphi_3'' \\ \alpha_3' &= \alpha_3 - \varphi_3''\end{aligned} \tag{25}$$

turns $I(\boldsymbol{\varepsilon}'')$ into

$$\begin{aligned} I = &\int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \exp\Big[- w_{12}''\cos(\alpha_1' - \alpha_2') - w_{13}''\cos(\alpha_1' - \alpha_3') \\ &- w_{23}''\cos(\alpha_2' - \alpha_3') - W_1''\cos(\alpha_1' - \varphi_1'' + \varphi_3'') \\ &- W_2''\cos(\alpha_2' - \varphi_2'' + \varphi_3'')W_3''\cos(\alpha_3')\Big]\,\mathrm{d}\alpha_1'\,\mathrm{d}\alpha_2'\,\mathrm{d}\alpha_3'. \end{aligned} \tag{26}$$

We define $\boldsymbol{\varepsilon}' = (W_1', W_2', W_3', w_{12}', w_{13}', w_{23}', \varphi_1', \varphi_2', \varphi_3')$ by

$$W_i' \equiv \begin{cases} W_i'' & \text{if } W_i'' \leq 0 \\ -W_i'' & \text{if } W_i'' > 0 \end{cases} \tag{27}$$

$$\varphi_i' \equiv \begin{cases} \varphi_i'' - \varphi_3'' & \text{if } W_i'' \leq 0 \\ \varphi_i'' + \pi - \varphi_3'' & \text{if } W_i'' > 0 \end{cases} \tag{28}$$

$$w_ij' \equiv w_ij'' \tag{29}$$

Now the I-equivalency of $\boldsymbol{\varepsilon}'$ with $\boldsymbol{\varepsilon}''$ is shown for the case of all W_1'', W_2'', W_3'' being nonpositive and the following property of the cos() function

$$W_i''\cos(\alpha_i' - \varphi_i'') = -W_i''\cos(\alpha_i' - \varphi_i'' - \pi) \tag{30}$$

shows that $\boldsymbol{\varepsilon}'$ is also I-equivalent with $\boldsymbol{\varepsilon}''$ for any positive W_i''. Since $\boldsymbol{\varepsilon}''$ is I-equivalent with $\boldsymbol{\varepsilon}$ and I-equivalency is transitive by definition, we obtain that $\boldsymbol{\varepsilon}'$ is I-equivalent with $\boldsymbol{\varepsilon}$. Clearly, all $W_1', W_2', W_3', w_{12}', w_{13}', w_{23}'$ up to w-least in $\boldsymbol{\varepsilon}'$ are nonpositive.

Since the proof is constructive, it provides a way of transforming any integral $I(\boldsymbol{\varepsilon})$ coming from real data to $I(\boldsymbol{\varepsilon}')$, reducing the definition range of I. Because φ_3' is fixed (zero), we could reduce the ennead $\boldsymbol{\varepsilon}'$ into an octad. However, this would break the formal symmetry of I, causing several formulae to become slightly more complicated. Therefore, the ennead form will be used throughout. For simplicity, the primes will be omitted and $\boldsymbol{\varepsilon}$ will be used instead of $\boldsymbol{\varepsilon}'$.

APPENDIX *D*
Integrand maximum localization

The following statement provides partial localization of the maximum position if certain conditions hold for $\boldsymbol{\varepsilon}$.

Statement. Let us have the function $F(\alpha_a, \alpha_b, \alpha_c) \equiv \exp[-w_{ab}\cos(\alpha_a - \alpha_b) - w_{ac}\cos(\alpha_a - \alpha_c) - w_{bc}\cos(\alpha_b - \alpha_c) - W_d\cos(\alpha_d - \varphi_d) - W_e\cos(\alpha_e - \varphi_e) - W_f\cos(\alpha_f - \varphi_f)]$, where

$\{a, b, c\} = \{d, e, f\} = \{1, 2, 3\}$, $|w_{ab}| = \max|w_{ij}| > 0$, $|w_{bc}| = \min|w_{ij}|$, $|W_d| = \max|W_i|$ and all W_d, W_e, W_f, w_{ab}, w_{ac} are nonpositive. If

$$w_{ab} \leq \sum_{i\neq d} W_i[1 - \cos(\varphi_i - \varphi_d)] \tag{31}$$

then the maximum of F is at most

$$\arccos\left\{1 - \frac{\sum_{i\neq d} W_i[1 - \cos(\varphi_i - \varphi_d)] - 2\max\{0, w_{bc}\}}{w_{ab}}\right\} \tag{32}$$

distant from the plane $\alpha_a = \alpha_b$ in the three-dimensional Cartesian coordinate system with axes α_a, α_b, α_c.

Proof. Since the exponential function (exp) is an increasing function, it is sufficient to prove the statement for the function $F'(\alpha_a, \alpha_b, \alpha_c) \equiv -w_{ab}\cos(\alpha_a - \alpha_b) - w_{ac}\cos(\alpha_a - \alpha_c) - w_{bc}\cos(\alpha_b - \alpha_c) - W_d\cos(\alpha_d - \varphi_d) - W_e\cos(\alpha_e - \varphi_e) - W_f\cos(\alpha_f - \varphi_f)$. Let us discuss the case when $w_{bc} \leq 0$ first. We need to show that

$$|\alpha_a^{\max} - \alpha_b^{\max}| \leq \arccos\left\{1 - \frac{\sum_{i\neq d} W_i[1 - \cos(\varphi_i - \varphi_d)]}{w_{ab}}\right\}. \tag{33}$$

Clearly,

$$F'(\alpha_a^{\max}, \alpha_b^{\max}, \alpha_c^{\max}) \leq -w_{ab}\cos(\alpha_a^{\max} - \alpha_b^{\max}) - w_{ac} - w_{bc} - \sum_i W_i \tag{34}$$

Take the function value at point $(\varphi_d, \varphi_d, \varphi_d)$:

$$F'(\varphi_d, \varphi_d, \varphi_d) = -w_{ab} - w_{ac} - w_{bc} - W_d - \sum_{i\neq d} W_i\cos(\varphi_i - \varphi_d). \tag{35}$$

Since $F'(\varphi_d, \varphi_d, \varphi_d) \leq F(\alpha_a^{\max}, \alpha_b^{\max}, \alpha_c^{\max})$ from (34) and (35) we obtain

$$-w_{ab} - w_{ac} - w_{bc} - W_d - \sum_{i\neq d} W_i\cos(\varphi_i - \varphi_d) \leq -w_{ab}\cos(\alpha_a^{\max} - \alpha_b^{\max}) - w_{ac} - w_{bc} - \sum_i W_i \tag{36}$$

leading to

$$1 - \frac{\sum_{i\neq d} W_i[1 - \cos(\varphi_i - \varphi_d)]}{w_{ab}} \leq \cos(\alpha_a^{\max} - \alpha_b^{\max}). \tag{37}$$

From the assumptions, $0 \leq 1 - \{\sum_{i\neq d} W_i[1 - \cos(\varphi_i - \varphi_d)]\}/w_{ab} \leq 1$; therefore, the arccosine of this expression is always well defined and (33) holds.

Let $w_{bc} > 0$. Then

$$F'(\alpha_a^{\max}, \alpha_b^{\max}, \alpha_c^{\max}) \leq -w_{ab}\cos(\alpha_a^{\max} - \alpha_b^{\max}) - w_{ac} + w_{bc} - \sum_i W_i, \tag{38}$$

which together with (35) means that

$$-w_{ab} - w_{ac} - w_{bc} - W_d - \sum_{i\neq d} W_i\cos(\varphi_i - \varphi_d) \leq -w_{ab}\cos(\alpha_a^{\max} - \alpha_b^{\max}) - w_{ac} + w_{bc} - \sum_i W_i, \tag{39}$$

$$1 - \frac{\sum_{i\neq d} W_i[1 - \cos(\varphi_i - \varphi_d)] - 2w_{bc}}{w_{ab}} \leq \cos(\alpha_a^{\max} - \alpha_b^{\max}). \tag{40}$$

The assumptions assure that $0 \leq 1 - \{\sum_{i\neq d} W_i[1 - \cos(\varphi_i - \varphi_d)] - 2w_{bc}\}/w_{ab} \leq 1$, thus

$$|\alpha_a^{\max} - \alpha_b^{\max}| \leq \arccos\left\{1 - \frac{\sum_{i\neq d} W_i[1 - \cos(\varphi_i - \varphi_d)] - 2w_{bc}}{w_{ab}}\right\}. \tag{41}$$

In Appendix *C*, we have shown that the validity of all the assumptions of the sentence except of the crucial assumption (31), which is equivalent to

$$|w_{ab}| - |\sum_{i\neq d} W_i[1 - \cos(\varphi_i - \varphi_d)]| > 0. \tag{42}$$

The larger the difference, the better the localization of the maximum. Now the question arises: what are the typical values of this difference in the case of protein SIRAS data? Typically, the structure-factor contributions of heavy atoms are much smaller than the contributions from protein atoms, $F_1 \simeq F_2 \simeq F_3$ and $\alpha_4 \simeq \alpha_5 \simeq \alpha_6$. From $F_1 \simeq F_2 \simeq F_3$ and definition (17),

$$w_{ij} \simeq ka_{ij}, \tag{43}$$

where k is constant for all w_{ij} in this broad approximation. Furthermore, from $\alpha_4 \simeq \alpha_5 \simeq \alpha_6$ and definition (19),

$$W_i \simeq \sum_{j=4}^{6} w_{ij} \simeq k\sum_{j=4}^{6} a_{ij}. \tag{44}$$

Let us now take the definition of the covariance matrix for the SIRAS function C_6 from (13) which must be positive definite. Using the analytical solution of the inverse of the covariance matrix, it can be shown that

$$a_3 = K[-(1 - D_1^2)\Sigma_{H1} + (1 - D_1^2)\Sigma_{H2} + (1 - D_2^2)(\Sigma_N - D_1^2\Sigma_P)], \tag{45}$$

$$a_{24} + a_{25} + a_{26} = a_{34} + a_{35} + a_{36} = KD_1(1 - D_2)(1 - D_1^2)\Sigma_{H1}, \tag{46}$$

where

$$\Sigma_{H1} = \Sigma_{N2} - \Sigma_{N2'} \tag{47}$$

$$\Sigma_{H2} = \Sigma_{N2} - \Sigma_N \tag{48}$$

$$K = \Sigma_{H1}\Sigma_P[\Sigma_{H1} - 2(\Sigma_{H2} + \Sigma_P - D_2^2\Sigma_P)] \times (\Sigma_N - D_1^2\Sigma_P)/\det\Sigma. \tag{49}$$

Usually, $0 < D_i < 1$, which means that

$$D_1(1-D_2)(1-D_1^2)\Sigma_{H1} < (1-D_1^2)\Sigma_{H1}. \quad (50)$$

Furthermore, the real part of the atomic scattering factor $f+f'$ of a heavy atom is typically much larger than its imaginary part f'' and subsequently $\Sigma_{H2} \simeq \sum f_i + f_i' \gg \Sigma_{H1} \simeq 2f_i''$. Since it can be proven that $(\Sigma_N - {D_1}^2\Sigma_P) > 0$ from the positive definiteness of Σ, we obtain

$$(1-D_1^2)\Sigma_{H2} + (1-D_2^2)(\Sigma_N - D_1^2\Sigma_P) \gg (1-D_1^2)\Sigma_{H1} \quad (51)$$

and because of (50) also

$$(1-D_1^2)\Sigma_{H2} + (1-D_2^2)(\Sigma_N - D_1^2\Sigma_P) \gg D_1(1-D_2)(1-D_1^2)\Sigma_{H1}, \quad (52)$$

meaning that $|a_{23}| \gg a_{24}+a_{25}+a_{26}|$, $|a_{23}| \gg a_{34}+a_{35}+a_{36}|$ and subsequently $|w_{23}| \gg |W_2|$, $|w_{23}| \gg |W_3|$ according to (43) and (44). This means that in a typical case $|w_{ab}| \gg |\sum_{i\neq d} W_i[1-\cos(\varphi_i - \varphi_d)]|$ and therefore the assumption (31) of the previous sentence should be fulfilled for typical SIRAS reflections. Indeed, the statistics from several SIRAS data sets shows that (31) is valid for the vast majority (over 99%) of reflections.

APPENDIX *E*
The SIRAS integral evaluation algorithm

Let us assume that the a and b indices from the Appendix D statement are equal to 2 and 3, respectively, *i.e.* the maximum lies close to the plane $\alpha_2 = \alpha_2$ (we have shown that the typical values of these indices are 2 and 3 later in Appendix D). Let us rotate the coordinate system $(\alpha_1, \alpha_2, \alpha_3)$ to $(\alpha_1', \alpha_2', \alpha_3')$ so that the plane $\alpha_2 = \alpha_3$ is equivalent to the plane given by coordinate axes α_1', α_2'. The following transformation can be used,

$$\begin{aligned} \alpha_1' &\equiv \alpha_1 \\ \alpha_2' &\equiv \tfrac{1}{2}(\alpha_2 + \alpha_3) \\ \alpha_3' &\equiv \tfrac{1}{2}(-\alpha_2 + \alpha_3), \end{aligned} \quad (53)$$

transforming the integral $I(\alpha_1, \alpha_2, \alpha_3)$ to

$$\begin{aligned} I = \int_0^{2\pi}\int_0^{2\pi}\int_0^{2\pi} \exp\Big[&- w_{12}\cos(\alpha_1' - \alpha_2' + \alpha_3') \\ &- w_{13}\cos(\alpha_1' - \alpha_2' - \alpha_3') - w_{23}\cos(2\alpha_3') \\ &- W_1\cos(\alpha_1' - \varphi_1) - W_2\cos(\alpha_2' - \alpha_3' - \varphi_2) \\ &- W_3\cos(\alpha_2' + \alpha_3' - \varphi_3)\Big]\, \mathrm{d}\alpha_1'\, \mathrm{d}\alpha_2'\, \mathrm{d}\alpha_3'. \end{aligned} \quad (54)$$

The maximum is now close to the plane given by axes α_1', α_2' and sampling of the variable α_2' over a short range around 0 in the numerical integration is sufficient to cover the peak. The largest required range can be estimated from the maximal distance of the maximum to the plane given by expression (32).

Based on the previous results and discussions, the following algorithm was implemented in *REFMAC*5 for the SIRAS function integral (and its first and second derivatives) calculation.

(i) The ennead $\boldsymbol{\varepsilon}$ is calculated using definitions (17) and (19) and, if required, transformations (22) and (25) are applied so that all the w parameters up to the w-least in $\boldsymbol{\varepsilon}$ are nonpositive.

(ii) The upper limit of the maximum peak height (let us denote it by ζ) is calculated as the sum of the absolute values of all w parameters. If this value is larger than a given threshold and $|w_{ab}|$ is larger than a given threshold, the reflection is classified into class A, otherwise into class B. The reflections in class A can be expected to give rise to larger peaks while the function $I(\alpha_1, \alpha_2, \alpha_3)$ is considered to be flat for class B reflections.

(iii) If the reflection belongs to class A, then the validity of assumption (31) is verified. If the assumption holds, the reflection is classified into class $A1$ and otherwise into class $A2$.

(iv) If the reflection belongs either to class B or to class $A2$ then the required sampling of both integration variables is estimated according to the value of ζ (the higher ζ, the denser the sampling) and numerical integration is performed according to (4), without the transformation (53) of the function and over the whole integration area.

(v) If the reflection belongs to class $A1$, the transformation (53) is performed. The variable α_3' is only sampled over a short range around 0. If the maximal peak-to-plane distance (32) is shorter than a given threshold, the approximation $\alpha_2 \simeq \alpha_3$ may be used, leading to a better estimation of ζ and hence a better sampling estimate.

We thank S. Sun and the JCSG for depositing both native and derivative data in the PDB. Funding for this work was provided by Leiden University and the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). GNM is funded by the Wellcome Trust.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Badger, J. (2003). *Acta Cryst.* D**59**, 823–827.
Beck, T., Krasauskas, A., Gruene, T. & Sheldrick, G. M. (2008). *Acta Cryst.* D**64**, 1179–1182.
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.
Boggon, T. J. & Shapiro, L. (2000). *Structure*, **8**, R143–R149.
Bricogne, G. & Irwin, J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. J. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. (1999). *Acta Cryst.* D**55**, 1555–1567.
Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Acta Cryst.* D**56**, 232–237.
Dauter, Z., Li, M. & Wlodawer, A. (2001). *Acta Cryst.* D**57**, 239–249.
DeLaBarre, B. & Brunger, A. T. (2006). *Acta Cryst.* D**62**, 923–932.
Graaff, R. A. G. de, Hilge, M., van der Plas, J. L. & Abrahams, J. P. (2001). *Acta Cryst.* D**57**, 1857–1862.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *Acta Cryst.* D**64**, 125–132.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.
Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
Matthews, B. W. (1966). *Acta Cryst.* **20**, 82–86.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Ness, S. R., de Graaff, R. A. G., Abrahams, J. P. & Pannu, N. S. (2004). *Structure*, **12**, 1753–1761.
North, A. C. T. (1965). *Acta Cryst.* **18**, 212–216.
Pannu, N. S., McCoy, A. J. & Read, R. J. (2003). *Acta Cryst.* D**59**, 1801–1808.
Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst.* D**54**, 1285–1294.
Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.
Pannu, N. S. & Read, R. J. (2004). *Acta Cryst.* D**60**, 22–27.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.
Skubák, P., Murshudov, G. N. & Pannu, N. S. (2004). *Acta Cryst.* D**60**, 2196–2201.
Skubák, P., Ness, S. & Pannu, N. S. (2005). *Acta Cryst.* D**61**, 1626–1635.
Sun, S., Kondabagil, K., Gentz, P. M., Rossmann, M. G. & Rao, V. B. (2007). *Mol. Cell*, **25**, 943–949.
Szczepanowski, R. H., Filipek, R. & Bochtler, M. (2005). *J. Biol. Chem.* **280**, 22006–22011.
Taylor, G. (2003). *Acta Cryst.* D**59**, 1881–1890.
Wernimont, A. K., Huffman, D. L., Lamb, A. L., O'Halloran, T. V. & Rosenzweig, A. C. (2000). *Nature Struct. Biol.* **7**, 766–771.