

# Three-dimensional model-free experimental error correction of protein crystal diffraction data with free- $R$ test

**Zheng-Qing Fu**†

Department of Biochemistry and Molecular Biology, The University Of Georgia, Athens, GA 30602, USA

† Previous address: Bruker AXS Inc., 5465 East Cheryl Parkway, Madison, WI 53711, USA.

Correspondence e-mail: fuzq@uga.edu

Received 9 August 2005  
Accepted 18 October 2005

Experimental error correction and scaling is the last step in X-ray diffraction data processing. It is also critical in obtaining good-quality data. In this study, an algorithm is proposed to more generally and efficiently correct experimental error in X-ray crystal diffraction data. With this algorithm, the experimental error is represented by a symbolic three-dimensional function  $C(\xi, \eta, t)$  in the detecting space. Here,  $(\xi, \eta)$  are the coordinates of the diffraction spots on the image and  $t$  represents the data-collection time. While the theoretical form of  $C(\xi, \eta, t)$  is not known, it will be determined from the data through computer-aided analysis. The free  $R_{\text{merge}}$  is introduced to check the validity of the solution. The three-dimensional symbolic function does not carry any assumptions and thus can generally account for experimental errors in various X-ray crystal diffraction experiments that are normally too complicated to be described by any fixed formula. Tests will be given to compare the results from different algorithms.

## 1. Introduction

Data quality is the key to successful solution of protein structures by X-ray crystallography. At present, MAD (Phillips & Hodgson, 1980; Karle, 1980; Hendrickson, 1991) is the most commonly used phasing method to solve novel protein structures. It utilizes the multi-wavelength anomalous scattering signals from heavy atoms that naturally exist in or are purposely introduced into the protein. The anomalous scattering is very weak and only accounts for less than 5% of the diffraction intensity in most cases. This kind of weak signal can be easily lost during data processing if experimental error is not accurately treated and corrected, especially for crystals with moderate and low diffraction quality and ability. For the SAD method using single-wavelength anomalous signals from light atoms such as sulfurs in native proteins (Hendrickson & Teeter, 1981; Wang, 1985; Dauter *et al.*, 1999) high data quality is even more critical because the anomalous signal only accounts for less than 1% of the diffraction intensity.

The diffraction intensity of biological macromolecular crystals decreases with cumulative exposure time owing to radiation damage during data collection. To put the intensities on a common scale, in the 1960s Hamilton, Rollett and Sparks introduced a scaling algorithm that applies a pair of constant factors ( $S$ ,  $B$ ) to all observed reflections on each image (Hamilton *et al.*, 1965). Here,  $S$  is the scaling factor and  $B$  is the isotropic factor intent on correcting resolution-dependent errors from crystal decay and absorption. The set of refinable parameters  $\{S(i), B(i)|i = 1, \text{number of frames}\}$  is then determined by minimizing the target function (intensity differences

between the symmetry-equivalent reflections). While this simple algorithm (denoted hereafter as ‘isotropic scaling’ or ISOS) works adequately for scaling the data, it does not effectively correct experimental error. Programs implementing the ISOS algorithm include *SCALEPACK/HKL2000* (Otwinowski & Minor, 1997) and *SCALA* (with ‘batch’ scaling option; Evans, 2005). To deal with the anisotropy, the isotropic scaling was modified by replacing the isotropic *B* factor with the empirical spherical harmonic functions (Blessing, 1995), which is denoted hereafter as ‘empirical spherical harmonic scaling’ or ESHS. Empirical spherical harmonic scaling has been implemented by the programs *SCALEPACK*, *X-GEN* (Howard, 2000), *SADABS* (Sheldrick, 2000), *SCALA* (with ‘secondary’ scaling option) and *d\*TREK* (Pflugrath, 1999). Recently, Otwinowski proposed the use of the exponential function instead of the spherical harmonics (Otwinowski *et al.*, 2003), which has not yet been seen to be used. Another significant modification of isotropic scaling was made by applying a scale factor to each of a fixed number of points on the detector surface (Kabsch, 1988). The scale factor for a reflection is evaluated as the weighted average of all these factors. To make the scale factor position-dependent and vary with reflections in the detecting space, it has to employ a pre-defined weighting scheme. This algorithm is denoted hereafter as ‘detector scaling’ or DETS. It overcomes, to some extent, the insufficiency of other algorithms described above. One limitation of this algorithm is that the explicit pre-defined weighting function may not fit various kinds of diffraction-experiment errors (Fu *et al.*, 2000, 2002). In addition, it does not provide a method to check the validity of the error model. Programs using ‘detector scaling’ include *XDS* (Kabsch, 1988) and *SCALA* (with ‘detector’ scaling option). Owing to these limitations, DETS has not shown any advantage in general use (Evans, 2005), except for scaling data from soft X-ray diffraction (Mueller-Dieckmann *et al.*, 2004). These algorithms used in the current existing programs have been derived with some modification from ‘isotropic scaling’, which emphasizes the scaling instead of experimental error correction. They all employ a modeling approach that tries to simulate the experimental error by some explicitly pre-defined functions. The explicit model described by a pre-defined specific function may work appropriately for certain cases. However, the experimental error is far more complicated owing to the large number of variables in X-ray diffraction experiments on biological macromolecular crystals.

To overcome the insufficiency of any pre-defined error model, a model-free algorithm is proposed in this study to more generally and efficiently correct experimental error. In this algorithm, the experimental error is represented by a symbolic function  $C(\xi, \eta, t)$  in the detecting space, where  $(\xi, \eta)$  are the coordinates of a diffraction spot on the image and  $t$  represents the data-collection time. The undefined error function  $C(\xi, \eta, t)$  will be determined from the data through computer-aided analysis. The free  $R_{\text{merge}}$  will be introduced for a validity test of the determination of  $C(\xi, \eta, t)$ . This model-free algorithm is denoted hereafter as ‘three-dimensional model-free error correction and scaling’ or 3DCS.

## 2. Method

In X-ray crystal diffraction experiments, reflection intensities are not only determined by the crystal itself, but are also affected by many other intrinsic and non-intrinsic errors. For macromolecular crystals, these could include absorption, anisotropic mosaicity, crystal defects, radiation damage, X-ray beam instability, detector defects and other systematic errors. Because there are no generally applicable formulae for all these errors, the experimental error has to be corrected through computer-aided data processing. The error correction at this step greatly affects the quality of the reduced data, which is critical for the solution of structures, especially for weak anomalous data. The explicit error models used by the current existing data-processing programs may make physical sense and may be adequate in most cases. However, their insufficiency and lack of generality may lead to inefficient error description and correction for various diffraction experiments. A generally applicable approach is needed to more accurately correct the experimental error, especially for data of moderate or low quality.

Most of the experimental errors in X-ray diffraction are too complicated to be theoretically represented by generally applicable formulae. For example, among the influences on absorption are diffraction geometry, the nature of the molecules in the crystal, solvent content, crystal shape, crystal orientation, mounting scheme, X-ray wavelength and air in the path. The complexity of these contributing sources for biological macromolecular crystals will make it impossible to derive a general formula which is appropriate for various kinds of crystals and different diffraction experimental settings. However, no matter how complicated it is, a symbolic three-dimensional function  $A(\xi, \eta, t)$  can be assigned to represent the absorption correction without any intuitive or theoretical assumption about the actual form or detail of the function, where  $(\xi, \eta)$  are the coordinates of a diffraction spot on the image and  $t$  represents the data-collection time.

Similarly, for each of other experimental errors that cannot be theoretically formulated, we can assign a symbolic function. We also know for sure that no matter how complicated and what the sources of the experimental errors are, the total effect of these errors is also three-dimensional and thus can be represented as a symbolic function  $C(\xi, \eta, t)$ ,

$$C(\xi, \eta, t) = A(\xi, \eta, t) * R(\xi, \eta, t) * X(\xi, \eta, t) * D(\xi, \eta, t) * S(\xi, \eta, t) * \dots, \quad (1)$$

where  $A(\xi, \eta, t)$  represents absorption,  $R(\xi, \eta, t)$  radiation damage,  $X(\xi, \eta, t)$  X-ray source defects,  $D(\xi, \eta, t)$  detector defects and so on. The reflections can be corrected and scaled by

$$I_j = C(\xi, \eta, t)I_j^0, \quad (2)$$

$$\sigma_j = C(\xi, \eta, t)\sigma_j^0 + I_j^0\sigma_c. \quad (3)$$

Here  $(I_j^0, \sigma_j^0)$  and  $(I_j, \sigma_j)$  are the intensity and standard deviation of the  $j$ th reflection before and after applying error correction. Although the formula of this three-dimensional

error function  $C(\xi, \eta, t)$  is not known at the start, it can be parameterized and determined from the data in a least-squares procedure that minimizes the target function  $\chi^2$  (4). The three-dimensional error function  $C(\xi, \eta, t)$  thus determined will carry the characteristics of the experimental error and be unique for the data set.

To parameterize the error function, all images from one continuous batch can be viewed as a stack of consecutive frames equally separated in  $t$ , which can be divided into sectors and further divided into small angular and radial bins defined relatively to the detector surface. In other words, the  $(\xi, \eta, t)$  space is divided into  $N_t \times N_a \times N_r$  blocks.  $N_t, N_a$  and  $N_r$  are the number of sectors, number of angular bins and number of radial bins, respectively. A refinable parameter is assigned to each dividing edge of these blocks. Thus, the total number of parameters to be solved in the least-squares procedure is  $N_t \times N_a \times N_r$ . The correction factor for each reflection is evaluated from the parameters on the surrounding edges by a linear interpolation of the surrounding parameters. The entire data set is composed of one or more such stacks. The parameterization of the error function is important in the implementation of the 3DCS algorithm. The finer the division of  $(\xi, \eta, t)$  space, the better the description of experimental error. However, the finer division will lead to a larger number of parameters to be determined, which will reduce the data-to-parameter ratio in the least-squares procedure. To validate the parameterization and solution of the error function, a subset (by default 5%) of the unique reflections is randomly selected to calculate the free  $R_{\text{merge}}$ . These reflections are not used in determining the error-correction function, but will still be scaled and output. The remaining reflections are used in a robust least-squares refinement procedure to determine  $C(\xi, \eta, t)$  by minimizing the  $\chi^2$  function (Hamilton *et al.*, 1965)

$$\chi^2 = \sum_{i=1}^N \sum_{j=1}^{N_i} (I_j - I_i)^2 / \sigma_j^2, \quad (4)$$

$$I_i = \sum_{j=1}^{N_i} W_j I_j / \sum_{j=1}^{N_i} W_j, \quad (5)$$

$$W_j = \frac{1}{(E_1 \sigma_j^2 + E_2 I_j^2)}, \quad (6)$$

where  $N$  is the number of unique reflections,  $N_i$  is the number of equivalents for the  $i$ th unique reflection and  $(I_j, \sigma_j^2)$  are the intensity and intensity variance of the  $j$ th observed reflection after correction and scaling by  $C(\xi, \eta, t)$ .

At the start,  $E_1$  and  $E_2$  are initialized as 1.0 and 0.0, respectively. The error function  $C(\xi, \eta, t)$  is then solved by least-squares minimization.  $E_1$  and  $E_2$  are automatically determined by another least-squares refinement to make the following function approach 1.0 ( $\chi^2$  analysis),

$$\langle \chi^2(E_1, E_2) \rangle = \frac{1}{N} \sum_{i=1}^N \frac{\sum_{j=1}^{N_i} (I_j - I_i)^2}{\sum_{j=1}^{N_i} (E_1 \sigma_j^2 + E_2 I_j^2)}. \quad (7)$$

The new  $C(\xi, \eta, t)$ ,  $E_1$  and  $E_2$  values are then put into the next cycle of minimization. As the minimization converges, the  $(I_i, \sigma_i)$  and/or  $[I(+)_i, \sigma(+)_i, I(-)_i, \sigma(-)_i]$  (if anomalous signals are requested) of each unique reflection will be evaluated. Finally, the set of unique reflections or the whole set of all reflections (if requested) after experimental error correction and scaling will be saved for the structure-solution process.

### 3. Tests

The ‘three-dimensional model-free error correction and scaling’ method was implemented into the computer program *PROSCALE* (Fu *et al.*, 2000), which is part of the *PROTEUM* software suite (Bruker-AXS, Inc.) running on MS Windows. Recently, another program *3DSCALE* was developed using standard C/C++ to support multiple platforms including Unix and Linux. In both *PROSCALE* and *3DSCALE*, the three-dimensional error function is automatically parameterized as described in §2. By default,  $N_b, N_a$  and  $N_r$  are set to  $8/\text{OscAng}$ , 5 and 2 (where *OscAng* is the oscillation angle in degrees), respectively. The defaults were obtained through program training against a variety of data sets of different quality and work well. However, the user can easily select a different parameterization scheme if desired. *PROSCALE* has been used by different groups in error correction and scaling of integrated data to solve structures (such as Schubot *et al.*, 2004; Sims *et al.*, 2003; Vanhooke *et al.*, 2004; Wu *et al.*, 2005; Lee *et al.*, 2004; Chen *et al.*, 2005; Ren *et al.*, 2005; Thoden & Holden, 2005). Listed in the next section are test results on some data sets comparing different scaling methods.

#### 3.1. Test data

**3.1.1. Zn-free insulin.** SAD data were collected from a Zn-free insulin crystal (space group  $I2_13$ ,  $a = 78.089$  Å) in the resolution range 39.0–2.15 Å using a Bruker Proteum-R CCD (a.k.a. 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 Bruker’s *PROTEUM* data-reduction package. The anomalous scattering from S atoms was used to solve the structure.

**3.1.2. C-terminal domain of a corrinoid-binding protein (CBP).** The protein contains 125 amino acids, with an unknown number of Co site(s). SAD data were collected from one crystal (space group  $P2_12_12$ ,  $a = 55.52$ ,  $b = 62.65$ ,  $c = 34.45$  Å) to a resolution of 2.30 Å using a Bruker Proteum-R CCD detector mounted on a Rigaku FRD generator using Cu  $K\alpha$  X-rays with Rigaku/MSC HiRes<sup>2</sup> optics. A total of 1200 0.3° oscillation images were recorded using an exposure time of 30 s. The intensities were indexed and integrated using the Bruker *PROTEUM* data-reduction package.

**3.1.3. Pfu631545.** The Se-containing protein contains 133 amino acids (including an N-terminal 6-His tag) with one Se site. MAD data were collected (space group  $P2_1$ ,  $a = 36.35$ ,  $b = 61.09$ ,  $c = 51.58$  Å,  $\beta = 97.62^\circ$ ) to about 2.0 Å resolution on

**Table 1**

Results in solving the structures from test data.

'Method' gives the algorithms used for experimental error correction and scaling (see §3.2 for details). The values in parentheses are for the highest resolution shell: 2.24–2.15 Å for insulin, 2.45–2.30 Å for CBP and 2.09–2.0 Å for Pfu631545.  $N_{\text{tr}}$  and TR% are the number and percentage of amino-acid residues automatically traced by *RESOLVE* or *ARP/wARP* during the structure-solution process.

Method	$R_{\text{merge}}$	Complete-ness (%)	Redundancy	$I/\sigma(I)$	$N_{\text{tr}}$	TR%
Insulin						
ISOS	0.057 (0.103)	96.9 (83.6)	15.8 (3.8)	50.4 (7.7)	48	94.1
ESHS	0.035 (0.077)	96.7 (81.3)	15.8 (3.7)	57.4 (13.6)	49	96.1
DETS	0.034 (0.089)	96.9 (83.6)	15.8 (3.8)	82.1 (12.6)	49	96.1
3DCS	0.034 (0.074)	96.9 (82.0)	15.8 (3.8)	60.6 (11.5)	49	96.1
CBP						
ISOS	0.053 (0.098)	99.2 (92.1)	7.7 (6.3)	28.4 (15.3)	95	76.0
ESHS	0.046 (0.067)	99.0 (92.0)	7.5 (6.2)	33.5 (17.4)	88	70.4
DETS	0.049 (0.061)	99.2 (92.1)	7.7 (6.3)	37.7 (19.8)	93	74.4
3DCS	0.049 (0.063)	99.2 (92.1)	7.6 (6.4)	34.0 (16.8)	102	81.6
Pfu631545						
ISOS	0.084 (0.281)	99.2 (96.0)	3.5 (3.2)	17.9 (4.1)	126	47.4
ESHS	0.072 (0.276)	99.3 (96.1)	3.6 (3.2)	17.0 (3.2)	147	55.3
DETS	0.077 (0.275)	99.2 (96.0)	3.5 (3.2)	18.5 (4.2)	141	53.0
3DCS	0.070 (0.254)	99.2 (96.1)	3.5 (3.2)	12.2 (3.3)	160	60.2

a MAR CCD 225 at APS SERCAT beamline 22-ID at three wavelengths: 0.97828, 0.97941 and 0.98086 Å. However, phasing with the 180° SAD data collected at 0.97941 Å turned out to give the best initial model. The SAD data, integrated using *HKL2000*, will be used in the test.

### 3.2. Results

Each of the above integrated data sets was scaled using different algorithms and reduced to separate unique data sets. The error-correction and scaling algorithms tested include ISOS, ESHS, DETS and 3DCS. ISOS was performed by *SCALA* 'batch mode' with 'bfactor on'. ESHS was performed by *SCALEPACK* on Pfu631545 data and by *SADABS* on insulin and CBP data. The default up to fourth-order and sixth-order harmonics were used with *SCALEPACK* and *SADABS*, respectively. DETS scaling was performed by *SCALA* 'detector 3' mode with 'rotation spacing 5'. 3DCS was performed by *3DSCALE* with defaults described above.  $E_1$  and  $E_2$  in (6) were manually adjusted to make  $\chi^2$  approach 1.0 in *SCALEPACK* scaling. This was performed automatically by other programs mentioned above. Finally, each unique data set obtained was used to solve the structure using *SGXPRO* (Fu *et al.*, 2005), which searches both program and parameter space to try building the structural model using *SHELXD* (Schneider & Sheldrick, 2002), *ISAS* (Wang, 1985), *SOLVE* (Terwilliger & Berendzen, 1999), *RESOLVE* (Terwilliger, 2000) and *ARP/wARP* (Perrakis *et al.*, 1999) for heavy-atom site searching, handedness test, phasing and autotracing, respectively. The results are listed in Table 1.

As expected, the insulin data are of high quality. There is no significant difference between the unique data sets from the four different scaling algorithms tested (see Table 1). All four data sets produced almost complete initial structure models with over 94.1% of the amino-acid residues automatically

traced. The CBP data has moderate quality. The reduced data sets from the four different algorithms all produced traceable electron-density maps. However, the three-dimensional model-free error correction and scaling generated a better electron-density map, leading to a more complete initial structure model autotraced by *RESOLVE*. The model has been used to complete the structure which was refined against a higher resolution data set (see PDB entry 1y80). In the Pfu631545 case, which has low-quality data, 3DCS scaling led to a traceable electron-density map with 60.2% of the amino-acid residues automatically traced. The structure has been solved from this initial model with help of manual fitting (see PDB entry 1ztd). Using the unique data sets from the ISOS, ESHS and DETS algorithms, autotracing built 47.4, 55.3 and 53.0% of the residues, respectively. However, the connectivity is lower than that from 3DCS. In other words, only small fragments were built, which is one indication of a poor structure model and electron-density map. In fact, the electron-density map from DETS is hard to trace and that from ISOS is not traceable at all. Fig. 1, drawn using *XTALVIEW* (McRee, 1992), shows the electron-density maps with the autotraced structure models built by using the unique data sets from the four tested algorithms.

### 4. Discussion

The 3DCS algorithm emphasizes the experimental error correction during scaling by using a general symbolic three-dimensional error function  $C(\xi, \eta, t)$  that is determined from the data through computer-aided minimization of the target function. Compared with modeling approaches, 3DCS does not make any theoretical or intuitive assumption about the variation of experimental error in the detecting space. The  $C(\xi, \eta, t)$  thus determined will uniquely represent and correct the experimental error in the data. It overcomes the insufficiency and lack of generality which may arise from the imprecise description of experimental error by explicitly pre-defined error model such as the spherical harmonic functions. In two of the test cases with moderate or low data quality, 3DCS scaling produced data that led to better electron-density maps and initial structure models, which have made the structure-solution processes easier. The free  $R_{\text{merge}}$  values are 0.035, 0.051 and 0.071 for insulin, CBP and Pfu631545 data, respectively. They agree well with the  $R_{\text{merge}}$  values calculated from the working sets, which are 0.034, 0.049 and 0.070, respectively (see Table 1). This good agreement suggests that the three-dimensional error functions determined in the least-squares procedure are valid in correcting experimental errors of the diffraction data. The 3DCS scaling provides a robust approach generally applicable to correct experimental errors in various X-ray diffraction data by using a model-free algorithm.

As in any least-squares procedure, a finer detecting-space sampling is better to describe the experimental error. However, finer division will reduce the data-to-parameter ratio, which may affect the accuracy of the process. The default sampling obtained through program training against a variety

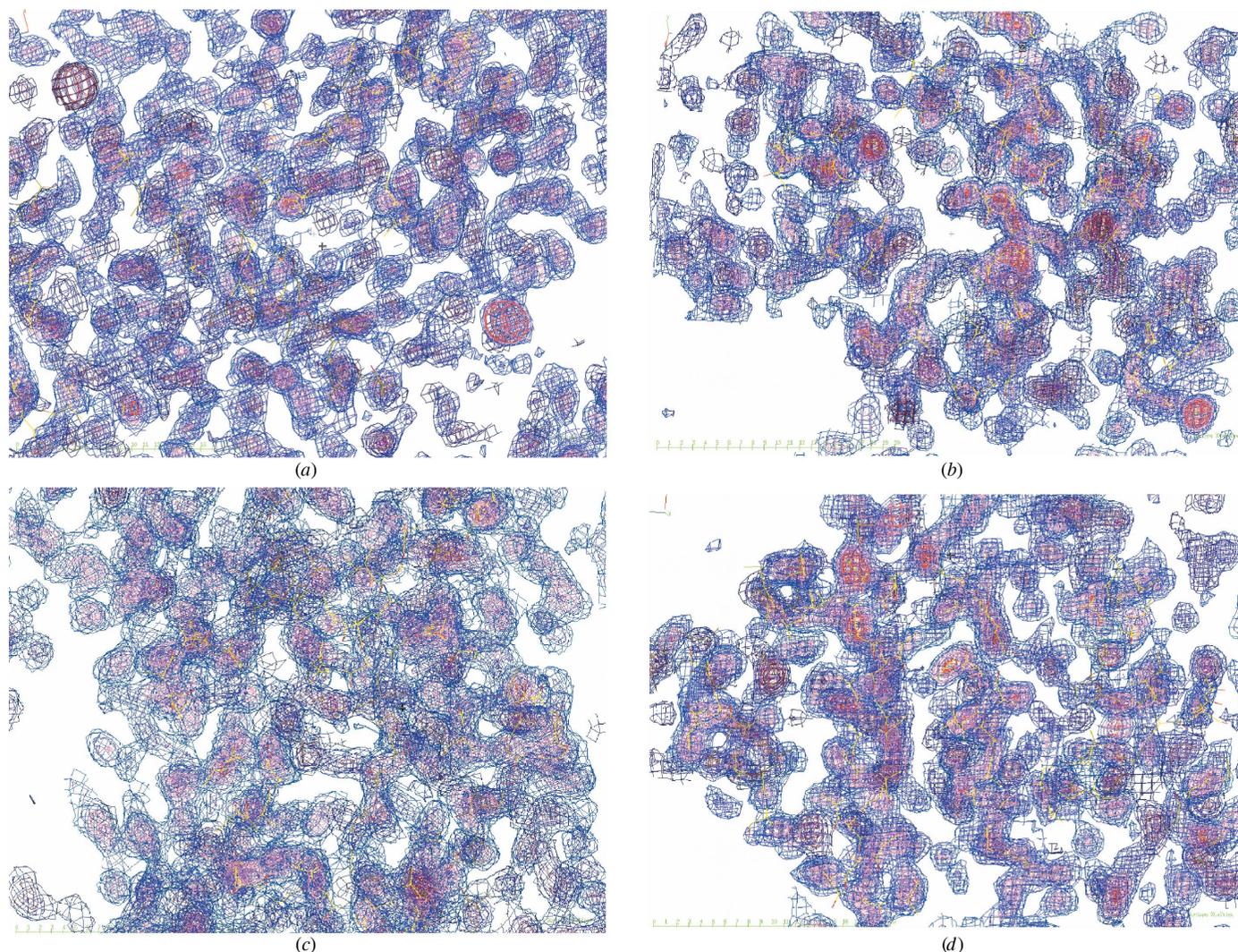
of data sets works well. Sampling variation around the default did not show a great impact on the final results in the three test cases (see Table 2). Careful design of the parameterization scheme can also enhance the data-to-parameter ratio, which makes the least-squares procedure more stable. For example,

the total numbers of parameters with the default sampling in 3DCS scaling are 220, 450 and 210 for insulin, CBP and Pfu631545 data, respectively. In ISOS scaling with two parameters  $\{S, B\}$  for each frame, these numbers were 1800, 2400 and 360, respectively. ESHS may also need more parameters if

**Table 2**  
Effects of detecting-space sampling.

$N_s$ ,  $N_a$  and  $N_r$  represent the number of sectors, number of angular bins and number of radial bins, respectively.  $N$  is the total number of parameters used in determining  $C(\xi, \eta, t)$ .  $N_{tr}$  and TR% are the number and percentage of amino-acid residues automatically traced by *RESOLVE* or *ARP/wARP*.

Insulin						CBP						Pfu631545					
$N_t$	$N_a$	$N_r$	$N$	$N_{tr}$	TR%	$N_t$	$N_a$	$N_r$	$N$	$N_{tr}$	TR%	$N_t$	$N_a$	$N_r$	$N$	$N_{tr}$	TR%
11	5	2	110	48	94.1	22	5	2	220	102	81.6	10	5	2	100	139	52.3
22	3	2	132	49	96.1	45	3	2	270	99	79.2	21	3	2	126	138	51.9
22	5	2	220	49	96.1	45	5	2	450	102	81.6	21	5	2	210	160	60.2
22	5	3	330	50	98.0	45	5	3	675	102	81.6	21	5	3	315	159	59.8
22	5	4	440	49	96.1	45	5	4	900	99	79.2	21	5	4	420	165	62.0
22	10	2	440	49	96.1	45	10	2	900	107	85.6	21	10	2	420	147	55.3
44	5	2	440	50	98.0	90	5	2	900	100	80.0	42	5	2	420	156	58.6
89	5	2	890	49	96.1	180	5	2	1800	101	80.8	85	5	2	850	145	54.5



**Figure 1**  
The electron-density maps and autotraced models of Pfu631545 from the structure-solution process using the unique data sets scaled by the four different algorithms. (a) Data from ISOS scaling. (b) Data from ESHS scaling. (c) Data from DETS scaling. (d) Data from 3DCS scaling.

frame grouping, as implemented in *SCALA* and *3DSCALE*, is not used.

Finally, the current implementation of 3DCS uses the traditional least-squares target (4) to determine the error function  $C(\xi, \eta, t)$ . A maximum-likelihood target may be used instead if proved to be beneficial.

The author thanks Dr Bi-Cheng Wang for encouraging discussions during this study and also thanks Dr John Rose and researchers at the Southeast Collaboratory for Structural Genomics for use of test data.

## References

- Blessing, R. H. (1995). *Acta Cryst.* **A51**, 33–38.
- Chen, Q., Liang, Y., Su, X., Gu, X., Zheng, X. & Luo, M. (2005). *J. Mol. Biol.* **348**, 1199–1210.
- Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.
- Evans, P. R. (2005). In the press.
- Fu, Z.-Q., Pressprich, M., Sparks, R. A., Foundling, S. & Phillips, S. (2000). Am. Crystallogr. Assoc. Annu. Meet., St Paul, Minnesota, USA. Abstract P066.
- Fu, Z.-Q., Rose, J. P. & Wang, B.-C. (2002). *Acta Cryst.* **A58**, C83.
- Fu, Z.-Q., Rose, J. P. & Wang, B.-C. (2005). *Acta Cryst.* **D61**, 951–959.
- Hamilton, W. C., Rollett, J. S. & Sparks, R. A. (1965). *Acta Cryst.* **18**, 129–130.
- Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
- Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
- Howard, A. J. (2000). In *Crystallographic Computing 7*, edited by K. D. Watenpaugh. Oxford University Press.
- Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.
- Karle, J. (1980). *Int. J. Quant. Chem.* **7**, 357–367.
- Lee, B. I., Kim, K. H., Park, S. J., Eom, S. H., Song, H. K. & Suh, S. W. (2004). *EMBO J.* **23**, 2029–2038.
- McRee, D. E. (1992). *J. Mol. Graph.* **10**, 44–47.
- Mueller-Dieckmann, C., Polentarutti, M., Carugo, K. D., Panjikar, S., Tucker, P. A. & Weiss, M. S. (2004). *Acta Cryst.* **D60**, 28–38.
- Otwinowski, Z., Borek, D., Majewski, W. & Minor, W. (2003). *Acta Cryst.* **A59**, 228–234.
- Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
- Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
- Pflugrath, J. W. (1999). *Acta Cryst.* **D55**, 1718–1725.
- Phillips, J. C. & Hodgson, K. O. (1980). *Acta Cryst.* **A36**, 856–864.
- Ren, H., Wang, L., Bennett, M., Liang, Y., Zheng, X., Lu, F., Li, L., Nan, J., Luo, M., Eriksson, S., Zhang, C. & Su, X. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 303–308.
- Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* **D58**, 1772–1779.
- Schubot, F. D., Kataeva, I. A., Chang, J., Shah, A. K., Ljungdahl, L. G., Rose, J. P. & Wang, B.-C. (2004). *Biochemistry*, **43**, 1163–1170.
- Sheldrick, G. M. (2000). *SADABS* v.2.03. Bruker ANS Inc., Madison, Wisconsin, USA.
- Sims, P. A., Larsen, T. M., Poyner, R. R., Cleland, W. W. & Reed, G. H. (2003). *Biochemistry*, **42**, 8298–8306.
- Terwilliger, T. C. (2000). *Acta Cryst.* **D56**, 965–972.
- Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* **D55**, 849–861.
- Thoden, J. B. & Holden, H. M. (2005). *J. Biol. Chem.* **280**, 21900–21907.
- Vanhooke, J. L., Benning, M. M., Bauer, C. B., Pike, J. W. & DeLuca, H. F. (2004). *Biochemistry*, **43**, 4101–4110.
- Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
- Wu, Y., Qian, X., He, Y., Moya, I. A. & Luo, Y. (2005). *J. Biol. Chem.* **280**, 722–728.