Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Zero-dose extrapolation as part of macromolecular synchrotron data reduction

Kay Diederichs, Sean McSweeney and Raimond B. G. Ravelli

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Zero-dose extrapolation as part of macromolecular synchrotron data reduction

Kay Diederichs,[a]* Sean McSweeney[b] and Raimond B. G. Ravelli[c]

[a]Fachbereich Biologie, Universität Konstanz, D-78457 Konstanz, Germany, [b]European Synchrotron Radiation Facility (ESRF), BP 220, F-38043 Grenoble CEDEX, France, and [c]European Molecular Biology Laboratory (EMBL) Grenoble Outstation, 6 Rue Jules Horowitz, BP 181, 38042 Grenoble CEDEX 9, France

Correspondence e-mail: kay.diederichs@uni-konstanz.de

Received 13 March 2003
Accepted 20 March 2003

Radiation damage to macromolecular crystals at third-generation synchrotron sites constitutes a major source of systematic error in X-ray data collection. Here, a computational method to partially correct the observed intensities during data reduction is described and investigated. The method consists of a redundancy-based zero-dose extrapolation of a decay function that is fitted to the intensities of all observations of a unique reflection as a function of dose. It is shown in a test case with weak anomalous signal that this conceptually simple correction, when applied to each unique reflection, can significantly improve the accuracy of averaged intensities and single-wavelength anomalous dispersion phases and leads to enhanced experimental electron-density maps. Limitations of and possible improvements to the method are discussed.

1. Introduction

Radiation damage to protein crystals can be observed at room temperature even on rotating-anode laboratory X-ray sources (Helliwell, 1988). It had already been found in the 1960s that to a good approximation the intensity of most strong reflections, if measured repeatedly during a data-collection run, decreases monotonically with the dose that the crystal has received up to the time when the measurement of the reflection took place. While diffractometers were in use, it was a general practice to correct for the average decrease in intensity by determining the slope of the decay curve for a few strong reflections scattered throughout reciprocal space. These slopes (called 'decay factors' in the following) were interpolated or extrapolated to all other reflections and used to approximately correct for their change in intensity arising from radiation damage. With the introduction of cryogenic crystal cooling and its routine application to protein crystallography (Garman & Schneider, 1997), radiation damage at laboratory X-ray sources and at beamlines of first-generation and many second-generation synchrotrons seemed to be under control.

The effects of radiation damage in cryogenic protein crystallography at third-generation synchrotron sites have been assessed recently by Ravelli & McSweeney (2000), Weik *et al.* (2000) and Burmeister (2000). Whereas in the past it was believed that radiation damage would only affect the resolution at which a structure can be determined (Nave, 1995), it was shown in these papers that radiation damage can also lead to localized modifications of the protein structure. These modifications were most noticeable for disulfide bonds, but were also significant for glutamate and aspartate side chains. Additionally, small translations and rotation of molecules were observed and significant cell changes can occur (Ravelli

et al., 2002). Together, these changes have important implications for the refinement and analysis of macromolecular structures.

Radiation damage also leads to diminished phasing power in experimental methods for phase determination (Rice *et al.*, 2000) as these methods (in particular multiwavelength anomalous dispersion, MAD) rely on an accurate determination of differences between reflection intensities at different wavelengths (dispersive differences) or of differences between the intensities of reflections belonging to Bijvoet pairs (anomalous differences). The photoelectric cross-section of the heavy atom will be very large around the absorption edge at which the MAD data are collected, making the data prone to primary damage. Together with the global and localized changes detailed above, this results in non-isomorphism between data sets that can easily swamp the dispersive differences (Ravelli & McSweeney, 2000). Often, potentially useful complete data sets at the inflection-point and remote wavelengths are therefore not taken into account in phasing and the structure is solved with a SAD (single-wavelength anomalous dispersion) data set using the peak alone (Rice *et al.*, 2000).

The specific modifications of the macromolecular structures induced by radiation damage have been shown to be dependent on dose. We believe that these changes will occur monotonically with dose, thus turning the electron density of a macromolecule into a dose-dependent function in real space. Owing to the nonlinear nature of the Fourier transform, the intensities of the majority of reflections will then change smoothly, but not necessarily monotonically, with dose.

Qualitatively, as radiation damage blurs and weakens electron density, the Fourier coefficients become weaker on average owing to Parseval's Theorem. This is an effect that data-reduction programs model and compensate for using a fall-off factor similar to an overall temperature factor (*SCALEPACK*; Otwinowski & Minor, 1997) or by resolution-dependent scale factors [*XDS*/*XSCALE* (Kabsch, 1988); *SCALA* (Collaborative Computional Project, Number 4, 1994)]. By these means, the average decay of intensity (as a function of resolution and dose) can be compensated for but no correction for the specific changes can be made.

Most scaling programs have other parameters that are used to model, for example, non-uniformity in the absorbance of the macromolecular crystal or non-uniformity in the response of the detector. Refining these parameters for a data set that clearly suffers from radiation damage can improve the apparent quality of the data set somewhat in terms of better values for $R_{\rm meas}$ (Diederichs & Karplus, 1997). However, systematic errors could also be introduced since an incorrect model has been used to deal with radiation damage.

The fact that the intensities of different observations of a unique reflection are expected to vary smoothly with dose can be used as the basis of a new method for compensation for their dose-dependency. The simplest model for a smooth change of the intensities is a linear function, which can also be considered as a first-order approximation of an exponential function. Henderson (1990) has predicted a limit of an absorbed dose of 2×10^7 Gy at which the total crystalline diffractive power of a protein crystal would be completely lost. Teng & Moffat (2000) have shown that some characteristics of radiation damage, such as unit-cell volume increase, change linearly with absorbed dose up to a limit of 1×10^7 Gy, whereas above this dose the crystal starts to decay in a non-linear fashion. We extrapolated that up to a limit of 1×10^7 Gy a linear variation of intensities with dose might be assumed.

The method that we present exploits redundancy in the data by fitting a least-squares line to all equivalent reflections of each unique reflection. To test and assess its suitability, we apply this conceptually simple procedure to a SAD data series collected on a crystal of a selenomethionine-derivatized (SeMet) protein that is highly susceptible to radiation damage.

2. Experimental procedures

2.1. Algorithm and computer program

We have written a computer program 0-*dose* (which can be obtained upon request from KD) in Fortran95 that implements a computational algorithm that is detailed in the following.

2.1.1. Input of data. The basis of the analysis is a single data file containing scaled non-merged intensity observations of one or more data sets. The current version of the program can read the formats written by *XDS* (XDS_ASCII.HKL) and *XSCALE* (option 'MERGE=FALSE'), whereas a newer version that reads *SCALEPACK* (option 'no merge original index') format and *CCP*4 MTZ files is in preparation. The file should be sorted on unique reflection indices and should contain for each observation the original indices, the intensity, its standard deviation (σ), a data-set identifier and the spindle φ value (or any other quantity related to dose) at which the reflection was observed during the measurement of the data set. In principle, the cumulative dose up to this observation could also be given as input.

For each data-set identifier occurring in the file, the program requires as input the total dose that the irradiated part of the crystal had seen at the beginning of the data set ('starting dose') and the dose that the crystal absorbed during each frame ('absorbed dose per frame'). Within this framework, each data set could for example correspond to one wavelength of a MAD data collection, to one of several MIR derivatives or to a low- or high-resolution pass of a native data set.

2.1.2. Evaluation of decay factors. For each unique reflection (hkl) in the file, suppose there are m data sets, with an individual data set denoted by the index j. Within each data set j, there are n_j observations of the intensity of a reflection (hkl), an individual observation being denoted by its intensity y_{ij}, a weight p_{ij} (p_{ij} is σ_{ij}^{-2}) and x_{ij}, the dose at which it was observed. It is assumed that the observations y_{ij} within data set j belong to a common α_j, the (extrapolated) zero-dose intensity of reflection (hkl) in data set j. If we assume a damage factor β common to all observations of (hkl) in the m data sets, we have to minimize the error function

$$S = \sum_j^m \sum_i^{n_j} p_{ij}(y_{ij} - \alpha_j - \beta \cdot x_{ij})^2.$$

Solving

$$\frac{\partial S}{\partial \alpha_j} = -2 \sum_i p_{ij}(y_{ij} - \alpha_j - \beta \cdot x_{ij}) \stackrel{!}{=} 0, \qquad \text{(I)}$$

$$\frac{\partial S}{\partial \beta} = -2 \sum_j \sum_i p_{ij}(y_{ij} - \alpha_j - \beta \cdot x_{ij}) \cdot x_{ij} \stackrel{!}{=} 0, \qquad \text{(II)}$$

it follows that

$$\alpha_j \sum_i p_{ij} - \sum_i p_{ij} \cdot y_{ij} + \beta \sum p_{ij} x_{ij} = 0 \qquad \text{(I}'\text{)}$$

or

$$\alpha_j = \frac{\sum_i p_{ij} y_{ij} - \beta \sum_i p_{ij} x_{ij}}{\sum_i p_{ij}} \qquad \text{(I}''\text{)}$$

and

$$\sum_j \left(\alpha_j \sum_i p_{ij} x_{ij} - \sum_i p_{ij} y_{ij} x_{ij} + \beta \sum_i p_{ij} x_{ij}^2 \right) = 0. \qquad \text{(II}'\text{)}$$

Using (I″), we obtain from (II′)

$$\beta = \frac{\sum_j \sum_i p_{ij} x_{ij} y_{ij} - \sum_j \left(\sum_i p_{ij} y_{ij} \cdot \sum_i p_{ij} x_{ij} / \sum_i p_{ij} \right)}{\sum_j \sum_i p_{ij} x_{ij}^2 - \sum_j \left[\left(\sum_i p_{ij} x_{ij} \right)^2 / \sum_i p_{ij} \right]} \qquad \text{(III)}$$

and

$$\sigma_\beta = \left[\frac{S}{\left(\sum_j n_j \right) - 1 - m} \right]^{1/2} \left\{ \frac{\sum_j \sum_i p_{ij}}{\sum_j \left[\sum_i p_{ij} x_{ij}^2 \sum_i p_{ij} - \left(\sum_i p_{ij} x_{ij} \right)^2 \right]} \right\}^{1/2}.$$

For clarity, the index (*hkl*) has been omitted.

In the case of data sets with anomalous signal, a common decay factor β but different intercepts a_j^+, a_j^- are assigned to the intensities of the Friedel pairs; thus, an anomalous data set is equivalent to two independent data sets without anomalous signal.

The formulas above are general for use with single (native or SAD) and multiple (MIR or MAD) data sets with or without anomalous signal. In addition to the assumption of linearity of the decay, we assume that the decay factor is a property of each unique reflection and is conserved among data sets and at different wavelengths. We therefore make the potentially incorrect assumption that the decay factor does not change when, for example, anomalous data are collected at the peak of the absorption edge of a heavy atom compared with data collected remote from this edge. It is possible to evaluate this assumption during the statistical analysis of the decay (see below).

A straightforward generalization of the current program would be to join data sets with common zero-dose intensity, for example when the same wavelength is collected twice, as in a high- and a low-dose pass. In the current version, these data sets are given different α_j.

2.1.3. Weighting of decay factors. The decay factors and their standard deviations can optionally be down-weighted by multiplication with the factors $1/(1 + \sigma_\beta/|\beta|)$ or $\max(0, 1 - \sigma_\beta/|\beta|)$. This serves to avoid overcorrecting the effects of radiation damage, as it is always safe not to correct at all.

2.1.4. Statistical analysis of the decay. The program performs a statistical analysis of the decay factors as a function of resolution and, if several data sets are present in the input file, a comparison of the decay factors of the data sets. This analysis gives a quantitative summary of the actual radiation damage in terms of fractional average decay, $\langle|\beta|\rangle x_{\max}/\langle\alpha\rangle$, as a function of resolution. The analysis is also performed after separately evaluating the decay factors β_j within each individual data set j. Those reflections for which decay factors can be calculated (*i.e.* those that occur at least twice in each of the data sets) are then used for calculating the ratio of the sums of absolute values of the decay factors,

$$f_{jk} = \sum |\beta_j| / \sum |\beta_k| \quad \text{for reflections common to data sets } j \text{ and } k.$$

Ideally, the observed decay factors β_j should be the same in all data sets and therefore f_{jk} should be 1 when calculated for all common reflections. In practice, values of f_{jk} calculated in resolution shells will differ from 1 owing to noise. Systematic deviations from the ideal value could arise from higher absorption at one wavelength compared with the other, leading to greater radiation damage. In this case, the input value of 'absorbed dose per frame' needs to be adjusted.

The shell-wise values of f_{jk} can also be used to check for deviations from the assumption of linear decay. As an example, the breakdown of linearity can become notable at high resolution when, owing to severe radiation damage, most higher resolution reflections drop to zero intensities.

2.1.5. Output of corrected intensities. The program computes the corrected intensities and their standard deviations as

$$y_{ij}^{\rm corr} = y_{ij} - \beta x_{ij},$$
$$\sigma_{ij}^{\rm corr} = [\sigma_{ij}^2 + (\sigma_\beta x_{ij})^2]^{1/2}.$$

Instead of directly using the extrapolated zero-dose intensities, we thus choose to extrapolate each observation to zero dose and do not make any further reference to the α_j. This serves to preserve the spread among the observations of a unique reflection within a data set. These corrected intensities are written to a file in a format that can be used for post-correction scaling in *XSCALE*.

During zero-dose extrapolation, the $\sigma_{ij}^{\rm corr}$ values are inflated with respect to the σ_{ij} to reflect the uncertainties of the extrapolation, resulting in a data set with an overall weaker $\langle I/\sigma(I)\rangle$. However, since the agreement between the corrected intensities $y_{ij}^{\rm corr}$ becomes better, one would rather expect an overall decrease of the σ_{ij} values based on the average agreement of symmetry-related observation in a resolution shell. This is achieved by re-running *XSCALE* after the radiation-damage correction ('post-correction scaling'), which

Table 1
Data-set statistics.

240 frames of 0.5° per frame were collected from a $P6_5$ crystal of tubulin in complex with RB3-SLD. 'Corrected' refers to zero-dose extrapolated data. The R factor is defined as $\sum|(I(h,i)-\hat{I}(h)|/\sum I(h,i)$, $R_{\rm meas}$ is the redundancy-independent R factor based on intensities, whereas $R_{\rm mrgd\text{-}F}$ gives the quality of amplitudes (F) in the scaled data set (Diederichs & Karplus, 1997).

		R factor (%)		$R_{\rm meas}$ (%)		$R_{\rm mrgd\text{-}F}$ (%)	
Resolution limit (Å)	Completeness (%)	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected
11.70	96.6	3.5	2.2	4.2	2.6	2.4	1.3
8.39	99.5	3.3	2.3	3.9	2.7	2.4	1.6
6.89	99.8	3.9	2.6	4.5	3.0	3.0	2.0
5.98	99.4	6.0	4.3	7.0	5.0	4.8	3.7
5.36	99.6	7.9	6.2	9.2	7.2	6.1	5.0
4.90	99.6	9.5	7.7	11.0	8.9	7.3	6.0
4.54	100.0	11.4	9.3	13.2	10.7	8.4	7.1
4.25	99.5	17.4	14.8	20.1	17.2	13.0	11.5
4.00	90.5	28.7	24.3	33.8	28.6	24.9	20.5
Total	98.0	6.2	4.7	7.2	5.5	7.0	5.7

again corrects the σ_{ij} to make them on average, within a resolution shell, consistent with the observed intensity differences between observations of a unique reflection.

2.2. Test data and computational procedure

Crystals of bovine brain tubulin (two α and two β units) in complex with a selenomethionine-derivatized stathmin-like domain of RB3 (RB3-SLD) were prepared as described previously (Gigant *et al.*, 2000). Crystals grow in space group $P6_5$, with typical unit-cell parameters $a = b = 328$, $c = 54$ Å. The best crystals do not diffract far beyond 4 Å and are highly radiation-sensitive. In total, there are four Se atoms to phase a complex of 210 kDa. Anomalous data were collected at the peak (0.9794 Å) of the absorption edge of Se. The fluorescence scan was measured on the same crystal with a strongly attenuated beam, which was approximated to provide zero dose. Data were collected at 15 K while cooling with He gas using a Helijet (Oxford Diffraction; http://www.oxford-diffraction.com/helijet.htm) in the hope that helium-cooling could extend the lifetime of the crystal in the beam. No systematic studies were performed to compare the lifetime of these crystals at 15 K compared with 100 K owing to the lack of sufficient crystals of constant quality. The data set at 15 K has been used as a test case because it shows a relatively good anomalous signal in the first half of the data, as well as clear signs of radiation damage while data collection continued.

Data were collected at the undulator MAD beamline ID14-4 of the European Synchrotron Radiation Facility (ESRF). An attenuator of about 0.3 mm Al was used, providing threefold attenuation. The exposure time per frame was 5 s and the crystal was rotated by 0.5° per frame. A total of 240 frames were used, which were collected in a continuous sweep. The sixfold crystallographic screw axis was almost parallel to the spindle axis, resulting in an almost complete data set for the first 120 frames and an average redundancy of 7.2 for the 240 frames (applying Friedel's law).

In the absence of a white line, the predicted anomalous signal $\Delta F/F$ at the peak wavelength is 1.2% for the four Se atoms present in the total of 2100 residues. The fluorescence scan showed a reasonable white line, from which a value of 6 e^- was derived for f''. The expected anomalous signal increases to 2% on taking the white line into account. To our knowledge, the ratio (No. of Se atoms)/(No. of residues) for this test case is by far the lowest of those reported so far. The data series was integrated using the program *XDS* (Kabsch, 1988) and scaled using *XSCALE* from the *XDS* suite. The merged data from *XSCALE* were used as the original uncorrected data set. *XSCALE* was also used to write an unmerged data set; zero-dose extrapolation was then performed on these unmerged data using the program 0-*dose* as described above.

Calculation of the f_{jk} factors requires, by definition, the existence of more than one data set. We therefore partitioned the 240 frames into two runs of 120 frames each and used the 0-*dose* program to calculate individual damage factors β for the two partitions, thus obtaining their average ratio f_{jk} in ten resolution shells. After these statistical calculations had been performed, the two runs were discarded.

To obtain the best estimates of the damage factors β, we then used the program on the unmerged intensities of the non-partitioned full data set. This calculation was followed by a rescaling and output of merged intensities using *XSCALE*, thereby obtaining corrected intensities for the ensuing crystallographic procedures.

With both the original and the unmerged data, five sites were found using the program *SHELXD* (Schneider & Sheldrick, 2002), which included all four Se sites plus a fifth site close to a cysteine; the latter most likely arises from reduction of mercurated RB3-SLD by tubulin. Phases were calculated using *SHARP* (de La Fortelle & Bricogne, 1997) and *SOLVE* (Terwilliger, 2002), using the sites found by *SHELXD*. A comparison between the corrected and the original data set was made using a number of criteria, such as number of solutions found by *SHELXD*, peak heights and Z score in *SOLVE* and phasing power and figure of merit as calculated by *SHARP*.

The quality of the resulting phases was assessed visually as well as by the calculation of map correlation coefficients. Only the final phases as obtained with *SHARP* were considered, both after solvent flattening and after additional non-crystallographic symmetry (NCS) averaging using *DM* (Collaborative Computational Project, Number 4, 1994). Two domains were used for the NCS averaging, where the electron density of the two α units was averaged as well as the electron-density of the two β units. As no refined model is currently available, the latter maps, which are visually far superior

to maps without density modification, were used as a reference.

Using an option of the 0-*dose* program, the calculations were repeated after treating the intercepts a_j^+, a_j^- as the same quantity, but the results differed only marginally. This is probably due to the weakness of the anomalous signal.

3. Results

The statistical output of the 0-*dose* program for the tubulin-RB3 data showed damage factors β that are distributed around an average of zero (data not shown), as can be expected for a data set which is scaled such that the average decay of intensity is corrected for. The fractional average decay of the reflections up to 6 Å was constant at 10%, doubling to 20% at 5 Å and rising steeply to 60% in the 5–3.8 Å resolution range.

The f_{jk} factors, which can be used to test the validity of a linear decay, were indeed close to 1, with variations of up to 0.05 except in the lowest resolution shell (50–30 Å; 41 reflections), where the f_{jk} was only 0.85.

Data-set statistics before (uncorrected) and after zero-dose extrapolation (corrected) are given in Table 1. A very clear improvement of data-set statistics can be seen, both in the traditional R factor and the redundancy-corrected $R_{\rm meas}$ (Diederichs & Karplus, 1997). Sites were found by *SHELXD* after conversion of the unmerged data to ΔF and $\sigma_{\Delta F}$ using *XPREP* (Schneider & Sheldrick, 2002). As expected, only a very weak anomalous signal could be found as judged by $\langle \Delta F/\sigma(\Delta F)\rangle$ ratios in resolution shells and by the low correlation coefficients (CC) between the observed and calculated E values (normalized structure factors) for the correct solutions. $\langle \Delta F/\sigma(\Delta F)\rangle$ was higher than 1.5 for reflections below 8 Å before and for reflections below 6 Å after zero-dose extrapolation. Despite this, no major differences were found between the uncorrected and the corrected data set when identifying the Se sites (data to 5 Å) with *SHELXD*.

The program *SOLVE* was run using data to 5 Å. The overall Z score as calculated with *SOLVE* using the sites found by *SHELXD* was somewhat higher after zero-dose extrapolation: 64.9 compared with 63.1 before correction. The peak heights and occupancies of the five sites were marginally larger after correction. The largest differences that were found using *SOLVE* were the figures of merit, which are given in Table 2. Especially between 10 and 6 Å, a clear difference was found. Overall, the mean figure of merit improved substantially from 0.26 before to 0.31 after correction. A slightly worse figure of merit in the lowest resolution bin is observed, but the significance of this is not clear.

The results as obtained by *SHARP* showed the largest improvements when comparing the original and the zero-dose extrapolated data. All data to 4 Å was used, together with the same sites, found by *SHELXD*, as were used in *SOLVE*. Both the occupancies and the positions x, y and z of the five anomalous scatterers were refined, whereas all atomic B factors were fixed at 100 Å^2. The refinements were stable, with one (uncorrected) or no (corrected) eigenvalues being filtered out. Table 3 shows $R_{\rm cullis}$, phasing power and figure of merit before solvent flattening of the acentric reflections, before and after correction. In contrast to *SOLVE*, the gain in phasing power after zero-dose extrapolation is remarkable. The improved phase quality after zero-dose extrapolation is even more evident from the map (Fig. 1) and the map correlation coefficients (Table 4).

Table 2
Mean figure of merit as calculated by *SOLVE* (Terwilliger, 2002).

Values are given based on the same input sites (command analyze_solve) and the data before and after zero-dose correction.

	Figure of merit	
Resolution limit (Å)	Uncorrected	Corrected
17.04	0.24	0.19
11.09	0.27	0.29
8.77	0.29	0.38
7.47	0.29	0.37
6.62	0.28	0.33
6.00	0.25	0.29
5.53	0.24	0.28
5.16	0.25	0.28
Total	0.26	0.31

4. Discussion

Data-reduction programs use data redundancy to correct groups of reflections for their average decay as a function of resolution and dose. Here, we propose using the redundancy to correct individual reflections for their specific decay, thereby extending the traditional strategy and making full use of the available data.

By correcting the observations for radiation damage, we expect to be able to extend the useful data-collection time of single crystals at third-generation synchrotron beamlines. Furthermore, we expect to arrive at more accurate averaged intensities and at better estimates of intensity differences at different wavelengths and thus ultimately to be able to solve and refine structures more quickly and reliably. Zero-dose extrapolation has the promise of providing unbiased intensities for refinement, thus providing a closer look at the 'true structure'.

The model that we investigate and use in this paper is a linear one, which has the advantage of simplicity. This minimizes the substantial danger of overfitting the data, as it adds only a single parameter to be determined for each unique reflection. The model obviously breaks down if the radiation damage is large, as in that case the intensities of many reflections will approach zero, an asymptotic behaviour that cannot be appropriately modelled with a linear function.

Whereas overfitting does not appear to be a problem with the almost eightfold-redundant data set used for this first study, the minimum redundancy for a successful application of the method needs to be investigated. We suggest that to find the best decay model for a given number of observations of a reflection, complete cross-validation (Brünger, 1992) could be used. In addition, and similar in spirit to the application of

R_{free} (Berglund *et al.*, 2002), the success of zero-dose extrapolation can be tested by setting aside a number (*e.g.* 500–1000) of observations that were collected at the beginning of data collection and predicting their intensities by extrapolation from the remaining observations.

As this paper is the first implementation of these ideas, we chose not to modify an existing scaling program but rather to design a new program 0-*dose* which requires the scaling procedure (*XSCALE* in this case) to be re-run after the program. This route gives more flexibility and was also chosen because only *SCALA* is currently available in source code. This results in the average trends of radiation damage *versus* resolution and dose being modelled by the scaling program, whereas our 0-*dose* program only models the deviations of individual observations from the average trend. However, we realise that the correction of single observations interferes to some extent with the other goals of scaling, namely to account for absorption effects and non-uniformity of the detector response.

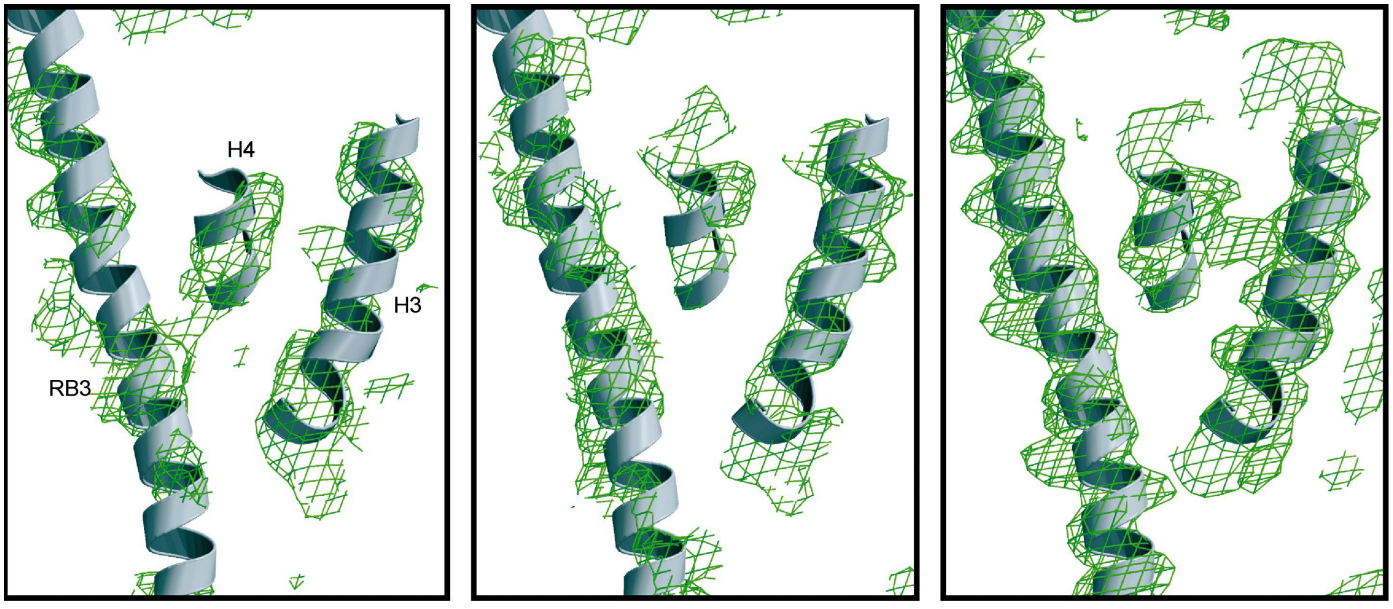
The experimental data to which our present algorithm was applied represent a case beyond the current limits of the SeMet-SAD/MAD phasing method. The improvements in figures of merit and phasing power seen in phasing statistics calculated in *SOLVE* and *SHARP*, however, demonstrate that the method is capable of providing improved intensity data, which in turn lead to better experimental phases. Somewhat surprisingly, the relative improvements in phasing statistics seem to be about equal in all resolution ranges. This might indicate that at low resolution, where the absolute damage is low, it is still substantial when compared with the accuracy of the data, so that a correction yields about the same relative improvement as at high resolution.

Although the quality of the final maps is highly limited by the resolution of the data, a clear improvement upon zero-dose extrapolation is observed in the Se-SAD map before NCS averaging (Fig. 1). The map correlation coefficients (Table 4) between the maps as shown in Figs. 1(*a*) and 1(*c*) is 0.297, whereas this improved spectacularly after zero-dose extrapolation to 0.510 (Figs. 1*b* and 1*c*). NCS averaging of the maps tends to improve them even further, but also makes the maps

Table 3
Phasing statistics as obtained for the acentric reflections using *SHARP* (de La Fortelle & Bricogne, 1997).

	R_{cullis}†		Phasing power‡		Figure of merit	
Resolution (Å)	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected
14.05	0.824	0.673	1.10	1.68	0.319	0.475
8.98	0.806	0.677	1.25	1.80	0.336	0.447
7.06	0.835	0.758	1.37	1.74	0.289	0.355
6.00	0.901	0.882	1.17	1.36	0.244	0.278
5.31	0.941	0.929	0.95	1.08	0.205	0.228
4.82	0.959	0.948	0.75	0.84	0.168	0.190
4.44	0.977	0.970	0.55	0.61	0.128	0.147
4.14	0.990	0.989	0.40	0.45	0.095	0.108
Total	0.932	0.904	0.86	1.04	0.195	0.235

† R_{cullis} = ⟨phase-integrated lack of closure⟩/⟨$|F_{ph} - F_{pl}|$⟩. ‡ Phasing power = ⟨[$|F_h$(calc)|/phase-integrated lack of closure]⟩.



(*a*) (*b*) (*c*)

Figure 1
Experimental 4 Å SAD electron-density maps of tubulin-RB3 . A part of the RB3-SLD helix is drawn, as well as a part of helix 3 and 4 of β_1 of tubulin. (*a*) shows the electron-density map in green after solvent flattening using the original data, whereas the map that was calculated using the zero-dose corrected data is shown in (*b*). (*c*) shows the electron density as calculated with the zero-dose extrapolated data after NCS averaging (α_1 with α_2, β_1 with β_2). The correlation (see also Table 4) between (*a*) and (*c*) is 0.297 and that between (*b*) and (*c*) is 0.510. All maps are contoured at 1.0σ. The figures were produced using *BOBSCRIPT* (Esnouf, 1999) and *RASTER3D* (Merritt & Bacon, 1997).

Table 4
Correlation coefficients of 4 Å electron-density maps.

Solvent flattening was applied using *DM* (Cowtan & Main, 1998) as directed by *SHARP*. NCS averaging was subsequently performed using two domains, one for the two α units and one for the two β units. 'Corrected' refers to zero-dose extrapolated data.

Data set	Correlation between α_1/α_2; solvent flattening only	Correlation between β_1/β_2; solvent flattening only	Correlation of maps before and after NCS averaging
Uncorrected	0.139	0.119	0.297
Corrected	0.311	0.332	0.510

calculated with the uncorrected and the corrected data more identical, so the benefits of zero-dose extrapolation are less obvious when more constraints useable for density modification are available. The correlation coefficient between the NCS-averaged maps using the original and zero-dose corrected data sets was 0.961; thus, they are virtually identical. The NCS relations that were used were based on the unrefined known structure of tubulin-RB3 (Gigant *et al.*, 2000). The correlation coefficients between the two α and the two β units before NCS averaging were 0.139 and 0.119 before, and 0.311 and 0.332 after zero-dose correction (Table 4). This large difference indicates that one would have a much better chance of solving the structure after zero-dose extrapolation if no prior information was available.

Possible applications of our method include the calculation of the intensities that would be observed if all reflections could be measured at the same dose (dose interpolation). This would be useful for the observation and analysis of radiation-damage effects to the protein, as it could provide 'snapshots' of structures during exposure, which are difficult to obtain experimentally (Berglund *et al.*, 2002). These snapshots can be used to assess which parts of a protein are most susceptible to radiation damage and what the chemical reactions induced by radicals are. Another possible application is the generation of two 'before' and 'after' X-ray burn data sets that could be used for phasing using the RIP method (Ravelli *et al.*, 2003).

As interpolation is more robust than extrapolation, interpolation of intensities to a dose value near the middle of data collection might be the preferred method for the correction of MAD data sets from a single crystal. This is easily achieved in the framework of the current program by using a negative 'starting dose' for the first measured data set.

Future work will explore more elaborate non-linear functions for describing and correcting the decay of individual reflections. It is anticipated that the algorithm outlined in this paper, or rather improvements of it, will be incorporated into data-reduction and scaling programs. This will enable its integration with the traditional scaling approaches, resulting in better compensation for decay and more accurate standard deviations of the corrected intensities, and will make its application straightforward.

To enable the future use of this and similar algorithms, we would like to encourage crystallographers to deposit their unmerged data sets with the PDB, together with a description that can be used for calculation of the dose that the crystal has absorbed up to the time a certain observation is made.

The crystals of tubulin in complex with a selenomethionine-derivatized stathmin-like domain of RB3 (RB3-SLD) were kindly provided by Patrick Curmi, Marcel Knossow and Benoit Gigant. We thank Vincent Favre-Nicolin for valuable discussion and André Schiefner for critical reading of the manuscript. The help of David Warner in installing and operating the Helijet is gratefully acknowledged.

References

Berglund, G. I., Carlsson, G. H., Smith, A. T., Szoke, H., Henriksen, A. & Hajdu, J. (2002). *Nature (London)*, **417**, 463–468.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Burmeister, W. P. (2000). *Acta Cryst.* D**56**, 328–341.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. & Main, P. (1998). *Acta Cryst.* D**54**, 487–493.
Diederichs, K. & Karplus, P. A. (1997). *Nature Struct. Biol.* **4**, 269–275.
Esnouf, R. M. (1999). *Acta Cryst.* D**55**, 938–940.
Garman, E. F. & Schneider, T. R. (1997). *J. Appl. Cryst.* **30**, 211–237.
Gigant, B., Curmi, P. A., Martin-Barbey, C., Charbaut, E., Lachkar, S., Lebeau, L., Siavoshian, S., Sobel, A. & Knossow, M. (2000). *Cell* **102**, 809–816.
Helliwell, J. R. (1988). *J. Cryst. Growth*, **90**, 259–272.
Henderson, R. (1990). *Proc. R. Soc. London Ser. B*, **248**, 6–8.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol.* **277**, 505–524.
Nave, C. (1995). *Radiat. Phys. Chem.* **45**, 483–490.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Ravelli, R. B. G., Leiros, H.-K. S., Pan, B., Caffrey, M. & McSweeney, S. (2003). *Structure Fold. Des.* **11**, 217–224.
Ravelli, R. B. G. & McSweeney, S. M. (2000). *Structure Fold. Des.* **8**, 315–328.
Ravelli, R. B., Theveneau, P., McSweeney, S. & Caffrey, M. (2002). *J. Synchrotron. Rad.* **9**, 355–360.
Rice, L. M., Earnest, T. N. & Brunger, A. T. (2000). *Acta Cryst.* D**56**, 1413–1420.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Teng, T. & Moffat, K. (2000). *J. Synchrotron Rad.* **7**, 313–317.
Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1937–1940.
Weik, M., Ravelli, R. B., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J. & Sussman, J. L. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 623–628.