Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Monitoring the anomalous scattering signal and noise levels in X-ray diffraction of crystals

Zheng-Qing Fu, John P. Rose and Bi-Cheng Wang*

Southeast Collaboratory for Structural Genomics, Department of Biochemistry and Molecular Biology, The University of Georgia, Athens, GA 30602, USA

Correspondence e-mail: wang@bcl1.bmb.uga.edu

Received 15 September 2003
Accepted 8 January 2004

A statistical index R_{as} is proposed in order to monitor the overall signal-to-noise ratio in an anomalous scattering data set. In this approach, symmetry-equivalent reflections are merged and grouped into centric and non-centric subsets. Reflections in the centric subset, which in theory should be equal, are used to estimate the noise level in the data. This approach differs from that used by most data-processing programs, in which the centric reflections are merged and averaged. By preserving the differences in centric reflections during data processing, an internal measure of the noise level can be estimated and used to analyze the quality of the anomalous signal in the data. An index R_{as} is defined as the ratio of the average Bijvoet difference of merged acentric reflections to merged centric reflections. Test results on a variety of data show that R_{as} has good correlation with the capability to determine the anomalous scattering substructure from the data. R_{as} can also be useful in monitoring the quality of the data in terms of the data-collection strategy, instrument settings and data-processing software used. R_{as} analysis has been implemented in the program *3DSCALE* as part of a data-processing program suite that is under development in our laboratory.

1. Introduction

Although first demonstrated by Hendrickson & Teeter (1981) and Wang (1982, 1985), protein-structure determination using the weak anomalous scattering signal from elements such as sulfur that are naturally present in most proteins has attracted little attention until recently (Wu *et al.*, 1999, 2001; Dauter *et al.*, 1999, 2002; Liu *et al.*, 2000; Gordon *et al.*, 2001; Li *et al.*, 2002). Common to these studies is the use of single-wavelength anomalous scattering (SAS or SAD) data recorded at or beyond the characteristic Cu $K\alpha$ wavelength (λ = 1.5418 Å).

Compared with traditional MAD (multi-wavelength anomalous dispersion; Hendrickson, 1985) and MIR (multiple isomorphous replacement; Blow & Crick, 1959; Dickerson *et al.*, 1961; North, 1965; Matthews, 1966) methods, the SAS method using S atoms as phasing probes should be more economical and more efficient as the necessity of preparing selenium-labeled protein or the preparation of heavy-atom derivatives is avoided. This approach, once refined, could have significant impact in the large-scale macromolecular structural analysis associated with structural genomics ventures.

The Bijvoet ratio for light atoms such as phosphorus and sulfur is small, of the order of 1% for most protein data (see Ramagopal *et al.*, 2003, and references therein). Thus, an accurate measure of the SAS signal and a means of monitoring the noise level in the data are essential for successful structure

determination from sulfur or phosphorous anomalous scattering data.

Traditionally, the quantity $R_{\rm sym}$ (also called $R_{\rm merge}$),

$$R_{\rm sym}(I) = \sum_h \sum_i |I(h,i) - \langle I(h)\rangle| / \sum_h \sum_i I(h,i), \qquad (1)$$

is used for describing X-ray diffraction quality (Stout & Jensen, 1968; Blundell & Johnson, 1976; McRee, 1993; Drenth, 1994; Ladd & Palmer, 1994). Here, the summation $\sum_h$ runs over the unique reflections, $\sum_i$ runs over all the symmetric equivalents of h and $\langle I(h)\rangle$ is the statistical average. Unfortunately, $R_{\rm sym}$ thus defined has proved to be a poor criterion for assessing the quality of X-ray data (Diederichs & Karplus, 1997; Weiss & Hilgenfeld, 1997; Weiss, 2001), since it generally increases (gets worse) as the data quality improves with increasing redundancy. In recent years, a number of alternative measures have been suggested; these include $R_{\rm meas}$ (or $R_{\rm r.i.m.}$), $R_{\rm mrgd\text{-}F}$ and $R_{\rm p.i.m.}$ (Diederichs & Karplus, 1997; Weiss & Hilgenfeld, 1997; Weiss *et al.*, 1998; Weiss, 2001).

$$R_{\rm meas} = \sum_h [m/(m-1)] \sum_i |I(h,i) - \langle I(h)\rangle| / \sum_h \sum_i I(h,i). \quad (2)$$

$R_{\rm meas}$ (or $R_{\rm r.i.m.}$) overcomes the problem of increasing $R_{\rm sym}$ with increasing data redundancy by including a correction 'm' for redundancy. $R_{\rm mrgd\text{-}F}$, by assessing the quality of the reduced data, enables a direct comparison with the refinement indices $R_{\rm cryst}$ and $R_{\rm free}$ (Kleywegt, 2000). $R_{\rm p.i.m.}$, the so-called precision-indicating merging R index, describes the precision of the averaged measurement (Weiss *et al.*, 1998). These improved R indices provide more reliable indications of the overall data quality of X-ray diffraction experiments and should be used in place of $R_{\rm merge}$ as indicators of overall X-ray data quality. Another generally used statistical parameter for data-quality assessment is $\langle I/\sigma(I)\rangle$, which describes the average strength or significance of the observed intensities. However, none of the above statistical parameters have direct correlation with the anomalous signal in an X-ray diffraction data set. Thus, none of them can practically serve as an efficient indicator of the data quality in terms of anomalous signal or the anomalous signal-to-noise ratio.

Currently, several methods have been proposed to evaluate or estimate the anomalous signal strength; these include $R_{\rm anom}$ (the Bijvoet difference ratio), Δ, the δR plot (normal probability plot) and χ^2.

The Bijvoet difference ratio is defined as

$$R_{\rm anom} = \langle |F_+(h) - F_-(h)|\rangle / \langle F(h)\rangle. \qquad (3)$$

For protein crystals, $R_{\rm anom}$ can be estimated as $(2N_a/N_p)^{1/2} f''/Z_{\rm eff}$ (Hendrickson & Teeter, 1981; Dauter *et al.*, 1999, 2002) or more precisely as $(2N_a/N_p)^{1/2} f''/f_{\rm eff} C_A \exp[\Delta B(\sin\theta/\lambda)^2]$ (Shen *et al.*, 2003). $R_{\rm anom}$ describes the average ratio of Bijvoet difference to structure factor. The error-correction and scaling program *PROSCALE* (Fu *et al.*, 2000) uses another parameter Δ, which is defined as the average ratio of Bijvoet differences in intensity (ΔI) and the standard deviation of the intensity [$\sigma(I)$] calculated using acentric reflections or all reflections,

$$\Delta = \langle |\Delta I|/\sigma(I)\rangle. \qquad (4)$$

Both $R_{\rm anom}$ and Δ describe the averaged Bijvoet differences in a given X-ray diffraction data set, which can be used as an indicator of the anomalous signal level. The reliability of Δ is dependent on the accuracy of evaluation of $\sigma(I)$ and can be improved by the data-collection strategy used (Popov & Bourenkov, 2003). However, Δ as defined above is not really an anomalous signal-to-noise ratio because $\sigma(I)$ does not represent the noise in measured Bijvoet differences. Without the evaluation of the noise level in the Bijvoet differences, $R_{\rm anom}$ and Δ can be misleading.

An alternate approach for estimating the anomalous signal in a data set is the δR plot (normal probability plot) of anomalous differences $\Delta I/\sigma(I)$ suggested by Howell & Smith (1992) and implemented in the scaling program *SCALA* (Evans, 1993) to evaluate anomalous signal. Finally, χ^2 statistics have been used in estimating anomalous signal strength by *SCALEPACK* (Otwinowski & Minor, 1997) with the anomalous flag turned on and off.

While both the δR plot and χ^2 can be used to identify data with anomalous signal, they do not provide a quantitative indication of the anomalous signal level. Thus, an index that directly measures the anomalous signal-to-noise ratio is needed. *XPREP* (Sheldrick, 2000) outputs two different sets of anomalous signal-to-noise ratios (denoted here as SN_1 and SN_2). Test results show that the proposed $R_{\rm as}$ index defined in §2 is a better indicator of the capability of a data set to solve the anomalous scatterer substructure than either SN_1 and SN_2. These results will be discussed in more detail in §3. We propose that $R_{\rm as}$ be used as an informative index in data-reduction programs to monitor the anomalous signal-to-noise ratio and have implemented this approach in the program *3DSCALE* as part of a data-processing program suite currently under development in our laboratory.

2. Methods

For protein crystals, which always belong to a non-centrosymmetric space group, Friedel's law is not obeyed in the presence of anomalous dispersion, *i.e.* $I(\mathbf{h}) \neq I(-\mathbf{h})$ for certain classes of reflections. However, there exists for most non-centrosymmetric space groups (with the exception of P_1, P_3, P_{31}, P_{32} and R_3) a class of reflections that are always centrosymmetric. For these centrosymmetric reflections, $I(\mathbf{h}) = I(-\mathbf{h})$ is still obeyed. The Bijvoet difference, the intensity differences between reflections that are space-group symmetry equivalents to the two members of a Fridel pair, will be zero [ΔIc = $I_{(+)} - I_{(-)} = 0$] for centric reflections and non-zero [ΔIa = $I_{(+)} - I_{(-)} \neq 0$] for acentric reflections. For the following discussion and in our calculations, all $I_{(+)}$ and $I_{(-)}$ represent merged reflections according to space-group symmetry. In principle, the accuracy of $I_{(+)}$ and $I_{(-)}$ should improve with increased data redundancy (which we define as the total number of observations, including symmetry-related ones, per unique reflection). This expectation together with others will be tested in §3.

Table 1
Data statistics.

The values in parentheses for the Δa, Δc and R_{as} columns are the in-shell values that were calculated with reflections in the listed resolution shell. All other values were calculated with reflections up to the resolution shell listed. CC is the correlation coefficient of the top solutions from *SHELXD*, with two values CC/all and CC/weak (in parentheses).

(*a*) Statistics from *3DSCALE* and *SHELXD* for the 60° data set.

Resolution (Å)	R_{sym} (%)	Completeness (%)	Redundancy	$\langle I/\sigma(I)\rangle$	Δa	Δc	R_{as}	CC/all (CC/weak)
4.78	2.48	98.18	6.32	83.31	4.34 (4.34)	2.24 (2.24)	1.94 (1.94)	37.5 (5.4)
3.76	2.35	99.08	6.52	84.57	3.74 (3.20)	1.99 (1.61)	1.88 (1.99)	39.7 (12.6)
3.28	2.39	99.00	6.61	78.57	3.28 (2.44)	1.92 (1.67)	1.71 (1.46)	36.0 (13.1)
2.97	2.44	99.25	6.64	72.43	2.92 (1.93)	1.79 (1.19)	1.63 (1.61)	33.7 (14.9)
2.75	2.52	98.76	6.67	65.68	2.69 (1.81)	1.73 (1.35)	1.55 (1.33)	29.2 (13.0)
2.58	2.60	98.78	6.68	60.39	2.48 (1.48)	1.64 (0.99)	1.51 (1.49)	29.3 (12.7)
2.45	2.67	98.95	6.50	55.45	2.33 (1.42)	1.59 (1.20)	1.46 (1.18)	25.8 (11.9)
2.34	2.71	98.97	6.17	50.85	2.20 (1.28)	1.59 (1.60)	1.38 (0.80)	25.3 (10.2)
2.24	2.75	97.64	5.85	46.66	2.08 (1.09)	1.56 (0.89)	1.33 (1.23)	24.7 (8.2)
2.15	2.76	92.67	5.52	43.93	2.00 (0.83)	1.56 (1.00)	1.28 (0.83)	21.2 (7.6)

(*b*) Statistics from *3DSCALE* and *SHELXD* for the 120° data set.

Resolution (Å)	R_{sym} (%)	Completeness (%)	Redundancy	$\langle I/\sigma(I)\rangle$	Δa	Δc	R_{as}	CC/all (CC/weak)
4.78	2.85	98.18	12.63	105.90	5.36 (5.36)	2.03 (2.03)	2.64 (2.64)	43.8 (19.2)
3.76	2.79	99.09	13.05	107.33	4.40 (3.57)	1.97 (1.87)	2.23 (1.90)	45.4 (18.3)
3.28	2.82	99.00	13.20	99.93	3.85 (2.82)	1.87 (1.53)	2.06 (1.84)	41.0 (16.1)
2.97	2.89	99.28	13.28	92.45	3.43 (2.27)	1.75 (1.22)	1.96 (1.85)	38.5 (18.7)
2.75	2.97	99.27	13.32	84.41	3.11 (1.91)	1.67 (1.18)	1.86 (1.62)	34.0 (14.6)
2.58	3.06	98.89	13.35	77.54	2.85 (1.60)	1.60 (1.04)	1.78 (1.54)	31.8 (12.1)
2.45	3.14	99.05	13.10	71.34	2.67 (1.64)	1.56 (1.21)	1.71 (1.36)	30.2 (12.0)
2.34	3.19	99.00	12.33	65.41	2.49 (1.26)	1.51 (0.95)	1.65 (1.33)	28.7 (10.8)
2.24	3.22	98.11	11.54	59.65	2.34 (1.20)	1.47 (0.77)	1.59 (1.56)	27.9 (10.1)
2.15	3.24	95.89	10.66	55.12	2.22 (0.86)	1.45 (0.77)	1.53 (1.12)	27.4 (9.2)

(*c*) Statistics from *3DSCALE* and *SHELXD* for the 180° data set.

Resolution (Å)	R_{sym} (%)	Completeness (%)	Redundancy	$\langle I/\sigma(I)\rangle$	Δa	Δc	R_{as}	CC/all (CC/weak)
4.78	2.95	99.29	18.89	109.51	5.71 (5.71)	1.58 (1.58)	3.60 (3.60)	44.7 (15.2)
3.76	2.93	99.64	19.50	110.94	4.52 (3.47)	1.57 (1.56)	2.87 (2.23)	47.1 (26.1)
3.28	3.00	99.76	19.72	104.46	3.90 (2.77)	1.55 (1.46)	2.52 (1.89)	44.1 (21.9)
2.97	3.07	99.58	19.83	98.01	3.55 (2.54)	1.50 (1.29)	2.37 (1.96)	40.9 (22.0)
2.75	3.16	99.67	19.91	90.60	3.26 (2.22)	1.48 (1.37)	2.20 (1.62)	36.7 (17.6)
2.58	3.25	99.56	19.94	84.15	3.04 (1.95)	1.46 (1.28)	2.09 (1.53)	35.2 (17.0)
2.45	3.33	99.32	19.42	78.07	2.89 (2.04)	1.45 (1.40)	1.99 (1.45)	32.2 (15.0)
2.34	3.38	99.41	18.37	71.93	2.71 (1.51)	1.43 (1.24)	1.89 (1.22)	31.1 (15.9)
2.24	3.42	98.72	17.19	65.81	2.56 (1.40)	1.41 (1.01)	1.82 (1.39)	28.6 (14.3)
2.15	3.44	97.01	15.84	60.61	2.41 (0.90)	1.40 (1.20)	1.71 (0.75)	28.3 (13.7)

(*d*) Statistics from *3DSCALE* and *SHELXD* for data set 2.

Resolution (Å)	R_{sym} (%)	Completeness (%)	Redundancy	$\langle I/\sigma(I)\rangle$	Δa	Δc	R_{as}	CC/all (CC/weak)
4.78	3.80	98.02	19.41	85.53	4.62 (4.62)	1.84 (1.84)	2.52 (2.52)	43.3 (16.8)
3.76	3.23	98.02	20.07	86.97	3.77 (3.01)	1.68 (1.35)	2.24 (2.24)	38.7 (20.7)
3.28	3.28	98.58	20.31	80.15	3.14 (1.97)	1.64 (1.42)	1.92 (1.39)	36.7 (18.5)
2.97	3.37	98.50	20.48	73.66	2.77 (1.75)	1.61 (1.34)	1.72 (1.30)	34.2 (17.6)
2.75	3.52	98.80	20.56	66.56	2.51 (1.54)	1.57 (1.20)	1.61 (1.28)	31.8 (17.0)
2.58	3.70	99.00	20.60	60.55	2.32 (1.41)	1.60 (1.75)	1.45 (0.81)	28.4 (13.3)
2.45	3.84	98.93	19.94	55.01	2.18 (1.38)	1.58 (1.37)	1.38 (1.01)	22.8 (11.6)
2.34	3.94	98.44	18.81	50.31	2.06 (1.19)	1.59 (1.57)	1.30 (0.76)	24.1 (12.8)
2.24	4.02	98.62	17.51	45.76	1.95 (1.15)	1.59 (1.56)	1.23 (0.74)	22.2 (10.3)
2.15	4.07	98.29	16.17	42.03	1.89 (1.19)	1.57 (1.14)	1.20 (1.04)	14.0 (5.68)

Strictly speaking, ΔIc will not be zero owing to experimental and counting-statistics errors. Based on this, we suggest that ΔIc can be used as an internal indicator to estimate the noise level in the data from non-centrosymmetric crystals which have centric reflections. Conversely, ΔIa, calculated using only the acentric reflections, should give an

estimate for both the anomalous signal and noise in the data; that is,

$$\Delta I\mathrm{a} = \text{signal} + \text{noise}, \tag{5}$$

$$\Delta I\mathrm{c} = \text{noise}. \tag{6}$$

Similarly, the Δ term (4) can be recast to yield the following two parameters that can be calculated during data processing,

$$\Delta a = \langle|\Delta I\mathrm{a}|/\sigma(I)\rangle, \tag{7}$$

$$\Delta c = \langle|\Delta I\mathrm{c}|/\sigma(I)\rangle. \tag{8}$$

Here, Δa and Δc are calculated by using acentric reflections and centric reflections, respectively. Δa represents the measured Bijvoet difference of merged acentric reflections, which contains both signal and noise. Δc represents the measured Bijvoet difference of merged centric reflections, which comes from noise alone.

As a first approximation, if one assumes that the noise level in the same data set is the same for the acentric and centric reflections within the same resolution shell, then we can define

$$R_{\mathrm{as}} = \Delta a/\Delta c. \tag{9}$$

While Δa gives the measured magnitude of the average signal in the data, R_{as} provides a measurement of the significance of Δa in terms of the measured noise level, Δc, in the data. Here, R_{as} can be simply regarded as the measured anomalous signal-to-noise ratio. The larger the R_{as} value, the stronger the anomalous signal, while R_{as} values ≤ 1 would indicate a lack of anomalous signal in the data set. From this it is easy to see that without the knowledge of Δc, the value of Δa alone is not adequate to assess the anomalous signal strength.

Two alternate approaches were also explored. In these approaches, (7) and (8) were redefined as follows:

$$\Delta a = \langle|\Delta I\mathrm{a}|\rangle,$$

$$\Delta c = \langle|\Delta I\mathrm{c}|\rangle$$

or

$$\Delta a = \langle|\Delta I_{\mathrm{a}}|/\sigma^2(I)\rangle,$$

$$\Delta c = \langle|\Delta I_{\mathrm{c}}|/\sigma^2(I)\rangle.$$

However, based on the limited tests described below it appears that (7) and (8) generate a better indicator R_{as} of the anomalous signal-to-noise ratio.

The estimated anomalous scattering signal Δs after correcting for noise can be defined as

$$\Delta s = [(\Delta a)^2 - (\Delta c)^2]^{1/2}. \tag{10}$$

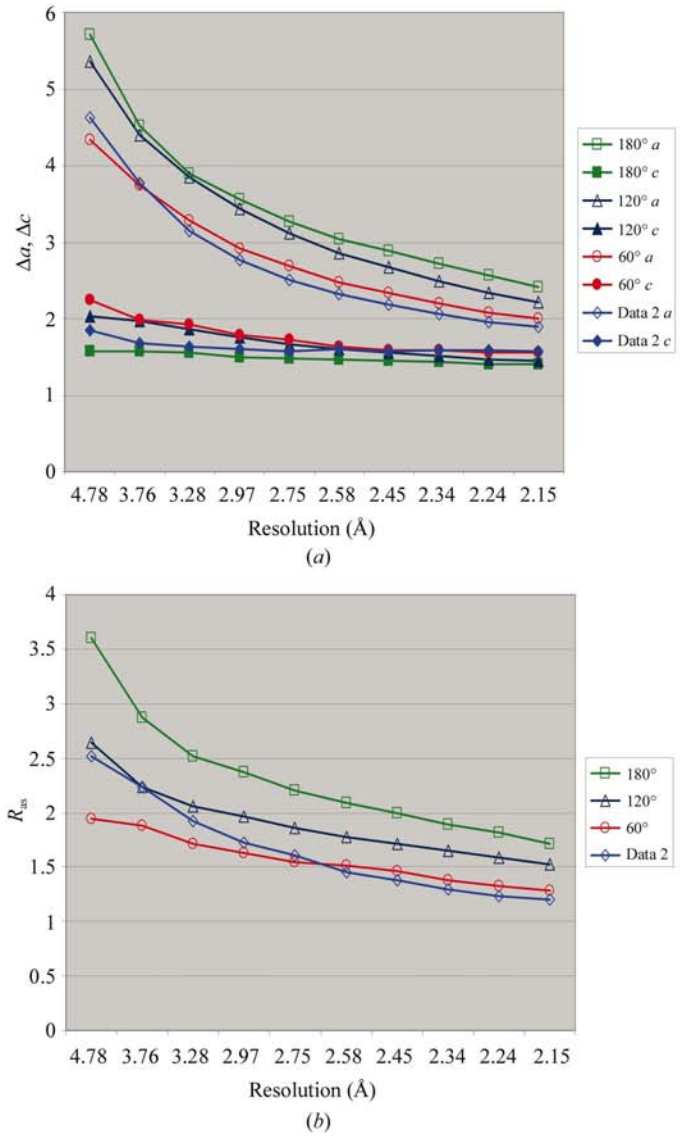


Figure 1
Red, dark blue and green lines represent 60, 120 and 180° data sets from the good-quality Zn-free insulin crystal (crystal 1), respectively. Light blue lines represent data 2 from the moderate-quality Zn-free insulin crystal (crystal 2). The lines with open markers represent Δa. The lines with solid markers represent Δc. (*a*) A plot of Δa and Δc *versus* resolution for the four data sets. (*b*) A plot of R_{as} values *versus* resolution for the four data sets. (*c*) A plot of in-shell R_{as} *versus* resolution shell for the four data sets.

Δs can be used to replace Δa in (9) to calculate another parameter $R'_{\rm as}$ that equals $(R^2_{\rm as} - 1)^{1/2}$.

The $R_{\rm as}$ index described above has been incorporated into *3DSCALE* (Fu, unpublished work), a data-scaling and error-correction program for area-detector data that also uses three-dimensional error models (Fu *et al.*, 2000).

3. Tests and results

3.1. $R_{\rm as}$ *versus* data redundancy

From a theoretical point of view, Δc or noise level should decrease with increased redundancy if the additional observations per unique reflections do not introduce additional systematic error into the data. In the case of Δa, since it is a combination of both signal and noise, it is difficult to anticipate how this value would change with redundancy. To investigate how $R_{\rm as}$ responds to changes in data redundancy, we used three sets of data collected with increasing data redundancy (1×, 2× and 3×) from the same crystal. Two Zn-free insulin crystals (space group $I2_13$; a = 77.95 Å) of different diffraction quality (good and moderate) were chosen for the analysis. The data from the good-quality crystal (crystal 1) were collected to 2.15 Å resolution (lower limit 39 Å) using a Bruker Proteum-R CCD (also known as Smart 6000) detector mounted on a Rigaku RUH3R rotating-anode generator using 5 kW focused (MSC/Blue confocal optics) Cu $K\alpha$ X-rays. A total of 900 0.2° oscillation images were recorded using an exposure time of 1 min. The intensities were indexed and integrated using the *PROTEUM* data-reduction package (Bruker). Data were scaled and merged using *3DSCALE*. Three data sets representing 60, 120 and 180° of crystal rotation were generated by scaling 1–300, 1–600 and 1–900 data frames respectively, which are denoted hereafter as the 60, 120 and 180° data sets.

Data from the moderate-quality crystal (crystal 2) were collected to 2.15 Å (lower limit 39 Å) using the same Bruker Proteum-R CCD detector mounted on a Rigaku FRD generator with Rigaku/MSC HiRes^2 optics. A total of 720 0.25° oscillation images were recorded with an exposure time of 30 s. As above, *PROTEUM* and *3DSCALE* were used to process the data (denoted hereafter as data 2). For each of the above four data sets, values for Δa, Δc and $R_{\rm as}$ were calculated using ten different resolution shells (see Table 1).

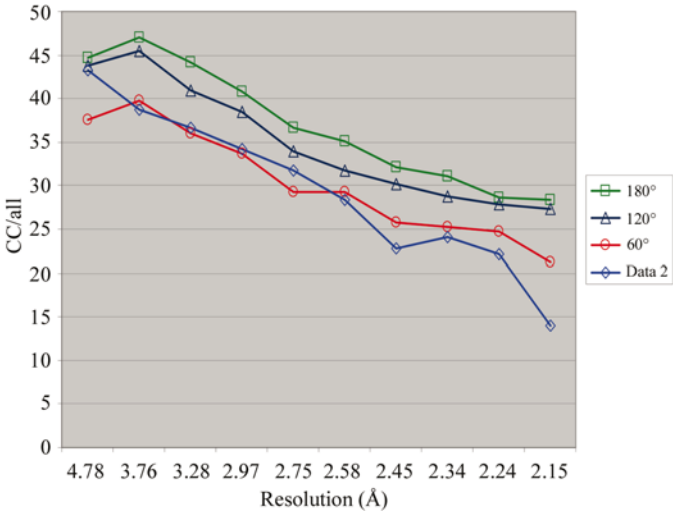


Figure 2
Plot of correlation coefficients CC/all of the top solutions from *SHELXD* *versus* resolution for the four data sets. Red, dark blue, green and light blue lines represent 60, 120 and 180° data sets and data 2, respectively.

For crystal 1, both the 120 and 180° data show a similar Δa distribution *versus* resolution, with the more highly redundant 180° data set having slightly higher Δa values and much lower Δc values. The 60° data set, however, has much lower Δa values with higher Δc values. From the $R_{\rm as}$ plot (Fig. 1*b*), it is interesting to note that there is a clear differentiation between the 60, 120 and 180° data that correlates well with what one would expect in terms of signal strength or data quality with increasing redundancy. In addition, comparing the $R_{\rm as}$ values for the 60, 120 and 180° data sets collected from the same 'good' crystal, it can be seen that all data sets contain anomalous signal that is clearly above the noise level.

If one looks at the plot of Δa *versus* resolution (Fig. 1*a*), one would assume that data 2 is worse than all the other three sets. However, if one looks at $R_{\rm as}$, which accounts for the noise in the data (Δc), which differs significantly from the theoretical value of 0.0, a different picture emerges. According to $R_{\rm as}$, the quality of data 2 is worse than that of either the 120° or the 180° data sets, but it is somewhat better than that of the 60° data set at resolution higher than 2.75 Å. This conclusion is supported by the correlation coefficients observed from *SHELXD* (Sheldrick & Schneider, 2001) described in the next section. Fig. 1(*b*) clearly shows that the signal-to-noise ratio in data 2 drops off much quicker with resolution than those of the three data sets from crystal 1, indicating that crystal 2 is not as good as crystal 1. The redundancy in data 2 is about three times that of the 60° data set. Fig. 1(*b*) indicates that it is possible to produce a set of data from a 'not-so-good' crystal with a better signal-to-noise ratio than that of a 'good' crystal simply by increasing the redundancy of the data set. This possibility is certainly in agreement with the intuition of most experimentalists, but it can now be monitored quantitatively by $R_{\rm as}$.

$R_{\rm as}$ can also provide information on the 'useful' resolution, in terms of anomalous signal, of the data set. If one assumes that an $R_{\rm as}$ value above a certain value, say 1.5, is needed for successful phasing, then a data cutoff of 2.58 Å is the 'useful' resolution for both the 60° and data 2 data, while data cutoffs of 2.15 Å and beyond 2.15 Å (the data limit) are the 'useful' resolutions for the 120° and 180° data sets, respectively. $R_{\rm as}$ can thus serve as a tool for monitoring the change of 'useful' data resolution with reflection redundancy.

Analysis of in-shell $R_{\rm as}$ values *versus* resolution (Table 1 and Fig. 1*c*) reveals that the anomalous signal-to-noise ratios improve more significantly with the 100% increase in redundancy in going from the 60° to the 120° data set than the 50% increase in redundancy observed in going from the 120° to the 180° data set. Again, this observation is in agreement with correlation coefficient from *SHELXD* and is what one would expect.

Table 2
Signal-to-noise ratios from *XPREP*.

SN_1 denotes the ratio based on input σ values and SN_2 the ratio on variances of F_+ and F_- as described in the output of *XPREP*.

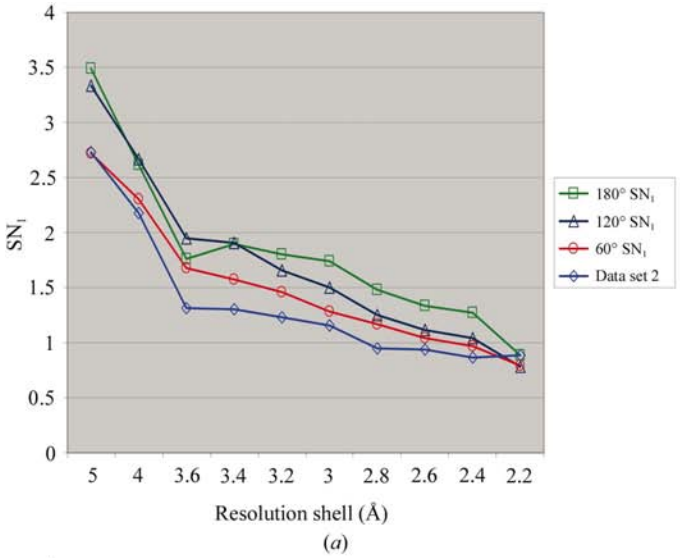
	60° data set		120° data set		180° data set		Data set 2	
Resolution shell	SN_1	SN_2	SN_1	SN_2	SN_1	SN_2	SN_1	SN_2
5.0	2.72	2.83	3.33	3.06	3.49	3.45	2.73	3.13
4.0	2.30	2.34	2.67	2.40	2.61	2.53	2.18	2.92
3.6	1.68	1.75	1.95	1.85	1.76	1.67	1.31	1.72
3.4	1.57	1.87	1.91	2.02	1.90	1.92	1.30	1.86
3.2	1.46	1.66	1.66	1.76	1.80	1.90	1.23	1.79
3.0	1.28	1.68	1.50	1.79	1.74	2.03	1.16	1.73
2.8	1.17	1.64	1.25	1.61	1.48	1.85	0.95	1.43
2.6	1.04	1.44	1.11	1.44	1.33	1.67	0.94	1.31
2.4	0.97	1.44	1.04	1.46	1.27	1.67	0.86	1.17
2.2	0.79	1.34	0.78	1.28	0.89	1.36	0.89	1.07

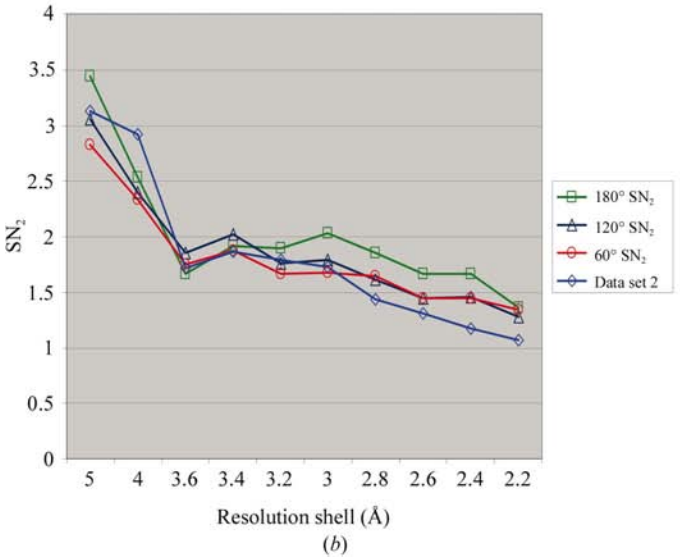
3.2. R_{as} *versus* the ability of the data to produce the anomalous scatterering substructure

This test was designed to evaluate the usefulness of R_{as} as an index for judging the data quality with respect to its usefulness in producing the anomalous scatterering substructure, the first step in the SAS or MAD structure determination.

The correlation coefficients (CC/all and CC/weak) produced by *SHELXD* were used for the analysis. The correlation coefficients are associated with the sets of heavy-atom sites found by the program and provide statistical estimates of the quality of the heavy-atom solution and by inference the strength of the anomalous scattering signal in the data. Our previous experience with *SHELXD* indicates that the correlation coefficients are reliable indicators for the certainty of the heavy-atom site(s) located by the program and are useful indicators for the strength of the anomalous scattering signal in the data, *i.e.* the larger the correlation coefficients the higher the certainty of the heavy-atom site and the larger the signal in the data.

SHELXD was used with 200 trials per run to determine the sulfur anomalous substructure (three disulfides) with all the above four data sets at ten different resolutions. The correlation coefficients of the best solutions for the given resolutions and data sets are listed in Table 1 and plotted in Fig. 2. As can be seen from Fig. 2, the R_{as} predictions are in good agreement with the *SHELXD* correlation coefficients, including the comparison of data 2 with the 60° data from crystal 1 as discussed previously. Thus, R_{as} can serve as an indicator of the ability of a data set to produce the anomalous scatterering substructure, a critical step in successful SAS or MAD structure determination.





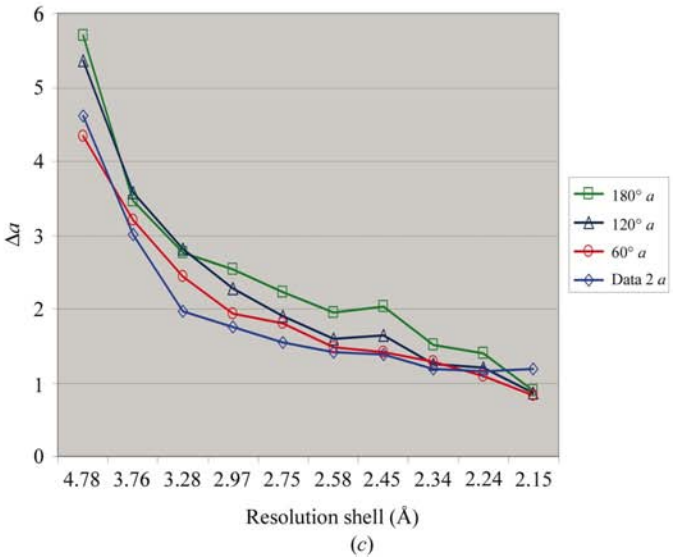


Figure 3
Plots of in-shell signal-to-noise ratios from *XPREP* and Δa values *versus* resolution shell. Red, dark blue, green and light blue lines represent 60, 120 and 180° data sets and data set 2, respectively. (*a*) SN_1, a signal-to-noise ratio based on input σ values from *XPREP*, *versus* resolution shell. (*b*) SN_2, a signal-to-noise ratio based on variances of F_+ and F_- from *XPREP*, *versus* resolution shell. (*c*) Δa *versus* resolution shell.

3.3. R_{as} *versus* the anomalous signal indicators used in *XPREP*

This test was designed to compare R_{as} with the two indicators of anomalous scattering signal-to-noise ratio produced by *XPREP* (Sheldrick, 2000). For the comparison, the unmerged data after error correction and scaling by *3DSCALE* were written out for each of the above four data sets and analyzed by *XPREP*. *XPREP* produces two indices, denoted here as SN_1 and SN_2, corresponding to input σ and variance of the Bijvoet differences, respectively. The resulting SN_1 and SN_2 values from the analysis are listed in Table 2 and plotted in Figs. 3(*a*) (SN_1 *versus* resolution shell) and 3(*b*) (SN_2 *versus* resolution shell). The SN_1 plot (Fig. 3*a*) shows that data 2 has a significantly lower signal-to-noise ratio than the other three data sets. This observation does not agree with R_{as} (Fig. 1*b*) and the correlation coefficients from *SHELXD* (Fig. 2). In general, the SN_1 plot (Fig. 3*a*) looks very similar to the Δa *versus* resolution shell plot (Fig. 3*c*). Although the exact definition of SN_1 was not found in the literature, its apparent similarity to Δa suggests that it is closely related to Δa. If this is true, SN_1 is not a good indication of signal-to-noise since Δa is a measure of signal plus noise as discussed earlier.

In addition, SN_2 (Table 2 and Fig. 3*b*), in our opinion also did not adequately predict the anomalous signal-to-noise ratio from our test data. Fig. 3(*b*), for example, suggests that the improvement of anomalous signal in going from the 60° to the 120° data set (100% increase in redundancy) is less significant than 50% increase in redundancy produced in going from the 120° to the 180° data set. This is not in agreement with either the R_{as} analysis or the *SHELXD* results (Fig. 2). Also, SN_2 predicts that anomalous signal in data 2 is as good as or better than the 120° data to a resolution of 3.0 Å. Again, this runs contrary to both the R_{as} analysis and *SHELXD* results, which indicate that the anomalous signal in data 2 is similar to that observed for the 60° data. Thus, both SN_1 and SN_2, the only other two quantities that have been suggested as indicators of the anomalous signal-to-noise ratio in data, failed to correlate well with the *SHELXD* results from our test data. Therefore, we believe that the proposed R_{as} index will serve to complement existing indices in data-reduction programs for detecting and monitoring the anomalous scattering signal-to-noise in the data.

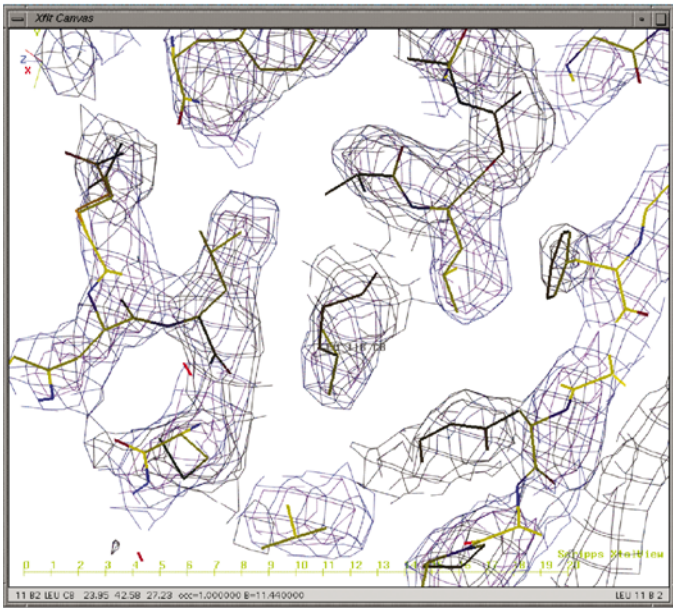


Figure 4
A representative section of electron density shown with *XTALVIEW* (McRee, 1992), centered on Leu11B of the Zn-free insulin structure, together with the refined insulin coordinates from the Protein Data Bank (Bernstein *et al.*, 1977) entry 9ins. The map is calculated using phases derived by *ISAS*2001 from the 60° data set to 2.58 Å resolution.

3.4. R_{as} value *versus* production of an interpretable electron-density map

In order to show at what minimum R_{as} level an interpretable electron-density map can be obtained by SAS phasing, we used the 60° data set to 2.58 Å as a test. *SHELXD* was used to find the anomalous scattering substructure (in this case three disulfide superatoms). *ISAS*2001 (Wang, 1985) was then used to resolve the SAS phase ambiguity and to produce the experimental phases that were used to produce the interpretable electron-density map shown in Fig. 4. This result confirms the validity of the R_{as} approach and suggests that an R_{as} value of 1.5 is sufficient for the production of an interpretable map from the test data. It is certainly possible that R_{as} values lower than 1.5 are sufficient to produce interpretable maps. However, more tests using a variety of data will be needed to confirm this. Tests with other applications of R_{as} will also be continued.

4. Discussion

The reliability of the measured anomalous signal is closely related to the noise level in the measurement of the data, especially when the signal is very weak, as in the case of P-, S- or Cl-containing native protein crystals. The strength of the current approach, as described in §2, is that we merge and group symmetry-equivalent reflections into centric and non-centric subsets. Those merged reflections in the centric subset that are equal in theory in the past have been merged into the same unique reflections by most data-processing programs, but we treat them as separate reflections and preserve them as 'internal probes' for noise level. To our best knowledge, this approach has not been applied in addressing the signal-to-noise ratio in anomalous scattering data. Here, we compare the statistics for the centric and non-centric subsets and this is the basis for our proposed R_{as}.

Unlike χ^2, $I/\sigma(I)$, R_{meas} or similar measures, R_{as} directly deals with Bijvoet differences (the anomalous signal) and unlike R_{anom} or the δR plot it evaluates both signal and noise. Test results show that predictions based on R_{as} are in better agreement with the trends observed in *SHELXD* correlation coefficients than those predicted using the *XPREP* indicators SN_1, SN_2 or anomalous differences (Δa), the only other quantitative indicators of anomalous signal-to-noise ratio currently in use. Test results also indicate that R_{as} can be used

to judge the 'useful' resolution of a data set for producing interpretable electron-density maps.

Since the proposed R_{as} index provides an effective way to quantitatively evaluate the anomalous signal-to-noise ratio in X-ray diffraction experiments, we propose that R_{as} be used in data-processing programs to provide an early indication of the quality of the anomalous scattering data, in particular as a quantitative measurement of the signal-to-noise ratio in the data.

A more reliable evaluation of data quality and data sufficiency will also help in decision-making. Thus, R_{as}-assisted decision-making could be very useful to the automation of data collection and structural analysis that is becoming more and more important in the era of structural genomics. For example, there is always the question whether the data quality will be improved if more observations are available owing to increased redundancy. In reality, collecting more and more data from the same crystal may not improve the data quality as expected intuitively because of the possibility of radiation damage. A reliable signal and noise evaluation index such as R_{as} described here could help to answer this and other related questions that arise during data acquisition. It is totally feasible that by using a robot with multiple crystals one could define a preset signal-to-noise target level in the data and using R_{as} to monitor data collection determine when a fresh crystal is needed and when this level has been reached. If this preset signal-to-noise level represents the minimum value required for solving a structure, then one will have a set of data that will most likely be accurate enough to solve the target structure.

The authors thank Drs Gary Newton and Wayne Hendrickson for comments and discussions, and Dr James Liu and Ms Lei Huang for providing the data used for the tests. The work was supported in part by National Institutes of Health Protein Structural Initiative Grant GM62407 from the National Institute of General Medical Sciences, University of Georgia Research Foundation and Georgia Research Alliance.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.

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

Dauter, Z., Dauter, M. & Dodson, E. (2002). *Acta Cryst.* D**58**, 494–506.

Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.

Dickerson, R. E., Kendrew, J. C. & Standberg, B. E. (1961). *Acta Cryst.* **14**, 1188–1195.

Diederichs, K. & Karplus, P. A. (1997). *Nature Struct. Biol.* **4**, 269–275.

Drenth, J. (1994). *Principles of Protein X-ray Crystallography*. New York: Springer–Verlag.

Evans, P. R. (1993). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by L. Sawyer, N. Isaacs & S. Bailey, pp. 114–122. Warrington: Daresbury Laboratory.

Fu, Z. Q., Pressprich, M., Sparks, R., Foundling, S. & Phillips, J. (2000). Am. Crystallogr. Assoc. Annu. Meet., St Paul, Minnesota, USA. Abstract P066.

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

Hendrickson, W. A. (1985). *Trans. Am. Crystallogr. Assoc.* **21**, 11.

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

Howell, L. & Smith, D. (1992). *J. Appl. Cryst.* **25**, 81–86.

Kleywegt, G. J. (2000). *Acta Cryst.* D**56**, 249–265.

Li, S., Finley, J., Liu, Z.-J., Qiu, S. H., Chen, H., Luan, C. H., Carson, M., Tsao, J., Johnson, D., Lin, G., Zhao, J., Thomas, W., Nagy, L. A., Sha, B., DeLucas, L. J., Wang, B.-C. & Luo, M. (2002). *J. Biol. Chem.* **277**, 48596–48601.

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

Ladd, M. F. C. & Palmer, R. A. (1994). *Structure Determination by X-ray Crystallography*, 3rd ed. New York: Plenum Press.

McRee, D. E. (1992) *J. Mol. Graph.* **10**, 44–46.

McRee, D. E. (1993). *Practical Protein Crystallography*. San Diego: Academic Press.

Matthews, B. W. (1966). *Acta Cryst.* **20**, 82–68.

North, A. C. T. (1965). *Acta Cryst.* **18**, 212–216.

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

Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* D**59**, 1145–1153.

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

Sheldrick, G. M. (2000). *XPREP* Version 6.10. Bruker AXS Inc., Madison, Wisconsin, USA.

Sheldrick, G. M. & Schneider, T. R. (2001). *Methods in Macromolecular Crystallography*, edited by D. Turk & L. Johnson, pp. 72–81. Amsterdam: IOS Press.

Shen, Q., Wang, J. & Ealick, S. E. (2003). *Acta Cryst.* A**59**, 371–373.

Stout, G. H. & Jensen, L. H. (1968). *X-ray Structure Determination. A Practical Guide.* London: Macmillan.

Wang, B.-C. (1982). Am. Crystallogr. Assoc. Summer Meet., San Diego, California, USA. Abstract P066.

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

Weiss, M. S. (2001). *J. Appl. Cryst.* **34**, 130–135.

Weiss, M. S. & Hilgenfeld, R. (1997). *J. Appl. Cryst.* **30**, 203–205.

Weiss, M. S., Metzner, H. J. & Hilgenfeld, R. (1998). *FEBS Lett.* **423**, 291–296.

Wu, C.-K., Burden, A., Rose, J. P., Ferrara, J., Dailey, H. A. & Wang, B.-C. (1999). *Acta Cryst.* A**55** (Suppl.), 255.

Wu, C.-K., Dailey, H. A., Rose, J. P., Burden, A., Sellers, V. M. & Wang, B.-C. (2001). *Nature Struct. Biol.* **8**, 156–160.