Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

New processing tools for weak and/or spatially overlapped macromolecular diffraction patterns

Dominique Bourgeois

ESRF, BP 220, 38043 Grenoble CEDEX, France and LCCP, UPR 9015, IBS, 41 Avenue des Martyrs, 38027 Grenoble CEDEX 1, France

Correspondence e-mail: bourgeoi@esrf.fr

Received 22 January 1999
Accepted 22 June 1999

Tools originally developed for the treatment of weak and/or spatially overlapped time-resolved Laue patterns were extended to improve the processing of difficult monochromatic data sets. The integration program *PrOW* allows deconvolution of spatially overlapped spots which are usually rejected by standard packages. By using dynamically adjusted profile-fitting areas, a carefully built library of reference spots and interpolation of reference profiles, this program also provides a more accurate evaluation of weak spots. In addition, by using Wilson statistics, it allows rejection of non-redundant strong outliers such as zingers, which otherwise may badly corrupt the data. A weighting method for optimizing structure-factor amplitude differences, based on Bayesian statistics and originally applied to low signal-to-noise ratio time-resolved Laue data, is also shown to significantly improve other types of subtle amplitude differences, such as anomalous differences.

1. Introduction

With the advent of third-generation synchrotron sources, challenging projects have been initiated in the field of protein crystallography. One example is 'real-time resolved' crystallography (Moffat, 1998; Stoddard, 1998), which provides insight into the structure of short-lived intermediates with lifetimes as short as a few nanoseconds (Srajer *et al.*, 1996; Perman *et al.*, 1998). In this technique, Laue diffraction patterns are collected from a single 100 ps X-ray pulse following reaction initiation in the crystal by a light pulse. Crystallographic data obtained in this way suffer from a very unfavourable signal-to-noise (S/N) ratio. This is because the signal is weak (small crystals have to be used to favour a homogeneous triggering of the reaction throughout the whole crystal volume, the excited structural state has a low occupancy owing to limited quantum efficiency of the reaction-initiation technique and structural modifications are faint), the noise is high (high-resolution data are needed and the X-ray background is polychromatic) and patterns are very severely overlapped (Laue patterns are crowded and transient crystal disorder leads to elongated spot-shapes). To extract the whole information content of such data, adequate processing tools have been developed. Most of them specifically deal with the caveats of Laue data (Helliwell *et al.*, 1989; Shrive *et al.*, 1990; Ren & Moffat, 1995*a*,*b*; Yang *et al.*, 1998; Campbell, 1995; Wakatsuki, 1993), but some may be extended to benefit the processing of difficult monochromatic data sets (Bourgeois *et al.*, 1998; Ursby & Bourgeois, 1997). This is the case for spatial overlap deconvolution and evaluation of low-intensity data.

Several fields in macromolecular crystallography indeed suffer from limitations which are similar to those of time-resolved Laue data. Despite the availability of high-brilliance undulator beams and better detectors, projects are often at the limit of the instrumental capability and produce diffraction patterns with poor signal-to-noise ratios, which are frequently overcrowded. This is the case with weakly diffracting crystals, with micro-crystals which necessitate a highly divergent micro-focused beam or with crystals of large unit-cell dimensions. MAD phasing is becoming increasingly popular, and anomalous signals may be swamped in the noise. Cryo-crystallography, which prolongs crystal lifetime, is frequently used to collect high-resolution data sets, and low S/N data are inevitably present in the higher resolution shells. Even 'standard' oscillation patterns collected on synchrotrons or home sources are often hampered by the presence of spatially overlapped spots, owing to developing mosaicity, unfavourable orientation of a crystal mounted in a cryo-loop or an inadequate data-collection strategy. Therefore, the possibility of evaluating weak data more accurately and deconvoluting overlaps may be instrumental in the success of many experiments. It will contribute to better high-resolution refinement, enhance the phasing power of anomalous data and increase the quality of data sets from viruses.

Spatially overlapped reflections are usually rejected by common software (Otwinowski, 1993; Leslie, 1992), which results in a more or less severe reduction in data completeness and redundancy. One might circumvent the problem by reducing the number of spots flagged as overlapped (by choosing a spot size smaller than the measured value), or by simply ignoring overlaps (*i.e.* integrating them as if they were not overlapped). These tricks might improve the overall results in some cases (badly measured reflections are often preferable to reflections not measured at all), but they are poorly justified and might introduce systematic errors which may not appear at the scaling stage. It is preferable to rely on a technique which performs genuine overlap deconvolution.

In this paper, it is shown how integration of monochromatic diffraction patterns with *PrOW* (Bourgeois *et al.*, 1998) significantly improves data quality by allowing deconvolution of spatially overlapped spots, by using an advanced profile fitting technique and by rejecting at an early stage strong outliers originating from zingers or ice-spots. We describe the latest developments made to the program, which include improvement in building the library of reference spots, accurate profile positioning by Fourier interpolation, rejection of non-redundant strong outliers by use of Wilson statistics and improved user-friendliness with a new graphical user interface. A method proposed recently (Ursby & Bourgeois, 1997) based on Bayesian theory (Gilmore, 1996) is also revisited, allowing the improvement of estimates of structure-factor amplitude differences. This technique, originally proposed for poorly accurate time-resolved Laue data, is shown to increase the S/N ratio in difference maps such as anomalous difference Fourier maps (Terwilliger, 1994).

2. Integration with *PrOW*

The program *PrOW* (Profile fitting for Overlapped and/or Weak data) was originally developed for the processing of Laue patterns. In such patterns, as many as 80% of the spots may be spatially overlapped, and a large majority of them have a low $I/\sigma(I)$ ratio, primarily owing to acute sensitivity to crystal disorder and to the presence of a strong polychromatic background. Two main advantages characterize *PrOW*: accurate evaluation of weak intensities by dynamic optimization of the profile-fitting area and the least-squares deconvolution of spatially overlapped spots. It was soon realised that these features could advantageously be extended to the case of monochromatic data. Recently, we have introduced additional improvements, which allow a further increase in the quality of data processed by *PrOW*.

2.1. Optimization of profile-fitting area

Profile fitting of a diffraction spot (q_i) involves determining a normalized model profile (p_i) (built from a library of well defined experimental spots) and minimizing the least-squares sum S:

$$S = \sum_{\mathcal{R}} w_i (I p_i - q_i)^2, \qquad (1)$$

where the q_is are the measured background-subtracted pixel intensities, w_i is a suitable weight (ideally the inverse variance of q_i), I is the integrated intensity to be determined and $\mathcal{R}$ is the profile-fitting area. Ideally, $\mathcal{R}$ should match the exact shape of the spot. However, it is often chosen to be constant, as a disc or an ellipse of fixed dimensions, during the integration of a whole data set. This is not satisfactory, since spot shapes may vary during data collection or within individual images, owing to developing mosaicity, diffraction anisotropy or position-dependant point-spread function (PSF) of the detector. Even if the spot shape did remain constant throughout the whole data set, it has been shown that the determination of $\mathcal{R}$ should take into account the signal-to-noise ratio of the spot under evaluation (Bourgeois *et al.*, 1998). This is justified as follows: suppose that an average bias ε results from statistical or systematic errors made in subtracting the background. This will give a spurious integrated intensity $\varepsilon \sum_{\mathcal{R}} p_i / \sum_{\mathcal{R}} p_i^2$ [solve (1) with $q_i = \varepsilon$], a quantity which diminishes as $\mathcal{R}$ becomes smaller. Conversely, the accuracy with which the real integrated intensity can be evaluated increases as $\mathcal{R}$ becomes larger, until $\mathcal{R}$ encompasses the true spot shape. Therefore, we need to find a compromise between these two effects by determining an optimum profile-fitting area. Practically, this can be achieved in the following way: $\mathcal{R}$ is chosen independently for each integrated spot as the area within a contour level $c_{\mathcal{R}}$ of the reference profile which minimizes an estimate of the expected uncertainty $\sigma_I(\mathcal{R})$. This estimate requires only the knowledge of the background level, the shape of the reference profile (p_i) and an estimate of the peak-to-background ratio of the spot (q_i). An example is shown in Fig. 1.

2.2. Deconvolution of spatially overlapped spots

If K spots $j = 1, 2, \ldots, K$ are predicted to overlap, and p_{ij} is the reference profile at pixel i for the spot j, the intensities (I_j) are found by minimizing the least-squares sum

$$S = \sum_{i \in \mathcal{R}_{1,2,\ldots,K}} w_i \Big(\sum_j I_j p_{ij} - q_i\Big)^2, \qquad (2)$$

where $\mathcal{R}_j$ is the optimized fitting area for the jth spot. Unless $i \in \mathcal{R}_j$, $p_{ij} = 0$. Owing to the difficulties in correctly estimating the variances of the q_is, w_i is chosen to be 1. The intensities (I_j) and their variances are then found by solving the normal equations derived from (2), as described in Bourgeois *et al.* (1998).

A difficulty in all integration programs is the proper flagging of spots as being overlapped or not. Strictly speaking, a very large number of reflections are, in general, spatially overlapped, for example because of a long tail in the detector PSF. However, flagging overlaps by using a drastic criterion such as the width at 0.1% of the PSF would be a non-useful and overdrastic method which would introduce unnecessary complications. A preferable approach consists of determining the level up to which overlapped reflections can be tolerated. Since spot shapes are not uniform and a higher degree of overlap can be accepted for weaker spots compared with stronger ones (because the tails of the former tend to disappear more into the background noise), this level should be evaluated independently for each reflection. In practice, this is a complicated task and an overall criterion based on a minimal distance from neighbouring spots is usually applied. In common programs, this 'overlap distance' usually corresponds to the diameter of the (fixed) fitting area. Since the latter is often chosen (somewhat arbitrarily) by the user, the amount of data flagged as overlapped strongly depends on the user choice. As shown in Fig. 2, this often results in a rather large underestimation of the quantity of truly overlapped spots. Alternatively, the overlap distance could be chosen as the average width of strong spots at a reasonable level, typically 5–10% of peak intensities. This would reduce the bias introduced by the user, but would still not be satisfactory. Here, the use of optimized profile-fitting areas has an obvious advantage: it gives a robust criterion for flagging spatially overlapped spots. If the two fitting areas $\mathcal{R}_1$ and $\mathcal{R}_2$ of two neighbouring spots do intersect, then the two spots will effectively be deconvoluted. If they do not intersect, then (2) reduces to

$$S = \sum_{i \in \mathcal{R}_1} w_i (I_1 p_{i1} - q_i)^2 + \sum_{i \in \mathcal{R}_2} w_i (I_2 p_{i2} - q_i)^2 \qquad (3)$$

and the two spots are not deconvoluted. It is, therefore, sufficient to largely overpredict the number of spatially overlapped spots and let the deconvolution algorithm sort out which spots actually need deconvolution, the only drawback being an increase in computing time.

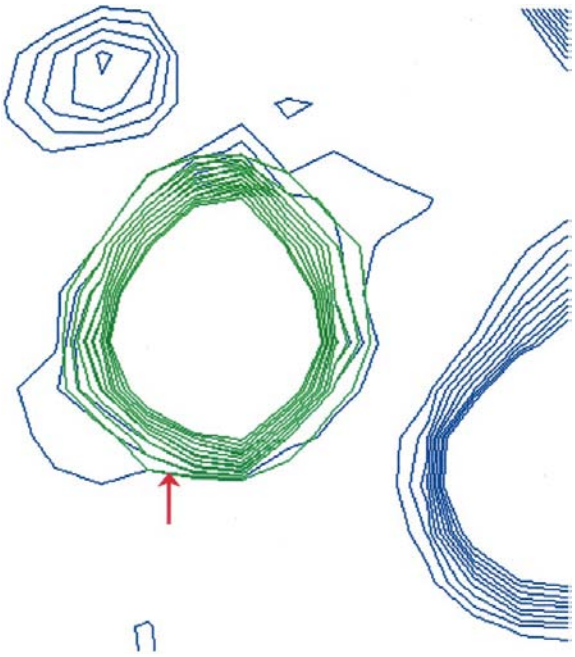

Figure 1
Optimized profile-fitting area. The contour plot of a small area extracted from a diffraction pattern is shown in blue. Three spots are clearly visible. The contour plot of the reference profile associated to the spot in the centre is shown in green. Green contour lines are drawn at levels of 1–10% of the reference profile-peak value. The external contour line (red arrow), corresponding in this case to the 1% level, delineates the fitting area associated to the spot.

2.3. Library of reference profiles

One of the most important steps in profile-fitting integration is the building of reliable well defined reference profiles which can be used to accurately model weak or overlapped spots. A common method consists of averaging a large number of diffraction spots located at a close distance to the reflection under evaluation. No particular constraint is used to select these spots, but the averaging process usually includes weighting schemes based, for example, on the distance to the spot under study. In the case of heavily overlapped patterns, this technique may produce incorrect reference profiles spoiled by satellite peaks. In *PrOW*, a different strategy was chosen: a library of reference spots is built from diffraction spots which fulfil a number of criteria. These take into account

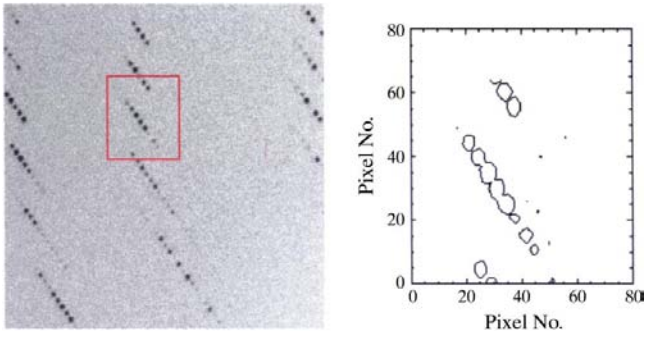

Figure 2
Plot of a region extracted from a monochromatic diffraction pattern (left). Although the spots appear to be well separated, most of them are not, as can be seen on the contour plot (right) corresponding to the area inscribed in the red rectangle and drawn at the 10% level of the highest peak in this area.

peak-to-background ratios, full-width at half-maximum (FWHM) discrepancy between predicted and observed positions, absence of saturated pixels, evenness of the background, the presence of hot or dead pixels and potential overlapping spots. These criteria are severe and may result in the selection of too few satisfactory spots, a majority of which are at low resolution and located close to the beam centre. Such a situation is undesirable, especially in the case of anisotropic patterns, as many weak or overlapped spots are poorly modelled by reference spots located too far away. To remedy this situation, three solutions are proposed. Firstly, reference spots are searched separately in a number of sectors in each image (typically 18) and the selection criteria to which they are submitted are progressively weakened until a decent number of satisfactory spots are selected. This ensures a relative evenness in the distribution of the reference spots within detector space. Secondly, the case of heavily overlapped patterns is handled in the following way: in such patterns, a fair number of spots, although predicted to be spatially overlapped, are of excellent quality because the spots with which they are supposed to overlap are extremely weak or absent. *PrOW* checks for this situation and allows the inclusion of such spots in the library. Thirdly, the library of reference spots can be extended to neighbouring frames. In principle, this should be performed as soon as partial reflections are allowed to be included in the library (the default case), in order to reconstruct reference profiles which have the shape of full reflections (Greenhough & Suddath, 1986). The angular range covered by the library should then primarily depend on crystal mosaicity. In practice, we observe that the difference between the shapes of partial and full reflections is rather faint and that the optimal range essentially depends on the stability of the diffraction power of the crystal. It is safer and computationally less expensive to extend the library backwards only, including the reference spots from the images which have been already successfully integrated. If this option is used, the weight given to a reference spot in building a reference profile includes both its spatial and angular distance to the spot being integrated.

2.4. Interpolation of reference profiles

Even with the highest positional accuracy of predicted patterns, sampling effects originating from image digitization deteriorate the quality of profile fitting (Leslie, 1991). A reference profile can only be positioned relative to a spot under evaluation in steps of one pixel, which makes the matching rather coarse when the spot FWHM corresponds to only a few pixels. Note that this problem does not appear when analytical reference profiles are used instead of experimental profiles. However, in the latter case, interpolation procedures can be used to position reference profiles with subpixel accuracy. In *PrOW*, a cubic interpolation method is used (Park & Schowengerdt, 1983), which closely approximates the theoretically optimum *sinc* interpolation [based on convolution with a $(\sin x)/x$ function]. The procedure relies on predicted spot positions and therefore works best when the root-mean-square deviation between predicted and observed spot positions is significantly smaller than the pixel size. Interpolation is used at two stages: firstly, to precisely centre each reference spot within its box, which reduces broadening effects during averaging processes, and secondly, to precisely position reference profiles relative to spots under evaluation. An example is shown in Fig. 3.

2.5. Rejection of strong outliers using Wilson statistics

Diffraction patterns may contain spurious spots which are located on top or close to predicted reflections. These reflections are assigned incorrect, often way too big, integrated intensities, as well as strongly biased, often way too small, σs. They are strong outliers. Spurious spots may originate from crystalline matter surrounding the sample (collimator, beamstop, satellite crystals or crystalline ice if the sample has been imperfectly flash-frozen), from cosmic rays hitting directly a CCD detector or from so-called 'zingers' which are produced by radioactive elements (essentially thorium) present in fibre-optic tapers. In favourable cases, they saturate the detector or show unacceptable background or spot profiles and the corresponding spoiled reflections are rejected at the integration stage. Often, outliers are only detected at the scaling stage, by comparison with redundant measurements of equivalent reflections. However, there are cases where outliers are not rejected during integration, are measured only once and never get thrown away. This situation is quite insidious, since it will generally not be easily detectable during scaling procedures. Only a few strong outliers may deteriorate the data quality to a considerable extent. This might be because of a general degradation of the scaling performance affecting the whole data set and/or to the presence of a few incorrect structure-factor amplitudes carrying an enormous weight during structure refinement or map calculation (R. Read, personal communication). An example of a zinger is shown in Fig. 4, which resulted in a high-resolution reflection – which happened to be measured only once in the data set – being assigned a huge integrated intensity. To eliminate such reflections, a coarse filter based on the use of Wilson statistics has been implemented in *PrOW*. We make the *a priori*

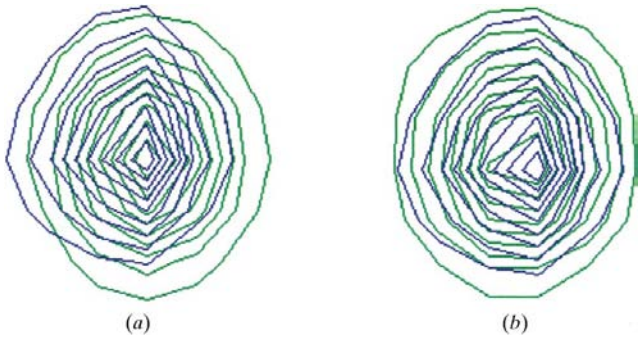

Figure 3
Effect of profile interpolation. Contour plots of an experimental spot are shown in blue. Contour plots of the associated reference profile are shown in green. (*a*) Without interpolation, (*b*) with interpolation. Interpolation clearly improves the matching between the two spots.

assumption that a sufficiently large set of measured intensities I is distributed according to the Wilson probability law

$$p(I) = \begin{cases} (\varepsilon\Sigma)^{-1}\exp(-I/\varepsilon\Sigma) & \text{for acentric reflections} \\ (2\pi\varepsilon\Sigma I)^{-1/2}\exp(-I/2\varepsilon\Sigma) & \text{for centric reflections} \end{cases}, \quad (4)$$

where $\Sigma = \langle I/\varepsilon \rangle$, I is the integrated intensity (corrected for Lorentz and polarization factors) and ε is the correction factor for the expected intensity in a reciprocal-lattice zone (Wilson, 1950). For our purpose, it is acceptable to estimate Σ independently for each image and in a few resolution bins.

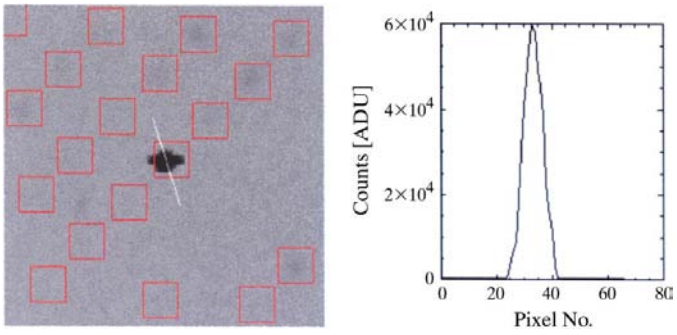

Figure 4
Plot of a high-resolution area extracted from a monochromatic diffraction pattern (left). Predictions are shown as red rectangles. A zinger is clearly visible in the centre of the area, falling on top of a predicted (weak) spot. A plot of a profile cut through this zinger along the white line is shown on the right. This profile is more than 10 pixels wide and almost reaches the detector saturation level. The corresponding integrated intensity is huge.

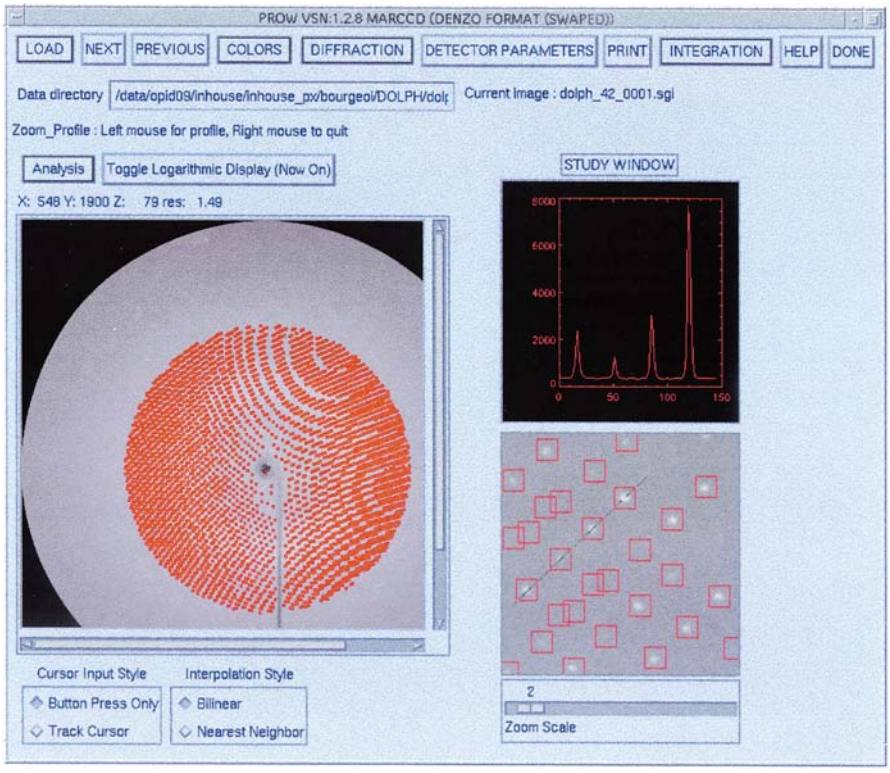

Figure 5
The *PrOW* graphical user interface.

Contrary to correctly measured integrated intensities, strong outliers are expected to deviate considerably from the Wilson law, showing an excessively small probability $p(I)$. Those measurements for which the integral $\int_I^\infty P(J)\,\mathrm{d}J$ {$= \exp(-I/\varepsilon\Sigma)$ for non-centric reflections and $1 - \mathrm{erf}[(I/2\varepsilon\Sigma)^{1/2}]$, where erf refers to the error function, for centric reflections} is smaller than a pre-defined small cutoff value, typically 1×10^{-9}, can thus be safely rejected. More precise estimations of the expected distribution of the integrated intensities could be made based on further *a priori* knowledge: for example, if a partial structure is known (R. Read, personal communication).

2.6. Graphical user interface

Image inspection and the correct choice of parameter values can be carried out using the *PrOW* graphical user interface (GUI) shown in Fig. 5. This GUI as well as a major part of the program has been written with the commercially available code IDL (version 5.1). CPU-demanding routines have been written in C to speed up calculations. The GUI has several attractive features, such as the possibility of drawing masks of any shape in order to prevent integration of unreliable parts of diffraction patterns, to visualize overlapping spots by three-dimensional or contour plots or to carefully inspect integration results. Several image formats are supported, including for example MAR image-plate detectors (18 cm, 30 cm or MAR 345), MAR CCD or FRELON and XRII CCD detectors developed at the ESRF. Predictions are read in either from *DENZO* '.x files' (monochromatic case) or from *LAUGEN* '.ge files' (Laue case). New integrated intensities and σs obtained by *PrOW* are output in the same formats. *PrOW* offers other new features specific to Laue patterns, the details of which will be published elsewhere.

2.7. Results with monochromatic data

PrOW was used to process a number of data sets from a transferase protein (whose structure is still under determination) recorded on flash-frozen crystals with a MAR CCD detector at the ESRF QUADRIGA (ID14/EH3) beamline. All *DENZO* processing sessions were performed by an independent skilled user. A first data set (referred to as #1, collected to 1.45 Å resolution) contained a significant amount of spatially overlapped spots (the detector was too small to collect such data) as well as scattered zingers and ice-spots, some of which (an average of one per frame) ended up right on top of predicted reflections. The presence of the resulting strong outliers, combined with a significant reduction in data redundancy at high resolution owing

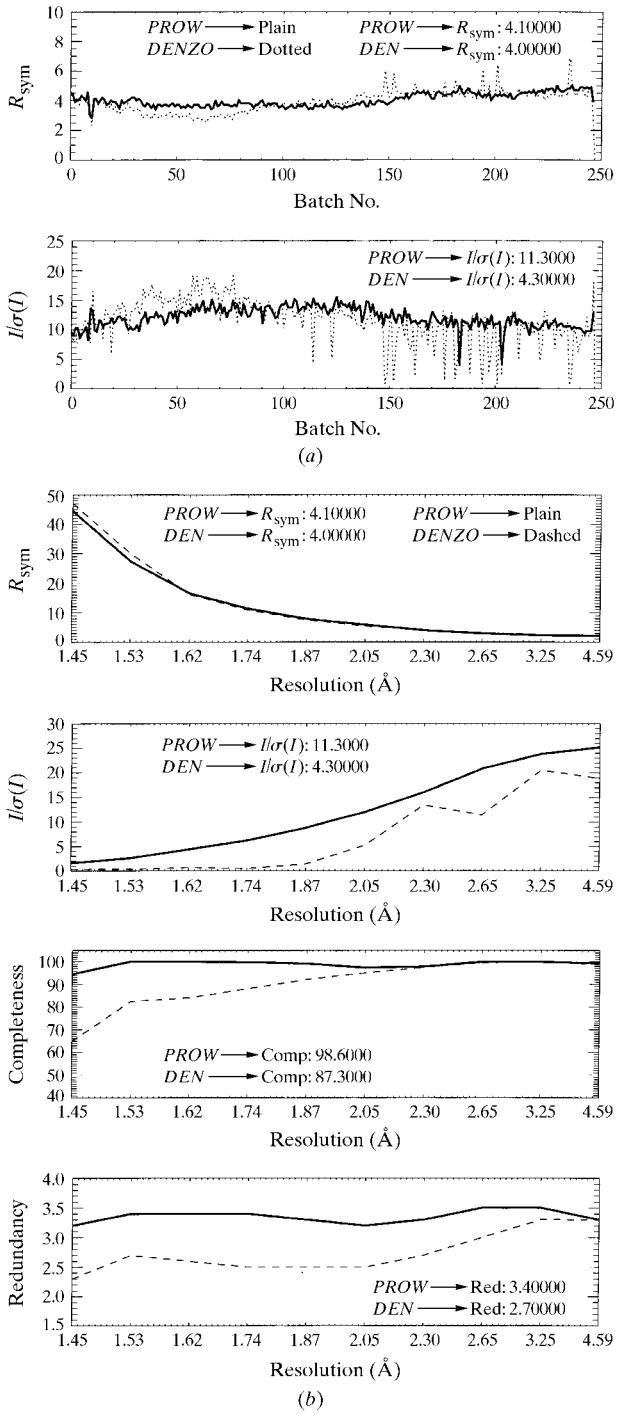

Figure 6
(*a*) Plots of R_{sym} and $I/\sigma(I)$ as a function of batch (or image) number for a data set (#1) of a transferase enzyme (unit-cell parameters a = 73.65, b = 133.54, c = 102.14 Å, α = 90, β = 105, γ = 90°, space group $P2_1$). Data were collected at 0.947 Å. Results obtained with *PrOW* and *DENZO* are shown as plain and dotted lines, respectively. A significant number of batches could not be scaled in a satisfactory way with *DENZO*, essentially as a result of the presence of few strong outliers. This resulted in R_{sym} spikes and $I/\sigma(I)$ glitches, which are almost fully recovered when *PrOW* is used. Data were scaled in the same way with *SCALA* (version 2.3.1). Both data treatments used the predictions from *DENZO*. (*b*) Statistical factors [R_{sym}, $I/\sigma(I)$, completeness and redundancy] are shown as a function of resolution for that same data set. The diffraction patterns contained a significant number of spatial overlaps at high resolution. Results obtained with *PrOW* and *DENZO* are shown in plain and dashed lines, respectively. Overall statistical factors are also shown.

to the rejection of spatial overlaps, produced an ill-behaved scaling after integration with *DENZO* (Fig. 6*a*), resulting in spikes in R_{sym} factors and glitches in $I/\sigma(I)$ ratios for a number of frames. The latter frames are not always those where strong outliers are observed. These abnormalities are almost fully recovered with *PrOW*. A concomitant improvement in data completeness and redundancy is also observed, at the expense of a marginal increase in the overall R_{sym} factor (Fig. 6*b*). The drastic improvement in $I/\sigma(I)$ is primarily a consequence of the significant gain in redundancy when spatial overlaps are deconvoluted, but also of the higher accuracy of the profile-fitting technique. An average of ∼1000 spots per image, corresponding to 25% of the total predicted pattern, were spatially overlapped and successfully deconvoluted. The improvement is exacerbated by the fact that many outliers (including those arising from zingers or ice-spots) are efficiently removed at the scaling stage, thanks to the higher redundancy. In this case, the influence of automatic strong outlier rejection in *PrOW* is minor. If strong outliers were not present in this data set, the improvement would not be as spectacular, but would most probably still be significant (see, for example, the results obtained in the absence of strong outliers by Bourgeois *et al.*, 1998).

In an attempt to search for heavy-atom derivatives, a second data set (#2) was collected to 2.0 Å resolution, in which there were no significant spatial overlaps. However, zingers and ice-spots were still present and, in order to detect anomalous contributions, Friedel mates were kept separated during scaling, which again resulted in a reduced data redundancy. Although it turned out that the derivative was poor and did not give any anomalous signal, better data quality was obtained with *PrOW* (Table 1). Most spikes in R_{sym} factors and glitches in $I/\sigma(I)$ ratios observed with *DENZO* were successfully retrieved with *PrOW*. Processing with *PrOW* was performed twice, with and without using the algorithm for strong outlier rejection. Table 1 demonstrates that about half of the improvement in statistical factors originates from the superior accuracy of the profile-fitting technique, the other half resulting from the successful recognition and rejection of few strong outliers (72 over the whole data set). It is also observed that when the strong-outlier rejection algorithm is not used, significantly fewer spikes and glitches are observed when compared with the case where *DENZO* is used (nine instead of 17). This again relates to the better accuracy of the measurements, leading to more efficient scaling, even though (contrary to data set #1) there was no increase in redundancy in this case. Normal probability plots (Howell & Smith, 1992) confirmed the superior quality of the processing with *PrOW* (Fig. 7). They showed that (spurious) anomalous differences measured with *PrOW* were significantly closer to the expected differences (which do not include any anomalous contribution and are based only on a normal distribution of errors) as compared with the case where *DENZO* is used. The normal probability plot obtained when strong outliers are not rejected (not shown) is almost indistinguishable from the one calculated with these outliers rejected. This highlights the well known fact that statistical factors derived from scaling

Table 1
Comparison of statistical factors for data set #2 collected to 2.0 Å resolution on a MAR CCD detector.

The values are given for the whole resolution range (30–2.0 Å) and, in parentheses, for the last resolution shell (2.11–2.0 Å). Data were scaled with *SCALA* (version 2.3.1). Both processes used predictions from *DENZO*. 'SOR' stands for 'strong outlier rejection'.

	R_{sym}†	R_{anom}‡	R_{meas0}§	PCV0¶	$I/\sigma(I)$	Completeness	Redundancy	Number of glitches††
PrOW (with SOR)	5.8 (40.6)	3.8 (24.6)	7.7 (51.6)	8.7 (59.6)	8.5 (1.5)	99.8 (99.8)	3.5 (3.5)	3
PrOW (without SOR)	6.1 (41.4)	4.1 (24.6)	8.2 (52.4)	9.4 (60.9)	5.3 (0.9)	99.8 (99.8)	3.5 (3.5)	9
DENZO	6.4 (46.1)	4.4 (27.9)	8.8 (58.8)	10.4 (70.5)	4.5 (0.8)	99.8 (99.8)	3.5 (3.5)	17

† $R_{sym} = \sum_{hkl}\sum_i |I_i - \langle I\rangle| / \sum_{hkl}\sum_i |I_i|$. ‡ R_{meas0}, multiplicity-weighted R_{sym} (relative to the overall mean). See Diederichs & Karplus (1997). § PCV0, pooled coefficient of variation (relative to the overall mean). See Diederichs & Karplus (1997). ¶ $R_{anom} = \sum_{hkl} |I_+ - I_-| / \sum_{hkl} |I_+ + I_-|$. †† A glitch is associated to a particular frame n if $[I/\sigma(I)]_n \leq 0.25\ \{[I/\sigma(I)]_{n-1} + [I/\sigma(I)]_{n+1}\}$.

procedures do not entirely reflect all aspects of data quality. Although the presence of few strong outliers may be extremely detrimental to further data treatment (*e.g.* model refinement with maximum likelihood) and therefore should be rejected whenever possible, they have very little impact on the overall distribution of anomalous pairs. In conclusion, for this data set, it is clear that the observed improvements were primarily a consequence of the superior accuracy of the whole integration process, in addition to the successful rejection of strong outliers.

A third data set (#3) was collected to 2.5 Å resolution with a crystal of a mercury derivative, which eventually led to successful structure determination. In this case, there was almost no spatial overlap and no detectable non-redundant strong outliers. The quality of the processing with *PrOW* was still significantly higher, the improvement being solely because of the more accurate profile-fitting technique. The overall $I/\sigma(I)$ ratio (derived by the program *SCALEPACK* version 1.4) increased from 10.3 to 11.6, the linear R_{sym} factor decreased from 5.8% (21.2% in the highest resolution shell) to 5.4% (19%) and there was a spectacular drop in the squared R_{sym} factor from 7.2 to 4.5%, suggesting a more efficient rejection of the (not necessarily strong) outliers. As expected, there was no noticeable change in completeness nor redundancy. A concomitant moderate improvement was observed in a Patterson anomalous difference map (not shown). Although this map was computed using low-resolution data (between 15 and 3.5 Å resolution) of high S/N ratio [$I/\sigma(I) > 3$], the S/N ratio of the main peak in the map increased by 4.6% (from 9.14σ to 9.57σ), indicating that the improvement obtained by *PrOW* not only concerns weak intensities, but also large ones. Further results on data set #3 are described in §3.2.

3. Signal-to-noise improvement of structure-factor amplitude differences with '*q* weighting'

3.1. Bayesian foundation for the *q*-weighting technique

The calculation of correct structure-factor amplitude differences (SFAD) is essential to obtain meaningful difference Fourier maps or to perform difference refinement (Terwilliger & Berendzen, 1995, 1996). The accuracy of amplitude differences can be altered in two ways: firstly by the presence of inaccurate measurements for which error estimates are correctly assessed and secondly by the presence of outliers for which both structure-factor amplitudes and error estimates are incorrect. Inaccurate measurements should be given less weight as they tend to be too large. This is because the distribution of the SFADs results from the convolution of true differences with false differences arising from statistical noise. Therefore, its width is increased by the presence of noise. Outliers (for example resulting from non-redundant spurious measurements, as described in §2.5) should be discarded. We have developed a simple technique based on Bayesian theory to efficiently perform these two actions. A thorough theoretical treatment was given in Ursby & Bourgeois (1997). In this paper, we describe the basic underlying ideas. Our weighting of SFADs is similar to the method proposed by French & Wilson (1978) to derive amplitudes from intensities and our way of discarding outliers is similar to the method used in *PrOW*. In both cases, we take advantage of the *a priori* knowledge we have of the data. Three reasonable assumptions are made. Firstly, we admit that for most of the data, measurement uncertainties $\sigma_{F_{obs}}$ are correctly estimated

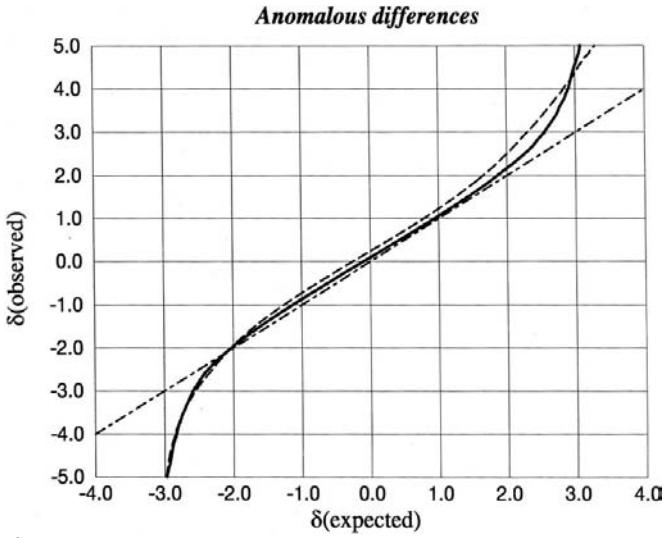

Figure 7
Normal probability plots of observed anomalous differences *versus* expected differences. Since there was no significant anomalous contributor in the crystal, the plots should ideally follow the straight dotted line. The normal probability plot obtained with *PrOW* (plain thick line) is much closer to this line than the one obtained with *DENZO* (dashed line). The normal probability plot obtained with *PrOW* in the case where strong outliers are deliberately not rejected, is barely deteriorated (not shown, see text).

by the scaling algorithm. Secondly, if a structure factor has amplitude F and estimated uncertainty σ_F, we state that experimental measurements of that structure factor have a Gaussian distribution centred around F and of width σ_F. This is an approximation of the fact that the measured intensity is distributed around F^2 according to a Poisson distribution. We call the associated conditional probability law $p_{\rm Poisson}$. Thirdly, we assume that our amplitude differences ΔFs obey Wilson statistics, meaning that they are distributed (in the non-centrosymmetric case) according to the conditional probability law

$$p_{\rm Wilson}(\Delta F/E) = 1/(\pi\varepsilon\sigma_D^2)^{1/2}\exp(-\Delta F^2/\varepsilon\sigma_D^2). \qquad (5)$$

Here, ε is the correction factor for the expected intensity in a reciprocal-lattice zone (Wilson, 1950), E relates to our prior assumed knowledge and we have made the reasonable assumption that the component of $\mathbf{\Delta F}$ collinear to $\mathbf{F}$ is almost equal to ΔF (Henderson & Moffat, 1971). The parameter σ_D [for a rigorous definition, see Ursby & Bourgeois (1997) and references therein] relates to the 'distance' between our related structures. It can be inferred from the data itself, essentially as the width of the distribution of the observed differences taken in a number of resolution bins and to which the contribution of the measurement uncertainties is subtracted. With these three assumptions in mind, we can apply Baye's theorem to determine the optimum probability distribution $p_{\rm opt}$ for each SFAD,

$$p_{\rm opt}(\Delta F/\Delta F_{\rm obs}, E) = p_{\rm Wilson}(\Delta F/E)p_{\rm Poisson}(\Delta F_{\rm obs}/\Delta F, E). \qquad (6)$$

Here, $\Delta F_{\rm obs}$ is our observation (= $F'_{\rm obs} - F_{\rm obs}$, where $F_{\rm obs}$ and $F'_{\rm obs}$ are the two measured amplitudes). Now, according to Blow & Crick (1959), the optimal (or most likely) value of our difference amplitude is at the centre of gravity of the distribution $p_{\rm opt}$,

$$\Delta F_{\rm opt} = \int \Delta F p_{\rm opt}(\Delta F/\Delta F_{\rm obs}, E)\mathrm{d}\Delta F. \qquad (7)$$

A simplification of (7) can be obtained if a few additional assumptions are made, namely $\langle\sigma_{F_{\rm obs}}\rangle << \langle F_{\rm obs}\rangle$ and $\sigma_{F_{\rm obs}} << F_{\rm obs}$. In this case,

$$\Delta F_{\rm opt} = q\Delta F_{\rm obs},$$
$$q = (\varepsilon\sigma_D^2/2)/[\sigma^2_{F_{\rm obs}} + \sigma'^2_{F_{\rm obs}} + (\varepsilon\sigma_D^2/2)]. \qquad (8)$$

In the centrosymmetric case, the corresponding expression for q is obtained by changing σ_D^2 to $2\sigma_D^2$. For the computation of Fourier difference maps involving structure-factor phases (generally calculated phases), the vector $\Delta F_{\rm opt}\exp(i\varphi_{\rm calc})$ also needs to be multiplied by the figure of merit m (Read, 1986). Note that the weights m and q play a parallel role: q relates to the reliability of structure-factor amplitudes, whereas m relates to the reliability of structure-factor phases. The expression (8) is only valid when our three major assumptions are fulfilled. This is not the case for outliers, for which the first assumption is violated: for such reflections, measurement uncertainties $\sigma_{F_{\rm obs}}$ are not reliable. They are generally strongly underestimated. However, it is always possible to assess how much $\Delta F_{\rm opt}$ calculated from (8) deviates from the Wilson law given by (5). In the case of a strong outlier, $\Delta F_{\rm opt}$ is expected to be abnormally large and the probability $p_{\rm Wilson}(\Delta F_{\rm opt}$ excessively small. We can thus safely reject those measurements for which the integral $\int_{|\Delta F_{\rm opt}|} p_{\rm Wilson}(\Delta F)\mathrm{d}(\Delta F)$ {$= 0.5(1 - \mathrm{erf}\,[|\Delta F_{\rm opt}|/(\varepsilon\sigma_D^2)^{1/2}]$ in the non-centrosymmetric case} is smaller than a pre-defined cutoff value, typically 1×10^{-5}.

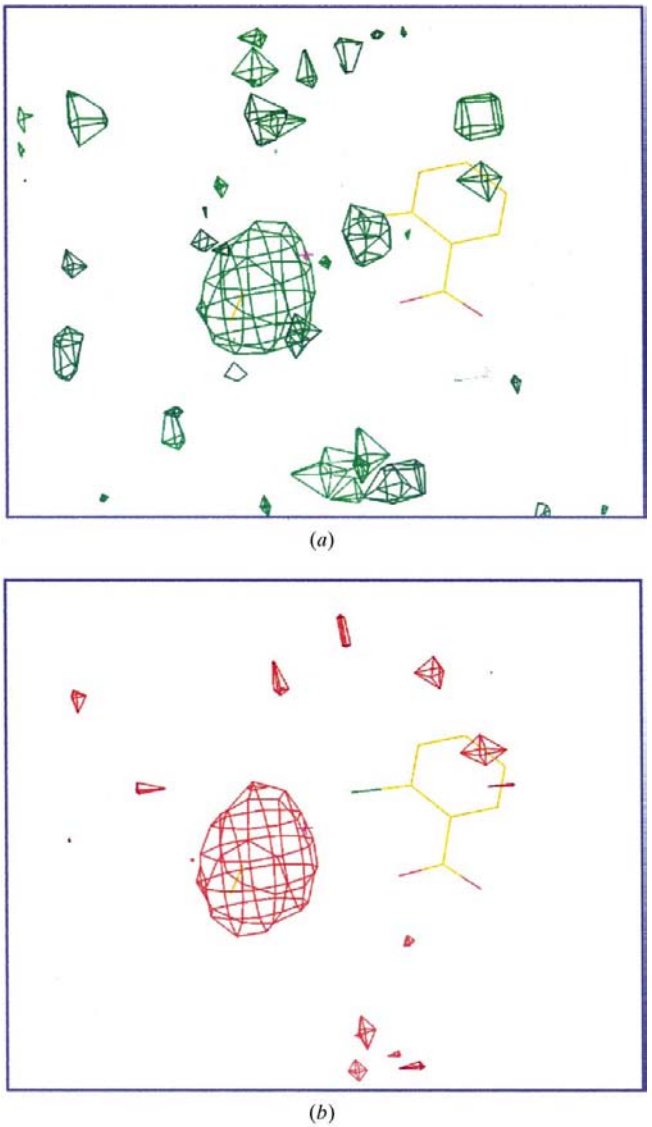

Figure 8
Anomalous Fourier difference-maps obtained with data set #3. The transferase enzyme was soaked with thiomersalate. Data were collected at 0.931 Å. Maps were calculated using the program *FFT* (Collaborative Computational Project Number 4, 1994). (*a*) Map obtained after processing with *DENZO*, (*b*) map obtained after processing with *PrOW* plus '*q* weighting'. The maps are shown at a contour level corresponding to the peak value of the anomalous peak (11.6σ for the first map, 13.0σ for the second) divided by 4.9. In these conditions, the displayed volumes of the anomalous peaks are roughly the same in the two maps. Anomalous peaks do not fall exactly on top of the Hg atom because model refinement was performed with an independent data set and did not take into account anomalous signals.

3.2. Results on anomalous data

The '*q*-weighting' technique was originally developed to improve noisy difference maps calculated from time-resolved Laue data. Here, it is shown how this technique, used in combination with *PrOW*, can significantly improve the quality of anomalous Fourier difference maps. We used data set #3 from our transferase enzyme, as described in §2.7, and computed such maps in order to study the signal-to-noise ratio of the anomalous peaks associated to the Hg atoms. Calculated phases were obtained from an independent data set collected on the same enzyme. This allowed preventing bias that would have resulted if phases from data set #3, processed either with *DENZO* or *PrOW*, had been used. Maps shown in Fig. 8 were obtained at one of the nine Hg sites which showed a low occupancy. A clear improvement in map quality is observed when *PrOW* and the '*q*-weighting' technique are used. The *S*/*N* ratio of the anomalous peak increased from 11.6σ to 13.0σ, *i.e.* a relative improvement of 11.4%. The relative contributions of *PrOW* and '*q* weighting' to this improvement were ~30 and ~70%, respectively (the *S*/*N* ratio increased by 7.85% when the '*q*-weighting technique was applied to the data processed with *DENZO*). One striking finding was that the average *q* value was only 0.34, a value which is of the same order as (if not smaller than) the one usually obtained for the weakest time-resolved Laue data. This confirms that anomalous data have an intrinsically unfavourable *S*/*N* ratio and largely deserve being processed with the tools presented in this article.

4. Conclusions

We have shown that some tools developed to tackle crowded and weak time-resolved Laue patterns can be adapted to significantly improve the processing of challenging monochromatic data sets. The importance of spatial overlap deconvolution to improve data completeness and/or multiplicity has been demonstrated on data sets collected on a reasonably large detector and with unit-cell dimensions which were not excessively large. A profile-fitting technique is described which takes into account modifications of spot shapes during data collection and which can lead to drastic improvements in key statistical factors such as $I/\sigma(I)$, especially at high resolution. We have also shown the importance of rejecting non-redundant outliers at the integration stage and proposed a method for doing so. Finally, we have outlined a simple Bayesian technique allowing the optimization of the calculation of structure-factor amplitude differences and have shown that significant improvements can be obtained in the calculation of anomalous difference Fourier maps.

DB is grateful to Ed Mitchell for the loan of the transferase data and to T. Ursby who has played the leading role in the development of the '*q*-weighting' technique.

References

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Bourgeois, D., Nurizzo, D., Kahn, R. & Cambillau, C. (1998). *J. Appl. Cryst.* **31**, 22–35.
Campbell, J. W. (1995). *J. Appl. Cryst.* **28**, 228–236.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Diederichs, K. & Karplus, P. A. (1997). *Nature Struct. Biol.* **4**, 269–275.
French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517–525.
Gilmore, C. J. (1996). *Acta Cryst.* A**52**, 561–589.
Greenhough, T. J. & Suddath, F. L. (1986). *J. Appl. Cryst.* **19**, 400–409.
Helliwell, J. R., Habash, J., Cruickshank, D. W. J., Harding, M. M., Greenhough, T. J, Campbell, J. W., Clifton, I. J., Elder, M., Machin, P. A., Papiz, M. Z. & Zurek, S. (1989). *J. Appl. Cryst.* **22**, 483–497.
Henderson, R. & Moffat, K. (1971). *Acta Cryst.* B**27**, 1414–1420.
Howell, P. L. & Smith, G. D. (1992). *J. Appl. Cryst.* **25**, 81–86.
Leslie, A. G. W. (1991). *Crystallographic Computing Number 5*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 50–61. IUCr/Oxford University Press.
Leslie, A. G. W. (1992). *Jnt CCP4 /EST–EACMB Newslett. Protein Crystallogr.* **26**.
Moffat, K. (1998). *Acta Cryst.* A**54**, 833–841.
Otwinowski, Z. (1993). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by L. Sawyer, N. W. Isaacs & S. Bailey, pp. 56–62. Warrington: Daresbury Laboratory.
Park, S. & Schowengerdt, R. (1983). *Comput. Vis. Graph. Image Process.* **23**, 256.
Perman, B., Srajer, V., Ren, Z., Teng, T. Y., Pradervand, C., Ursby, T., Bourgeois, D., Schotte, F., Wulff, M., Kort, R., Hellingwerf, K. & Moffat, K. (1998). *Science*, **279**, 1946–1950.
Read, R. (1986). *Acta Cryst.* A**42**, 140–149.
Ren, Z. & Moffat, K. (1995*a*). *J. Appl. Cryst.* **28**, 461–468.
Ren, Z. & Moffat, K. (1995*b*). *J. Appl. Cryst.* **28**, 482–494.
Shrive, A., Clifton, I. J., Hajdu, J. & Greenhough, T. J. (1990). *J. Appl. Cryst.* **23**, 169–174.
Srajer, V., Teng, T. Y., Ursby, T., Pradervand, C., Ren, Z., Adachi, S., Schildkamp, W., Bourgeois, D., Wulff, M. & Moffat, K. (1996). *Science*, **274**, 1726–1729.
Stoddard, B. L. (1998). *Curr. Opin. Struct. Biol.* **8**, 612–618.
Terwilliger, T. C. (1994). *Acta Cryst.* D**50**, 11–16.
Terwilliger, T. C. & Berendzen, J. (1995). *Acta Cryst.* D**51**, 609–618.
Terwilliger, T. C. & Berendzen, J. (1996). *Acta Cryst.* D**52**, 743–748.
Ursby, T. & Bourgeois, D. (1997). *Acta Cryst.* A**53**, 564–575.
Wakatsuki, S. (1993). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by L. Sawyer, N. W. Isaacs & S. Bailey, pp. 71–79. Warrington: Daresbury Laboratory.
Wilson, A. J. C. (1950). *Acta Cryst.* **3**, 258–261.
Yang, X., Ren, Z. & Moffat, K. (1998). *Acta Cryst.* D**54**, 367–377.