J. Appl. Cryst. (1999). **32**, 664–670

Difference structure-factor normalization for heavy-atom or anomalous-scattering substructure determinations

Robert H. Blessing[a*] and G. David Smith[a,b]

[a]*Hauptman-Woodward Institute, 73 High Street, Buffalo, New York 14203-1196, USA, and* [b]*Roswell Park Cancer Institute, Elm and Carlton Streets, Buffalo, New York 14263, USA. E-mail: blessing@hwi.buffalo.edu*

(*Received 23 December 1998; accepted 2 March 1999*)

Abstract

Procedures are described for normalizing structure-factor difference magnitudes, $|\Delta|F||_{\rm SIR} = ||F_{\rm Der}| - |F_{\rm Nat}|| \le |F_{\rm Heavy}|$ or $|\Delta|F||_{\rm SAS} = ||F_{+\bf h}| - |F_{-\bf h}|| \le 2|F''|$, to prepare data for probabilistic direct methods phasing to determine heavy-atom or anomalous-scattering substructures in SIR (single-derivative isomorphous replacement) or SAS (single-wavelength anomalous scattering) cases. Applications of the procedures in several recent determinations of multi-selenium substructures in selenomethionyl proteins *via SnB* direct-methods phasing are briefly summarized.

1. Introduction

Among early efforts to exploit probabilistic phasing methods in protein crystallography were applications of the *MULTAN* program using SIR (single-derivative isomorphous replacement) or SAS (single-wavelength anomalous scattering) difference-magnitude data to determine heavy-atom or anomalous-scattering substructures (Wilson, 1978; Mukherjee *et al.*, 1989). In connection with recent further work on difference-magnitude phasing methods (Langs *et al.*, 1995; Smith *et al.*, 1998; Turner *et al.*, 1998) we have developed a program *DIFFE* that implements the difference-magnitude normalization procedures described below. The *DIFFE* program presumes the results of our earlier data processing programs: *SORTAV* for merging multiple equivalent reflection measurements (Blessing, 1997*a*), *BAYES* for Bayesian post-processing to improve the weak-reflection data (French & Wilson, 1978; Blessing *et al.*, 1998), *LEVY* and *EVAL* for structure-factor normalization (Blessing *et al.*, 1996; 1998), and *LOCSCL* for local scaling to improve the accuracy of SIR or SAS difference magnitudes (Matthews & Czerwinski, 1975; Blessing, 1997*b*).

2. SIR differences

In the SIR case, illustrated schematically in Fig. 1, the derivative structure factor is a vectorial sum on the Argand plane of native and heavy-atom substructure components, $F_{\rm Der} = F_{\rm Nat} + F_{\rm Heavy}$, and the difference magnitudes of interest are

$$|\Delta|F||_{\rm SIR} = ||F_{\rm Der}| - |F_{\rm Nat}|| \le |F_{\rm Heavy}|. \tag{1}$$

Given the corresponding normalized structure-factor magnitudes defined by

$$|E_{\bf h}| = |F_{\bf h}|/\langle|F_{\bf h}|^2\rangle^{1/2} = |F_{\bf h}|/\left(\varepsilon_{\bf h}\sum_{a=1}^{N}|f_a|^2\right)^{1/2}, \tag{2}$$

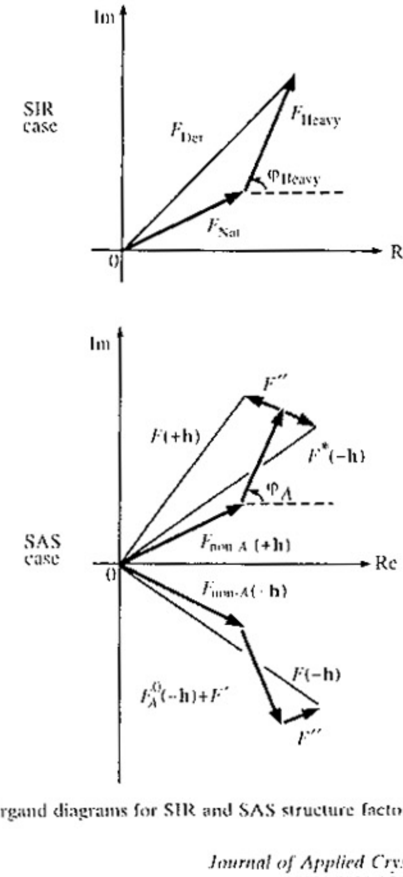

Fig. 1. Argand diagrams for SIR and SAS structure factors.

the difference magnitudes (1) can be calculated as

$$|\Delta|F|| = \varepsilon_{\mathbf{h}}^{1/2}\left|\left(\sum_{a=1}^{N_{\rm Der}}|f_a|^2\right)^{1/2}|E_{\rm Der}| - \left(\sum_{a=1}^{N_{\rm Nat}}|f_a|^2\right)^{1/2}|E_{\rm Nat}|\right|, \tag{3}$$

where the atomic scattering factor magnitudes, $|f_a| = |f_a^0 + f_a' + if_a'| = [(f_a^0 + f_a')^2 + (f_a'')^2]^{1/2}$, allow for the possibility of anomalous scattering, and the reciprocal-lattice point degeneracy factors, $\varepsilon_{\mathbf{h}} = 1$ in space groups $P1$ and $P\bar{1}$, and $\varepsilon_{\mathbf{h}} \geq 1$ in higher symmetry space groups, allow for multiple symmetry-enhancement of the statistically expected values of the squared magnitudes $|F_{\mathbf{h}}|^2$.

We recall that for structure factors $F = |F|\exp(i\varphi)$, with maximum possible magnitudes $|F| \leq \sum_a |f_a|$, the Wilson distributions give the squared-magnitude expectation values $\langle |F|^2\rangle = \varepsilon_{\mathbf{h}} \sum_a |f_a|^2$ that define the normalization (2). By analogy, referring to Fig. 1, we infer that for difference structure factors defined by

$$F_\Delta = |\Delta|F||\exp(i\varphi_{\rm Heavy}), \tag{4}$$

with

$$|\Delta|F|| \leq |F_{\rm Heavy}| \leq \sum_{a=1}^{N_{\rm Heavy}} |f_a|, \tag{5}$$

we should expect squared difference magnitudes with

$$\begin{aligned}\langle|\Delta|F||^2\rangle \leq \langle|F_{\rm Heavy}|^2\rangle &= \varepsilon_{\mathbf{h}}\sum_{a=1}^{N_{\rm Heavy}}|f_a|^2 \\ &= \varepsilon_{\mathbf{h}}\left[\left(\sum_{a=1}^{N_{\rm Der}}|f_a|^2\right) - \left(\sum_{a=1}^{N_{\rm Nat}}|f_a|^2\right)\right].\end{aligned} \tag{6}$$

Therefore, in order to use probabilistic methods to determine the heavy-atom substructure, we calculate greatest-lower-bound estimates of SIR difference E magnitudes (see also Dodson *et al.*, 1975) as

$$\begin{aligned}|E_\Delta| &= \frac{|\Delta|F||}{q\langle F_{\rm Heavy}|^2\rangle^{1/2}} \\ &= \frac{\left|\left(\sum_{a=1}^{N_{\rm Der}}|f_a|^2\right)^{1/2}|E_{\rm Der}| - \left(\sum_{a=1}^{N_{\rm Nat}}|f_a|^2\right)^{1/2}|E_{\rm Nat}|\right|}{q\left[\left(\sum_{a=1}^{N_{\rm Der}}|f_a|^2\right) - \left(\sum_{a=1}^{N_{\rm Nat}}|f_a|^2\right)\right]^{1/2}},\end{aligned} \tag{7}$$

where

$$q = q_0\exp(q_1 s^2 + q_2 s^4) \quad \text{with} \quad s = (\sin\theta_{\mathbf{h}})/\lambda \tag{8}$$

is a least-squares-fitted renormalization scaling function (Blessing *et al.*, 1996, 1998) that imposes the condition $\langle|E_\Delta|^2\rangle = 1$ and is intended to adjust empirically for effects of: imperfect isomorphism of the derivative and native crystals; inaccurately known heavy-atom content due to multiple derivative sites and/or disordered partial site occupancies; and differences between and among the unit-cell distributions of mean-square atomic displacements in the heavy-atom substructure, the derivative crystal and the native protein crystal. The coefficient q_1 in (8) approximates $(\langle B_{\rm Der}\rangle - \langle B_{\rm Nat}\rangle)$, and q_2 approximates $[\langle(B_{\rm Nat} - \langle B_{\rm Nat}\rangle)^2\rangle - \langle(B_{\rm Der} - \langle B_{\rm Der}\rangle)^2\rangle]$.

The *DIFFE* program applies user-supplied data-selection input-cutoff values $t_{\rm max}$, $x_{\rm min}$, $y_{\rm min}$, $s_{\rm min}$ and $s_{\rm max}$, and output cutoff values $z_{\rm min}$ and $z_{\rm max}$. The input cutoffs limit the processing to data pairs with

$$\frac{||E_{\rm Nat}| - |E_{\rm Der}| - \text{median}(|E_{\rm Nat}| - |E_{\rm Der}|)|}{1.25\,\text{median}\big[||E_{\rm Nat}| - |E_{\rm Der}| - \text{median}(|E_{\rm Nat}| - |E_{\rm Der}|)|\big]} \leq t_{\rm max} \tag{9}$$

$$\min\left[\frac{|E_{\rm Nat}|}{\sigma(|E_{\rm Nat}|)}, \frac{|E_{\rm Der}|}{\sigma(|E_{\rm Der}|)}\right] \geq x_{\rm min} \tag{10}$$

$$\frac{||E_{\rm Nat}| - |E_{\rm Der}||}{\left[\sigma^2(|E_{\rm Nat}|) + \sigma^2(|E_{\rm Der}|)\right]^{1/2}} \geq y_{\rm min}, \tag{11}$$

and

$$s_{\rm min} \leq (\sin\theta_{\mathbf{h}})\lambda \leq s_{\rm max}, \tag{12}$$

where typically $t_{\rm max} = 6$, $x_{\rm min} = 3$, $y_{\rm min} = 1$, $s_{\rm min} = 0$, and $s_{\rm max}$ is chosen by inspection of a plot of the spherical shell averages $\langle|E_\Delta|^2\rangle_s$, *versus* $\langle s\rangle$ from a preliminary run of the *DIFFE* program with unlimited $s_{\rm max}$.

The purpose of the $t_{\rm max}$ and $s_{\rm max}$ cutoffs is to prevent generating spuriously large $|E_\Delta|$ values for data pairs that were mismeasured or measured with large uncertainty due to imperfect isomorphism and/or to the general falloff of scattering intensity with increasing scattering angle. The data selection condition (9) corresponds to the assumption that the distribution of standardized values $t = (x - \mu)/\sigma$, corresponding to the differences $x = |E_{\rm Nat}| - |E_{\rm Der}|$, should approximate a zero-mean unit-variance normal distribution,

$$N[(x-\mu)/\sigma] = N(t) = (2\pi)^{-1/2}\exp\left(-t^2/2\right),$$

$$\mu = \langle x\rangle, \quad \sigma = \langle(x - \langle x\rangle)^2\rangle^{1/2} = 1.25\langle|x - \langle x\rangle|\rangle, \tag{13}$$

for which values of $|t| > t_{\rm max} > \sim 3$ are extremely improbable.

After renormalization of the data selected *via* (9)–(12), the program performs an extensive statistical analysis that compares the empirical distribution of the renormalized $|E_\Delta|$ values with theoretical distributions of $|E|$ values for ideal random-atom structures. The program propagates the experimental measurement uncertainties and the error-of-fit uncertainties of the difference renormalization scaling parameters into $\sigma(|E_\Delta|)$ values, and the output cutoffs $z_{\rm min}$ and $z_{\rm max}$ are used to accept, as being sufficiently reliable to be used in subsequent phasing calculations, only data for which

$$|E_\Delta|/\sigma(|E_\Delta|) \geq z_{\min} \tag{14}$$

and

$$(|E_\Delta| - |E_\Delta|_{\max})/\sigma(|E_\Delta|) \leq z_{\max}, \tag{15}$$

where typically $z_{\min} = 3$ and $z_{\max} \leq 3$, and

$$\begin{aligned} |E_\Delta|_{\max} &= \sum_{a=1}^{N_{\rm Heavy}} |f_a| \Big/ \left(\varepsilon_{\mathbf{h}} \sum_{a=1}^{N_{\rm Heavy}} |f_a|^2 \right)^{1/2} \\ &= \frac{\left| \left(\sum_{a=1}^{N_{\rm Der}} |f_a| \right) - \left(\sum_{a=1}^{N_{\rm Nat}} |f_a| \right) \right|}{\varepsilon_{\mathbf{h}}^{1/2} \left| \left(\sum_{a=1}^{N_{\rm Der}} |f_a|^2 \right) - \left(\sum_{a=1}^{N_{\rm Nat}} |f_a|^2 \right) \right|^{1/2}} \end{aligned} \tag{16}$$

is a physical SIR least upper bound.

Since successful phasing of heavy-atom or anomalous-scattering substructures typically requires only the $20n$ to $40n$ largest $|E_\Delta|$ values, where n is the number of independent substructure atoms, it is sometimes advisable to re-run the *DIFFE* program with higher data-selection thresholds $x_{\min}$, $y_{\min}$ and $z_{\min}$, and with $s_{\max} \simeq 0.167$ Å^{-1} ($d_{\min} \simeq 3$ Å), in order to improve agreement between the distribution statistics of the empirical substructure $|E_\Delta|$ values and ideal random-atom $|E|$ values.

3. SAS differences

In the SAS case, illustrated schematically in Fig. 1, the total structure factor is an Argand vectorial sum of normal and anomalous-scattering components, $F = F_{{\rm non}\ A} + F_A^0 + F' + F'' = F^0 + F' + F''$, in which each atomic f'' component has a $\pi/2$ phase lag with respect to its corresponding atomic $f^0 + f'$ component. Friedel pairs of reflections are therefore related by Argand vectorial differences, $F_{+\mathbf{h}} - F_{-\mathbf{h}}^* = 2F''$, and the difference magnitudes of interest are

$$|\Delta|F|| = ||F_{+\mathbf{h}}| - |F_{-\mathbf{h}}|| \leq 2|F''|, \tag{17}$$

which, given the corresponding $|E|$ values, can be calculated as

$$|\Delta|F|| = \left[\varepsilon_{\mathbf{h}} \sum_{a=1}^{N} (f_a^0 + f_a')^2 + (f_a'')^2 \right]^{1/2} ||E_{+\mathbf{h}}| - |E_{-\mathbf{h}}||. \tag{18}$$

Then, again by analogy with Wilson statistics, and again referring to Fig. 1, we infer that for difference structure factors defined by

$$F_\Delta = |\Delta|F|| \exp\{i[\varphi_A + (\pi/2)]\}, \tag{19}$$

with

$$|\Delta|F|| \leq 2|F''| \leq 2 \sum_{a=1}^{N} f_a'' \tag{20}$$

and with φ_A denoting the phase of the $F_A^0 + F'$ component of the structure factor due to the anomalous-scattering substructure, we should expect squared difference magnitudes with

$$\langle |\Delta|F||^2 \rangle \leq 4\langle |F''\rangle = 4\varepsilon_{\mathbf{h}} \sum_{a=1}^{N} (f_a'')^2. \tag{21}$$

Therefore, in order to use probabilistic methods to determine the anomalous-scattering substructure, we calculate greatest-lower-bound estimates of SAS difference E magnitudes as

$$\begin{aligned} |E_\Delta| &= |\Delta|F||/2q\langle |F''|^2 \rangle^{1/2} \\ &= \frac{\left[\sum_{a=1}^{N} (f_a^0 + f_a')^2 + (f_a'')^2 \right]^{1/2} ||E_{+\mathbf{h}}| - |E_{-\mathbf{h}}||}{2q \left[\sum_{a=1}^{N} (f_a'')^2 \right]^{1/2}}, \end{aligned} \tag{22}$$

where, again, $q = q_0 \exp(q_1 s^2 + q_2 s^4)$ with $s = (\sin \theta_{\mathbf{h}})/\lambda$ is a least-squares-fitted renormalization scaling function that imposes the condition $\langle |E_\Delta|^2 \rangle = 1$ and, in the SAS case, is intended to adjust empirically for effects of: inaccurately known chemical composition of the unit cell; multiple sites and/or disordered partial site occupancies in the anomalously scattering substructure; inaccuracies in the values of the anomalous-scattering corrections f' and f'' and in the assumption that they are independent of both the magnitude and direction of $\mathbf{h}$; and differences between the unit-cell distributions of mean-square atomic displacements in the anomalously scattering substructure and the structure overall.

The *DIFFE* program again applies user-supplied data-selection input cutoffs $t_{\max}$, $x_{\min}$, $y_{\min}$, $s_{\min}$ and $s_{\max}$, and output cutoffs $z_{\min}$ and $z_{\max}$. The SAS data selection conditions are analogous to the SIR conditions described above: In (9)–(12), $|F_{+\mathbf{h}}|$ and $|F_{-\mathbf{h}}|$ replace $|F_{\rm Nat}|$ and $|F_{\rm Der}|$; in (9), median$(|F_{+\mathbf{h}}| - |F_{-\mathbf{h}}|) = 0$, in general; and in (15), the physical SAS least upper bound is

$$|E_\Delta|_{\max} = \sum_{a=1}^{N} f_a'' \Big/ \left[\varepsilon_{\mathbf{h}} \sum_{a=1}^{N} (f_a'')^2 \right]^{1/2}. \tag{23}$$

4. Application examples

In an early application, an approximate version of the SIR *DIFFE* normalization procedure was employed with the *SnB* probabilistic phasing program (Miller *et al.*, 1994) to determine the Se$_4$ substructure in the 16 kDa core domain of selenomethionyl avian sarcoma virus

Table 1. *SAS data selection for the Se_{30} substructure in selenomethionyl S-adenosylhomocysteine hydrolase at the 0.9795 Å f″ peak wavelength adjacent to the 0.9802 Å Se K-edge (Turner et al., 1998)*

Space group $C222$; $a = 91.93$, $b = 168.02$, $c = 137.77$ Å.

25163	input reflection pairs with $\min\left[\frac{\lvert F_{+\mathbf{h}}\rvert}{\sigma(\lvert F_{-\mathbf{h}}\rvert)}, \frac{\lvert F_{-\mathbf{h}}\rvert}{\sigma(\lvert F_{-\mathbf{h}}\rvert)}\right] \geq 3.0$ and $39.9 > d_{\mathbf{h}} > 2.8$ Å
13	reflection pairs rejected with $\frac{\lvert\lvert E_{+\mathbf{h}}\rvert - \lvert E_{-\mathbf{h}}\rvert - \mathrm{median}(\lvert E_{-\mathbf{h}}\rvert - \lvert E_{-\mathbf{h}}\rvert)\rvert}{1.25\,\mathrm{median}[\lvert\lvert E_{+\mathbf{h}}\rvert - \lvert E_{-\mathbf{h}}\rvert - \mathrm{median}(\lvert E_{+\mathbf{h}}\rvert - \lvert E_{-\mathbf{h}}\rvert)\rvert]} > t_{\max} = 6.0$
506	additional pairs rejected with $\min\left[\frac{\lvert E_{+\mathbf{h}}\rvert}{\sigma(\lvert E_{-\mathbf{h}}\rvert)}, \frac{\lvert E_{-\mathbf{h}}\rvert}{\sigma(\lvert E_{-\mathbf{h}}\rvert)}\right] < x_{\min} = 3.0$
12187	additional pairs rejected with $\frac{\lvert\lvert E_{+\mathbf{h}}\rvert - \lvert E_{-\mathbf{h}}\rvert\rvert}{[\sigma^2(\lvert E_{+\mathbf{h}}\rvert) + \sigma^2(\lvert E_{-\mathbf{h}}\rvert)]^{1/2}} < y_{\min} = 1.0$
8853	additional pairs rejected with $\frac{\lvert E_\Delta\rvert}{\sigma(\lvert E_\Delta\rvert)} < z_{\min} = 3.0$
0	additional pairs rejected with $\frac{\lvert E_\Delta\rvert - \lvert E_\Delta\rvert_{\max}}{\sigma(\lvert E_\Delta\rvert)} > z_{\max} = 0.0$
3604	$\lvert E_\Delta\rvert$ values for phasing trials. The 600 largest $\lvert E_\Delta\rvert$ values were used for *SnB* phasing

Table 2. *SIR data selection for the $(Se–S)_{16}$ substructure in argininosuccinate synthetase at 1.5418 Å Cu Kα wavelength (Lemke & Howell, 1999)*

Space group $I222$; $a = 79.70$, $b = 105.84$, $c = 127.33$ Å.

35369	input reflection pairs with $\min\left[\frac{\lvert F_{\mathrm{Nat}}\rvert}{\sigma(\lvert F_{\mathrm{Nat}}\rvert)}, \frac{\lvert F_{\mathrm{Der}}\rvert}{\sigma(\lvert F_{\mathrm{Der}}\rvert)}\right] > 3.0$ and $12.0 > d_{\mathbf{h}} > 2.0$ Å
6	reflection pairs rejected with $\frac{\lvert\lvert E_{\mathrm{Nat}}\rvert - \lvert E_{\mathrm{Der}}\rvert - \mathrm{median}(\lvert E_{\mathrm{Nat}}\rvert - \lvert E_{\mathrm{Der}}\rvert)\rvert}{1.25\,\mathrm{median}[\lvert\lvert E_{\mathrm{Nat}}\rvert - \lvert E_{\mathrm{Der}}\rvert - \mathrm{median}(\lvert E_{\mathrm{Nat}}\rvert - \lvert E_{\mathrm{Der}}\rvert)\rvert]} > t_{\max} = 6.0$
2118	additional pairs rejected with $\min\left[\frac{\lvert E_{\mathrm{Nat}}\rvert}{\sigma(\lvert E_{\mathrm{Nat}}\rvert)}, \frac{\lvert E_{\mathrm{Der}}\rvert}{\sigma(\lvert E_{\mathrm{Der}}\rvert)}\right] < x_{\min} = 12.0$
13644	additional pairs rejected with $\frac{\lvert\lvert E_{\mathrm{Nat}}\rvert - \lvert E_{\mathrm{Der}}\rvert\rvert}{[\sigma^2(\lvert E_{\mathrm{Nat}}\rvert) + \sigma^2(\lvert E_{\mathrm{Der}}\rvert)]^{1/2}} < y_{\min} = 4.0$
12353	additional pairs rejected with $d_{\mathbf{h}} < d_{\min} = 3.0$ Å
4183	additional pairs rejected with $\frac{\lvert E_\Delta\rvert}{\sigma(\lvert E_\Delta\rvert)} < z_{\min} = 12.0$
0	additional pairs rejected with $\frac{\lvert E_\Delta\rvert - \lvert E_\Delta\rvert_{\max}}{\sigma(\lvert E_\Delta\rvert)} > z_{\max} = 0.0$
3045	$\lvert E_\Delta\rvert$ values for phasing trials. The 640 largest $\lvert E_\Delta\rvert$ values were used for *SnB* phasing

integrase (Jaskólski & Wlodawer, 1996). This determination utilized (Se–S) difference magnitudes from 3.7 Å resolution subsets of data sets from Se-Met (selenomethionyl) and S-Met (natural methionyl) crystals measured to 2.1 and 2.0 Å resolution, respectively, with a rotating anode Cu *Kα* X-ray source.

More recently, SAS *DIFFE* normalization and *SnB* phasing have been used to determine two larger, multiselenium anomalous-scattering substructures in Se-Met proteins: the Se_8 substructure in the 35 kDa selenomethionyl protein denoted C3d, a fragment of complement component C3 and ligand for complement receptor 2, from data measured to 2.0 Å resolution (Nagar *et al.*, 1998; Smith *et al.*, 1998); and the Se_{30} substructure in the 98 kDa asymmetric crystal chemical unit of selenomethionyl *S*-adenosylhomocysteine hydrolase from data measured to 2.8 Å resolution (Turner *et al.*, 1998). In both these cases, synchrotron data sets were measured at Se *K*-edge, peak, and remote wavelengths for MAD (multi-wavelength anomalous dispersion) phasing, and consistent determinations of the Se substructures were readily obtained treating either the edge or peak data sets as independent SAS cases for the *DIFFE* normalizations and the *SnB* phasing.

Very recently, SIR *DIFFE* normalization of (Se–S) difference magnitudes and *SnB* phasing were used to determine an Se_{16} substructure in the 50 kDa monomer that comprises the asymmetric crystal chemical unit of the tetrameric protein argininosuccinate synthetase (Lemke & Howell, 1999). In this case, 13 of 16 substructure atoms were located unambiguously using differences between data sets measured with Se-Met and S-Met crystals and a home-laboratory rotating-anode Cu *Kα* source before traveling to make synchrotron MAD measurements. The pre-determined Se_{13} substructure then gave sufficient leverage in MAD

Table 3. *Agreement statistics for empirical $|E_\Delta|$ versus ideal $|E_{Se}|$ values*

	Se_{30} SAS	$(Se–S)_{16}$ SIR
Number of reflection pairs	3604	3045
$\langle\|E_\Delta\|^2\rangle$	1.85	1.68
$\langle\|E_{Se}\|^2\rangle$	1.68	1.35
$\langle\|E_\Delta\|\rangle/\langle\|E_{Se}\|\rangle$	1.06	1.15
$\rho = \frac{\langle(\|E_\Delta\| - \langle\|E_\Delta\|\rangle)(\|E_{Se}\| - \langle\|E_{Se}\|\rangle)\rangle}{[\langle(\|E_\Delta\| - \langle\|E_\Delta\|\rangle)^2\rangle\langle(\|E_{Se}\| - \langle\|E_{Se}\|\rangle)^2\rangle]^{1/2}}$	0.62	0.46
$r = \frac{\langle\|\|E_\Delta\| - \|E_{Se}\|\|\rangle}{\langle 0.5(\|E_\Delta\| + \|E_{Se}\|)\rangle}$	0.24	0.36

phasing so that a clear tracing of the protein main chain, and of many side chains as well, was achieved almost as soon as the synchrotron MAD measurements were completed, while still at the synchrotron site. [*Ex post facto*, with the synchrotron MAD data in hand, *SnB* phasing of the peak anomalous SAS *DIFFE* data gave the complete Se_{16} rather than just the partial Se_{13} substructure (Lemke & Howell, 1999).]

Illustrations of *DIFFE* normalization results from the Se_{30} SAS example and the $(Se–S)_{16}$ SIR example are given in Tables 1–3 and Figs. 2 and 3. Tables 1 and 2 illustrate the operation of the data selection conditions (9)–(16) and (23). In the $(Se–S)_{16}$ SIR example (Table 2), the data selection thresholds are unusually high because both the Se-Met and S-Met crystals diffracted very strongly, and the statistical measurement uncertainties $\sigma(|F|^2)/|F|^2$ were unusually small. Tables 1 and 2 also note that in the Se_{30} SAS example the 600 (or $20n_{Se}$) largest $|E_\Delta|$ values were used for *SnB* phasing, while in the $(Se–S)_{16}$ SIR example the 640 (or $40n_{Se}$) largest $|E_\Delta|$ values were used. As will be discussed in a forthcoming paper on experiences with *SnB* phasing of Se substructures in Se-Met proteins using SAS peak, SAS edge and SIR edge-minus-remote *DIFFE* data (Howell *et al.*, 1999), the number of amplitude triplets $|E_\Delta(\mathbf{h})E_\Delta(\mathbf{k})E_\Delta(-\mathbf{h}-\mathbf{k})|$ corresponding to three-phase structure invariants $[\varphi(\mathbf{h}) + \varphi(\mathbf{k}) + \varphi(-\mathbf{h} - \mathbf{k})]$ generated among the selected data is as important as, or more important than, the number of data selected per substructure atom. In this connection, a few very low-order data [*e.g.* $|E_\Delta(222)|$ in the Se_{30} example] can be key links in triplets generation.

Table 3 and Figs. 2 and 3 indicate the overall levels of agreement between the empirically normalized $|E_\Delta|$ values and ideal values

$$|E_{Se}| = \frac{\left|\sum_{a-1}^{N_{Se}} |f_a| \exp[i\delta_a + 2\pi i h r_a - B_a(\sin\theta_{\mathbf{h}})^2/\lambda^2]\right|}{\left\{\varepsilon_{\mathbf{h}} \sum_{a=1}^{N_{Se}} |f_a|^2 \exp[-2B_a(\sin\theta_{\mathbf{h}})^2/\lambda^2]\right\}^{1/2}} \qquad (24)$$

calculated from the refined parameters of the Se substructures. The rather weak agreement statistics (Table 3) and the quite substantial scatter of the ideal *versus* empirical data (Figs. 2*a* and 3*a*) highlight the difficulty of extracting reliable difference magnitude data; however, the normal probability plots (Figs. 2*b* and 3*b*) show that for the most part the empirical $|E_\Delta$ values are in fact normally distributed about the ideal $|E_{Se}$

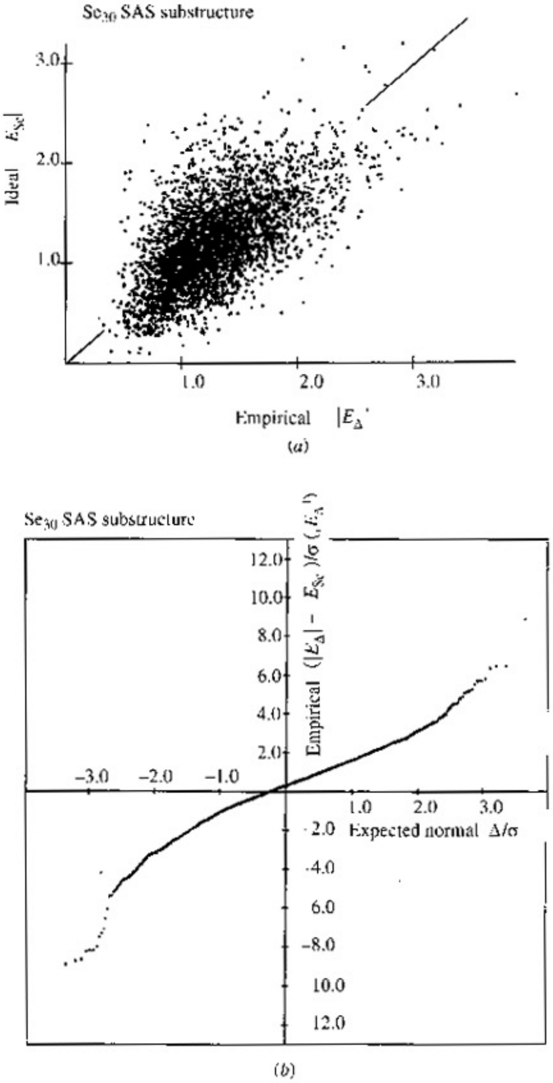

Fig. 2. For the Se_{30} SAS example (Turner *et al.*, 1998): (*a*) scatter plot of ideal $|E_{Se}|$ *versus* empirical $|E_\Delta|$ values about a diagonal line of unit slope; and (*b*) normal probability plot (Howell & Smith, 1992) of the empirical ranked standardized deviates $(|E_\Delta| - |E_{Se}|)/\sigma(|E_\Delta|)$ *versus* ranked standardized deviates Δ/σ expected for a normal distribution. A least-squares straight line through the central 50% of the 3604 normal probability plot points has slope 1.54 and intercept 0.27.

values, albeit with rather large variance. The plots exhibit some sigmoidal curvature, and the linear central portions of the plots have slopes that exceed unity. This shows that the estimated standard uncertainties, $\sigma(|E_\Delta|)$, generally underestimate the $(|E_\Delta| - |E_{Se}|)$ differences – greatly so in the $(Se–S)_{16}$ SIR example. The normal plots also exhibit positive, non-zero intercepts due to extreme, abnormally negative $(|E_\Delta| - |E_{Se}|)$ differences that are not balanced by positive $(|E_\Delta| - |E_{Se}|)$ differences because improbably large $|\Delta|E||$ values were eliminated by the t_{max} data selection (9) applied in deriving the $|E_\Delta|$ values.

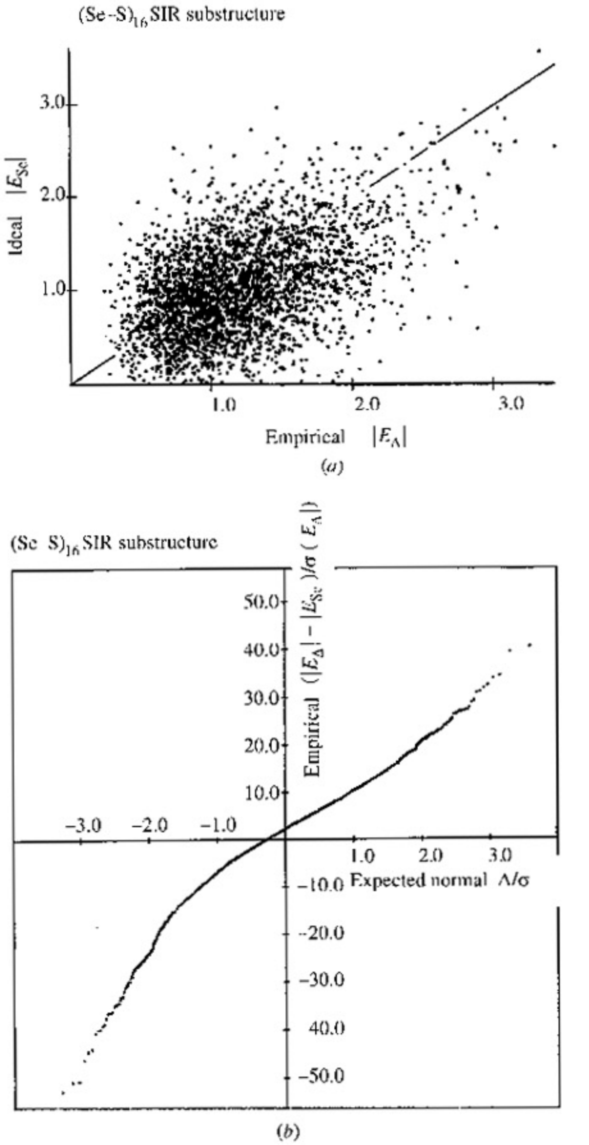

Fig. 3. As Fig. 2, for the $(Se–S)_{16}$ SIR example (Lemke & Howell, 1999). A least-squares straight line through the central 50% of the 3045 normal probability plot points has slope 8.2 and intercept 2.4.

The *DIFFE* normalization procedure is effective apparently because: (*a*) it can select small subsets of reflection pairs (on the order of only a few percent of the measured pairs) that are especially sensitive to the heavy-atom or anomalous-scattering substructure, and (*b*) it can provide usable, albeit rough, approximations to large, substructure $|E|$ values. Finally, we note that, in its present state of development, phasing by probabilistic direct methods requires data to atomic resolution, usually taken to be $d_{min} \leq 1.2$ Å; however, with respect to heavy-atom or anomalous-scattering substructures, resolution as coarse as $d_{min} \simeq 3$ Å is atomic resolution.

The Fortran source code and ASCII text files of users' instructions for the *DIFFE* program and its precursor data processing programs are available from RHB on request. The programs are also available in versions adapted into the *SnB* version 2.0 distribution package (see http://www.hwi.buffalo.edu/SnB/). We are grateful for research support from USDHHS PHS NIH grant No. GM46733, for helpful discussions with our colleagues James Rini, Mary Turner, Lynne Howell, David Langs, and Herbert Hauptman, and for the data provided by Lynne Howell, Mary Turner, and Chris Lemke for Tables 1–3 and Figs. 2 and 3.

References

Blessing, R. H. (1997*a*). *J. Appl. Cryst.* **30**, 421–426.
Blessing, R. H. (1997*b*). *J. Appl. Cryst.* **30**, 176 177.
Blessing, R. H., Guo, D. Y. & Langs, D. A. (1996). *Acta Cryst.* D**52**, 257–266.
Blessing, R. H., Guo, D. Y. & Langs, D. A. (1998). *Direct Methods for Solving Macromolecular Structures*, NATO ASI Series Volume, Series C, *Mathematical and Physical Sciences*, Vol. 507, edited by S. Fortier, pp. 47–71. Dordrecht: Kluwer Academic Publishers.
Dodson, E., Evans, P. & French, S. (1975). *Anomalous Scattering*, edited by S. Ramaseshan & S. C. Abrahams, pp. 423–436. Copenhagen: Munksgaard.
French, S. & Wilson, K. S. (1978). *Acta Cryst.* A**34**, 517–525.
Howell, P. L., Blessing, R. H., Smith, G. D. & Weeks, C. M. (1999). In preparation.
Howell, P. L. & Smith, G. D. (1992). *J. Appl. Cryst.* **25**, 81–86.
Jaskólski, M. & Wlodawer, A. (1996). *Acta Cryst.* D**52**, 1075–1081.
Langs, D. A., Guo, D. Y. & Hauptman, H. A. (1995). *Acta Cryst.* D**51**, 1020–1024.
Lemke, C. & Howell, P. L. (1999). Am. Crystallog. Assoc. Meet., Buffalo, New York, May 1999. In preparation.
Matthews, B. W. & Czerwinski, E. W. (1975). *Acta Cryst.* A**31**, 480–497.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *J. Appl. Cryst.* **27**, 613–621.

Mukherjee, A. K., Helliwell, J. R. & Main, P. (1989). *Acta Cryst.* A**45**, 715–718.

Nagar, B., Jones, R. G., Diefenbach, R. J., Isenman, D. E. & Rini, J. M. (1998). *Science*, **280**, 1277–1281.

Smith, G. D., Nagar, B., Rini, J. M., Hauptman, H. A. & Blessing, R. H. (1998). *Acta Cryst.* **D54**, 799–804.

Turner, M. A., Yuan, C.-S., Borchardt, R. T., Hershfeld, M. S., Smith, G. D. & Howell, P. L. (1998). *Nature Struct. Biol.* **5**, 369–376.

Wilson, K. S. (1978). *Acta Cryst.* B**34**, 1599–1608.