Modelling and refining site-specific radiation damage in SAD/MAD phasing

M. Schiltz[a]* and G. Bricogne[b]

[a]École Polytechnique Fédérale de Lausanne (EPFL), Laboratoire de Cristallographie, CH-1015 Lausanne, Switzerland, and [b]Global Phasing Ltd, Sheraton House, Castle Park, Cambridge CB3 0AX, England. E-mail: marc.schiltz@epfl.ch

Site-specific radiation damage on anomalously scattering sites can be used to generate additional phase information in standard single- or multi-wavelength anomalous diffraction (SAD or MAD) experiments. In this approach the data are kept unmerged, down to the Harker construction, and the evolution of site-specific radiation damage as a function of X-ray irradiation is explicitly modelled and refined in real space. Phasing power is generated through the intensity differences of symmetry-related reflections or repeated measurements of the same reflection recorded at different X-ray doses. In the present communication the fundamentals of this approach are reviewed and different models for the description of site-specific radiation damage are presented. It is shown that, in more difficult situations, overall radiation damage may unfold on a time scale that is similar to the evolution of site-specific radiation damage or to the total time that is required to record a complete data set. In such cases the quality of the phases will ultimately be limited by the effects of overall radiation damage.

Keywords: radiation damage; MAD phasing; heavy-atom refinement.

1. Introduction

The impact of damage induced by X-rays in a macromolecular crystal shows up in two distinct but possibly overlapping effects. Radiation damage first occurs at discrete well localized sites (Burmeister, 2000; Ravelli & McSweeney, 2000; Weik *et al.*, 2000, 2002; Ennifar *et al.*, 2002). Only in a second phase does it affect the whole of the crystal structure, causing long-range disorder with its characteristic consequences of overall intensity decay and loss of high-resolution reflections. Site-specific damage occurs in the form of breakage of disulfide bonds, decarboxylation of aspartate and glutamate residues, loss of hydroxyl groups from tyrosines and loss of methylthio groups of methionines. Heavier atoms such as selenium in selenosubstituted proteins (Rice *et al.*, 2000; Ravelli *et al.*, 2005), bromine in brominated nucleic acids (Ennifar *et al.*, 2002; Ravelli *et al.*, 2003; Schiltz *et al.*, 2004), metals in metalloproteins (Penner-Hahn *et al.*, 1989; Schlichting *et al.*, 2000; Berglund *et al.*, 2002; Yano *et al.*, 2005), iodine (Evans *et al.*, 2003; Zwart *et al.*, 2004) and mercury (Ramagopal *et al.*, 2005) in isomorphous derivatives often exhibit a particularly pronounced sensitivity to X-rays. Site-specific radiation damage can therefore be especially problematic in experimental (*de novo*) phasing methods since it has a deleterious effect on precisely those atoms on which one is relying to obtain the phasing signal.

On the other hand, the site-specific radiation damage that occurs on heavier atoms (including sulfurs) can be used to phase macromolecular crystal structures in a pseudo-SIR (single isomorphous replacement) fashion, where a diffraction data set is first collected on a fresh crystal (the pseudo-native) followed by intense exposure to X-ray or UV radiation and subsequent collection of a second data set containing site-specific radiation damage (a pseudo-derivative) (Ravelli *et al.*, 2003; Nanao & Ravelli, 2006). The method has been termed radiation-induced phasing (RIP) and its feasibility has been established in several case studies including both proteins and halogenated nucleic acids (Ravelli *et al.*, 2003; Evans *et al.*, 2003; Schiltz *et al.*, 2004; Banumathi *et al.*, 2004; Weiss *et al.*, 2004; Zwart *et al.*, 2004; Nanao *et al.*, 2005; Nanao & Ravelli, 2006). In many of these model studies, strongly diffracting and relatively radiation-resistant samples were used (*e.g.* lysozyme, elastase, insulin, thaumatin, trypsin, ribonuclease A) and it was thus possible to obtain essentially complete data sets that were almost unaffected by site-specific and overall radiation damage. In a second stage, site-specific radiation damage was induced by controlled irradiation under conditions that kept overall radiation damage at a low level. Thus, in its basic implementation, RIP requires that at least one complete data set can be collected before substantial site-specific radiation damage shows up. In real-life situations where, more often than not, one has to deal with weakly

diffracting samples, this requirement is frequently not satisfied and site-specific radiation damage may already occur during the early stages of the diffraction data collection. Further, it is not always the case that site-specific and overall radiation damage evolve on significantly different time scales. Thus, in practice, it may not be possible to generate localized damage by X-ray irradiation while at the same time preserving as much as possible the diffraction quality of the sample. These reasons may partly explain why there has so far been no account of a successful determination of a new protein structure by the RIP method, despite the many case studies on model compounds that have been reported. In those cases where site-specific radiation damage was exploited on real-life data (Schiltz *et al.*, 2004; Weiss *et al.*, 2004), it was used as a supplement to phase information generated by anomalous scattering.

There is thus a need and a potential for exploiting site-specific radiation-induced structural changes whenever they occur unavoidably and undesirably and not only in those cases where one is purposely creating radiation damage for RIP. In order to achieve this goal, it is necessary to model the site-specific changes in a continuous fashion as they evolve during a diffraction data collection. This entails that the diffraction data are kept unmerged and that a time or dose stamp is associated with each reflection measurement. In the case of a brominated RNA crystal structure determination, where standard three-wavelength MAD (multi-wavelength anomalous diffraction) phasing was unsuccessful because of fast X-ray-induced debromination (Ennifar *et al.*, 2002), we found that a substantial enhancement of the phasing power was achieved by modelling the site-specific changes in a continuous dose-dependent fashion and by keeping the diffraction data unmerged (Schiltz *et al.*, 2004). The evolution of site-specific radiation damage is explicitly modelled in real space (through atomic parameters) which contrasts with the so-called zero-dose extrapolation method (Diederichs *et al.*, 2003; Diederichs, 2006), where purely empirical corrections are applied to intensity measurements prior to merging.

In this report we review simple models for site-specific radiation damage as they are implemented, or in the process of being implemented, in the phasing program *SHARP* (La Fortelle & Bricogne, 1997; Bricogne *et al.*, 2003). We discuss how phase information is generated through the intensity differences of symmetry-related reflections or multiple measurements of the same reflection, recorded at different stages of the data collection (*i.e.* at different stages of the X-ray irradiation). Finally, we discuss an example of a selenomethionine SAD (single-wavelength anomalous diffraction) structure determination where radiation-induced changes were exploited to aid phasing.

2. Modelling site-specific radiation damage

2.1. The RIP and RIPAS methods in the general context of *de novo* phasing methods

All experimental (*de novo*) phasing methods are based on the exploitation of measurable intensity modulations. The experiment must, for each symmetry-unique reflection $\mathbf{h}$, yield two or more ($N_\mathbf{h}$) data items j corresponding to structure factor amplitudes

$$F_{\mathbf{h},j} = |\mathbf{F}_{\mathbf{h},j}|, \qquad j = 1 \ldots N_\mathbf{h}. \qquad (1)$$

A symmetry-unique reflection is understood to be the equivalence class of $\mathbf{h}$ over the Laue group, *i.e.* including Friedel opposites.

If these intensity modulations are to generate phase information, it is required that, for each reflection $\mathbf{h}$, the complex structure factor can be split up into a part $\mathbf{P}_\mathbf{h}$ that is constant between all data items j and which arises from a common structure, and a variable part $\mathbf{H}_{\mathbf{h},j}$, which is due to a subset of atoms often called the *substructure*,

$$\mathbf{F}_{\mathbf{h},j} = \mathbf{P}_\mathbf{h} + \mathbf{H}_{\mathbf{h},j}. \qquad (2)$$

This condition expresses the requirement that the observed modulations in intensities must be directly related to physical or chemical phenomena occurring in some subset of the crystal structure and are thus not just due to purely experimental or geometric factors that could be modelled by scale factors. To put it more succinctly, the modulations must take place on the complex plane, not just on the intensities. This is the reason why site-specific radiation damage has the potential to generate phase information (if it is correctly modelled), whereas this is not the case for overall (global) radiation damage.

It is further required that, for each data item j, the variable part of the structure factor can be expressed by an atomic model, *i.e.* by a set of parameters in real space (*e.g.* coordinates of atomic positions, occupancy factors, atomic scattering factors, atomic displacement parameters *etc.*), denoted by the parameter vector θ common to all reflections $\mathbf{h}$. For a given reflection, the set of $N_\mathbf{h}$ measured data items can then be written as

$$F_{\mathbf{h},j}(\theta) = |\mathbf{P}_\mathbf{h} + \mathbf{H}_{\mathbf{h},j}(\theta)|, \qquad j = 1 \ldots N_\mathbf{h}. \qquad (3)$$

We are thus in a situation where one part of the complex structure factor is common to all data items, but unknown, whereas, for the part which varies, an atomic model is available for each data item. In this system, each equation defines a circle on the complex plane with radius $F_{\mathbf{h},j}$ and centre $-\mathbf{H}_{\mathbf{h},j}(\theta)$. The set of all $N_\mathbf{h}$ circles represents what is known as the *Harker construction* (Harker, 1956). The system of equations can be solved for the complex quantity $\mathbf{P}_\mathbf{h}$ if $N_\mathbf{h} > 1$ in the case of a centric reflection, or if $N_\mathbf{h} > 2$ in the case of an acentric reflection and provided that the various $\mathbf{H}_{\mathbf{h},j}$ are not collinear in the complex plane. However, the simplicity of equation (3) is usually defaced by various sources of errors, including experimental errors on the measured quantities $F_{\mathbf{h},j}$, model errors on $\mathbf{H}_{\mathbf{h},j}$ and non-isomorphism. The extension of the Harker construction to data affected by errors has led to the formulation of probabilistic treatments culminating in the development of maximum-likelihood methods (La Fortelle & Bricogne, 1997; Bricogne *et al.*, 2003).

The above formulation is very general and encompasses the methods of single or multiple isomorphous replacement (SIR,

MIR), single- or multi-wavelength anomalous diffraction (SAD, MAD), radiation-induced phasing (RIP), or any combination of these methods (SIRAS, MIRAS, RIPAS, SADDAM[1]). In the MIR case, the label j would refer to the various heavy-atom derivatives. In the SAD case, j would stand for either the (+) or (−) Friedel mate. In the case of RIP, j would label the state of irradiation (*e.g.* simply denoting *before* and *after* irradiation).

It is often implicitly assumed that the label j refers to a *data set* (*e.g.* a particular heavy-atom derivative, a data set recorded at a particular wavelength *etc.*), so that the entirety of all structure factor amplitudes $F_{\mathbf{h},j}$ that share the same label j forms a complete and coherent data set, related to a particular crystal structure. However, this is by no means a necessary requirement. Since the Harker construction [equation (3)] is set up individually for each reflection $\mathbf{h}$, j is just a generic label which encodes book-keeping information about the experimental conditions under which each of the $N_{\mathbf{h}}$ data-items of a given reflection were obtained. As a consequence, instead of setting up the Harker construction with various structure factor amplitudes pertaining to different data sets, it can equally well be set up with several symmetry-related measurements or with repeats of the same measurement of a given reflection, all recorded on the same sample and at the same X-ray wavelength. These data items must, however, differ by a variable part $\mathbf{H}_{\mathbf{h},j}$ in their complex structure factor. Under usual circumstances, variations in the intensities of symmetry-related reflections or of repeated measurements of the same reflection are only due to geometric or experimental factors such as differences in absorption, irradiated crystal volume, incident beam fluctuations or decay due to overall radiation damage. These variations can be corrected for by means of multiplicative (scale) factors applied to the intensities. Such intensity differences can therefore not generate phase information and the data items are usually merged into a single structure factor amplitude after the correction factors have been applied. For the same reason, no phase information can be generated by the zero-dose extrapolation method of Diederichs *et al.* (2003), which consists of applying empirical corrections to intensity measurements prior to data merging. The situation is, however, different if site-specific radiation damage occurs during the X-ray data collection. In that case, the crystal structure varies continuously and symmetry-related reflections or repeated measurements of the same reflection which are recorded at different times are no longer equivalent because they pertain to different stages of the radiation-induced structural changes. If these changes are modelled and/or refined explicitly, the intensity differences can be exploited by means of the Harker construction to yield phase information.

It follows from the above considerations that radiation-induced phasing can be subsumed under the general scheme of *de novo* phasing methods (*i.e.* the Harker construction) subject to the condition that the measured data are kept unmerged and that a time or dose stamp is attached to each

intensity measurement. Phasing power can then be generated through the intensity differences of symmetry-related reflections recorded at different doses, *i.e.* corresponding to different states of the X-ray-induced structural changes. In this scheme the operations that are equivalent to data merging and correction (*e.g.* zero-dose extrapolation) are deferred to the phasing stage. Data merging is effectively carried out on the complex plane, *i.e.* through the Harker construction: from all the data items $F_{\mathbf{h},j}$, a single quantity $\mathbf{P}_{\mathbf{h}}$ is estimated, but as a complex value.

It is worth noting that the phase information which is generated through site-specific radiation damage is akin to isomorphous phase information and is thus orthogonal and complementary to the phase information generated by anomalous differences (North, 1965). This may explain why RIP-aided SAD (RIPAS) or SAD with radiation-damage (SADDAM) appears to be a particularly useful method on real-life data.

2.2. Atomic models for site-specific radiation damage

RIP phase information can only be extracted from the diffraction data if a model describing the site-specific structural changes in terms of refinable parameters is available. Such a model contains the usual atomic parameters (atomic positions, occupancy factors, atomic displacement parameters, anomalous scattering factors) as well as a time- or dose-dependent description for the evolution of those parameters that undergo changes as a result of X-ray irradiation. The variable part of the complex structure factor can then be written as a function of X-ray dose,

$$\mathbf{H}_{\mathbf{h},j} = \sum_i \mathbf{f}_i\big(\mathbf{h}, d_{\mathbf{h},j}\big), \qquad (4)$$

where the summation is (in the asymmetric unit of the crystal cell) over all atoms that are initially (before the irradiation) present in the substructure and $d_{\mathbf{h},j}$ is the dose at the jth measurement of reflection $\mathbf{h}$.

In many cases, site-specific radiation damage consists of the breaking of chemical bonds, followed by the apparent 'disappearance' of atoms that were engaged in these bonds (Burmeister, 2000; Ravelli & McSweeney, 2000; Weik *et al.*, 2000, 2002). The atoms do not necessarily disappear from the sample, but they become disordered and are thus no longer contributing to the diffraction signal. The dehalogenation of derivated nucleic acids (Ennifar *et al.*, 2002; Schiltz *et al.*, 2004) and the radiolysis of S—Hg bonds in mercury derivatives of proteins (Ramagopal *et al.*, 2005) are examples where radiation damage manifests itself as a gradual 'disappearance' of atoms. In such cases the substructure contributions can be modelled as follows,

$$\mathbf{f}_i(\mathbf{h}, d) = (f_i^\circ + f_i' \pm \imath f_i'')\big[1 - \xi_i(d)\big]O_i T_i(\mathbf{h})\exp\big(2\pi\imath\mathbf{h}^T\mathbf{x}_i\big),$$
$$(5)$$

where O_i is the occupancy of the atom i at zero dose, $T_i(\mathbf{h})$ its Debye–Waller factor (temperature factor) and other symbols have their usual meaning.

---

[1] SAD in the presence of radiation DAMage.

We here assume that, for each atom i that is subjected to radiation-induced changes, $\xi_i(d)$ represents the *extent of reaction* or *degree of advancement* of the reaction (de Donder, 1920; de Donder & Van Rysselberghe, 1936). Each of these reaction coordinates $\xi_i(d)$ can be represented by a continuous function of time or dose d that varies monotonously between the values 0 and 1. For substructure atoms that are not radiation-sensitive, ξ is fixed to 0. This model simply amounts to replacing the usual time-independent occupancy factor by an effective time- or dose-dependent occupancy $O_i[1 - \xi_i(d)]$. Note that the parameter O_i in this expression then takes the meaning of a zero-dose occupancy factor.

The functions $\xi_i(d)$ must be parametrized by a small number of refinable variables which we denote by the vector $\boldsymbol{\beta}_i$ (note that the various $\boldsymbol{\beta}_i$ are subsets of the general parameter vector $\boldsymbol{\theta}$). In many cases, very simple rate laws can be applied, based on considerations of reaction kinetics. Thus, for the radiation-induced debromination in nucleic acids, it was shown that the reaction follows a simple first-order kinetics (Ennifar *et al.*, 2002; Schiltz *et al.*, 2004). The reaction coordinates can then be parametrized as

$$\xi_i(d) = 1 - \exp(-\beta_i d), \tag{6}$$

where there is only one refinable parameter per atom: the decay constants β_i.

This is the parametrization that was used in the previous case study on a brominated RNA crystal structure determination (Schiltz *et al.*, 2004). It was found that a substantial enhancement of the phasing power could be achieved by modelling the site-specific changes by this simple model, which essentially only adds one new refinable parameter per atom. The fact that the phasing was successful with a minimum number of parameters gives strong support to the validity of the exponential decay model. This parametrization is fully implemented in a version of the program *SHARP* that has been upgraded for the use of scaled but unmerged data. An auxiliary program called *SCALA2SHARP* has been written which prepares a multi-record MTZ data file produced by the CCP4 (Collaborative Computational Project, Number 4, 1994) program *SCALA* (Evans, 1993) for input to *SHARP*. In particular, *SCALA2SHARP* adds a data column containing a dose stamp $d_{\mathbf{h},j}$ for each measurement to the reflection file. Both the upgraded version of *SHARP* and the program *SCALA2SHARP* are available to interested users who have a valid licence for *SHARP*. Requests should be made to sharp-develop@globalphasing.com.

In many cases the evolution of the radiation-induced changes may, however, not be as easily described by simple analytical laws. It happens frequently that the crystal is not fully bathed in the X-ray beam or that the intensity profile of the incident beam displays important variations. In such cases, not all parts of the sample are equally irradiated, and fresh (unexposed) parts of the sample may be continuously brought into the X-ray beam as the sample is rotated, so that the effective (average) occupancy of an atom in the sample does not follow a simple rate-law as a function of dose. In these situations it is not necessarily the case that $\xi(d)$ varies

monotonously as a function of d and it may be advantageous to resort to a purely empirical parametrization of the $\xi_i(d)$ functions, *e.g.* by polynomial interpolation or by quadratic or cubic B-splines with a small number of refinable control points. The testing of such parametrizations in *SHARP* is currently in progress.

2.3. More comprehensive models for site-specific radiation damage in various circumstances

It often happens that radiation-induced breakages of chemical bonds lead to displacements of atoms. This has, for example, been observed in the case of the cysteine sulfur atoms that are released upon the radiation-induced cleavage of disulfide bonds (Weik *et al.*, 2000). The sulfur atoms usually move away from each other but, owing to steric constraints, they do not become completely disordered. Similar radiation-induced atom displacements have also been observed in some cases for mercury atoms in protein derivatives (Ramagopal *et al.*, 2005) and for halogen atoms in nucleic acid derivatives (Ennifar *et al.*, 2003). Suppose that, as a result of a radiation-induced bond cleavage, an atom i moves from its initial position $\mathbf{x}_{i0}$ to a new position $\mathbf{x}_{i1}$. Its contribution to the substructure can be modelled as follows,

$$\begin{aligned}\mathbf{f}_i(\mathbf{h}, d) = (f_i^\circ + f_i' \pm \imath f_i'')\big\{&\big[1 - \xi_i(d)\big]O_{i0}T_{i0}(\mathbf{h})\exp\big(2\pi\imath\mathbf{h}^T\mathbf{x}_{i0}\big)\\&+ \xi_i(d)O_{i1}T_{i1}(\mathbf{h})\exp\big(2\pi\imath\mathbf{h}^T\mathbf{x}_{i1}\big)\big\},\end{aligned} \tag{7}$$

where both sites are linked to a common reaction coordinate $\xi_i(d)$. In most cases the atom which is displaced after a bond cleavage acquires a greater conformational freedom, resulting in a larger Debye–Waller factor. It thus makes sense to refine individual zero-dose occupancy factors and atomic displacement parameters for the initial and final states, respectively. It could even happen that, after the disruption of a chemical bond, the atom moves to two (or more) alternate positions, corresponding to different rotameric side-chain conformations. Thus, the atomic model can be further extended to include multiple final sites $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots$,

$$\begin{aligned}\mathbf{f}_i(\mathbf{h}, d) = (f_i^\circ + f_i' \pm \imath f_i'')\big\{&\big[1 - \xi_i(d)\big]O_{i0}T_{i0}(\mathbf{h})\exp\big(2\pi\imath\mathbf{h}^T\mathbf{x}_{i0}\big)\\&+ \xi_i(d)\big[O_{i1}T_{i1}(\mathbf{h})\exp\big(2\pi\imath\mathbf{h}^T\mathbf{x}_{i1}\big)\\&+ O_{i2}T_{i2}(\mathbf{h})\exp\big(2\pi\imath\mathbf{h}^T\mathbf{x}_{i2}\big) + \ldots\big]\big\}.\end{aligned} \tag{8}$$

Again, the parametrization of the initial and final sites are linked through the common reaction coordinate ξ_i. This model can be fully extended to the special case of disulfide bond breaks by assigning the same reaction coordinate to both sulfur atoms.

A different treatment is needed for the case of metal atoms in metalloproteins whose anomalous scattering is to be exploited at an absorption edge. The main effect of radiation damage in metalloproteins is a photoreduction of the metal centres (Penner-Hahn *et al.*, 1989; Schlichting *et al.*, 2000; Berglund *et al.*, 2002; Yano *et al.*, 2005). The near-edge (XANES) spectral features of the anomalous dispersion factors $f'(\lambda)$ and $f''(\lambda)$ of a metal centre are known to be very sensitive to its oxidation state. This effect has sometimes been used to study radiation-induced reduction of metal centres in

crystalline metalloproteins by X-ray absorption spectroscopy (Penner-Hahn *et al.*, 1989; Yano *et al.*, 2005). In the context of anomalous scattering experiments that are performed at an absorption edge of the metal, radiation-induced reduction can lead to a drift in the anomalous dispersion factors even if the X-ray wavelength remains constant. Both the real and imaginary components are affected. Photoreduction of the absorbing atoms usually leads to a shift of the absorption edge towards lower energies. The intensity of the so-called white line (if present) and other near-edge features are also modified. Depending on the initial position of the edge with respect to the X-ray photon energy, the f' and f'' factors can either increase or decrease. A suitable parametrization would therefore be as follows,

$$\mathbf{f}_i(\mathbf{h}, d) = \left\{ \left[1 - \xi_i(d) \right] (f_{i0}^\circ + f_{i0}' \pm \imath f_{i0}'') + \xi_i(d)(f_{i1}^\circ + f_{i1}' \pm \imath f_{i1}'') \right\} \\ \times O_i T_i(\mathbf{h}) \exp\left(2\pi \imath \mathbf{h}^T \mathbf{x}_i\right), \qquad (9)$$

where it is assumed that there is no change in the positional and atomic displacement parameters of the atom. In a standard MAD experiment, the initial anomalous dispersion factors f_{i0}' and f_{i0}'' can be determined from absorption spectra performed on the sample, prior to the diffraction data collection, and would thus not need to be refined. In some cases, experimental X-ray absorption data might also be available for the final state (they could be easily obtained by recording a fluorescence spectrum at the end of the diffraction data collection) so that f_{i1}' and f_{i1}'' would be known as well; otherwise they may be refined. It is interesting to note that, in this particular case, the radiation-induced changes mimic a multi-wavelength experiment: symmetry-related reflections that are recorded at different times will correspond to different anomalous dispersion factors f' and f'' of the anomalous scatterers, *i.e.* exactly the same effect that is obtained in a MAD experiment by changing the wavelength.

There can be more complicated cases where, for example, changes in the anomalous dispersion factors are concurrent with changes in the atom position *etc.* Clearly, all these cases can be easily parametrized with a limited number of reaction coordinate functions $\xi(d)$. If these functions can in turn be expressed as functions of just one or a small number of variables [as *e.g.* in equation (6)], the risk of overparametrization is minimized. As a general rule in *de novo* phasing, it is important to avoid overparametrization of the substructure model by limiting the number of refinable parameters as much as possible. In this sense the refinement of dose-dependent functional forms [*e.g.* $\xi(d)$] on unmerged data is more efficient than the practice of splitting up a data set into small chunks and refining individual atomic parameters (*e.g.* occupancies, coordinates, scattering factors *etc.*) for each of these partial data sets. In the case of the radiation-induced debromination in an RNA structure (Schiltz *et al.*, 2004), we found that the strategy of breaking up the full data set into 15 chunks and refining individual occupancy factors for the radiation-sensitive Br atoms in each of the partial data sets (*i.e.* a total of 15 occupancy factors per atom) resulted in less good phasing statistics as compared with the scenario where dose-depen-

dent modelling of heavy-atom occupancies in the form of equations (5) and (6) was used, even though the latter involved the refinement of only two occupancy parameters per atom (*i.e.* a zero-dose occupancy O_i and a decay parameter β_i). We therefore expect similar improvements through the implementation of the parametrizations given by equations (7), (8) and (9), which are attempts to devise physically sensible descriptions of radiation-induced effects with a minimal number of refinable parameters. Further, the refinement of such dose-dependent parameters does not critically depend on the data redundancy. In the above-mentioned case of the radiation-induced debromination in an RNA structure (Schiltz *et al.*, 2004), decay parameters (β_i) could be refined successfully in *SHARP* on data sets with a redundancy of less than 2 (M. Schiltz *et al.*, unpublished results). This can be traced back to the fact that the dose-dependent functions $\xi(d)$ are common to all reflections. Even if, for a given reflection $\mathbf{h}$, the redundancy ($N_\mathbf{h}$) is small, the $\xi(d)$ functions (or, conversely, the β_i parameters) are refined simultaneously on all reflections. From the set of all reflections a sufficient coverage of all dose values can be expected, even for low-redundancy data sets. The total number of $\xi(d)$ functions (*i.e.* the total number of radiation-sensitive atoms in the asymmetric unit) is usually rather small, compared with the total number of reflections. This contrasts with the zero-dose extrapolation method of Diederichs *et al.* (2003), where an individual empirical decay factor is refined for each reflection $\mathbf{h}$. Not surprisingly, the success of zero-dose extrapolation relies critically on the degree of redundancy of the data set (Diederichs, 2006).

The more refined atomic models for radiation-induced changes [*e.g.* the models described by equations (7), (8) and (9)] are in the process of being implemented in the program *SHARP*.

2.4. Parametrization of scale factors

Although the merging step is abandoned in the proposed scheme, scaling of the raw data by empirical functions depending on a limited number of refinable scale factors can be carried out in the usual way, *i.e.* by minimizing the disagreement between symmetry-equivalent reflections. It is, however, a well known fact in heavy-atom refinement that the scaling of derivative *versus* native data sets needs to be updated as soon as a model for the heavy-atom substructure is available. Thus, inter-crystal scale factors are usually refined alongside the heavy-atom parameters in phasing programs such as *SHARP*. By analogy, if time- or dose-dependent models for the substructure are refined, it may be advantageous to simultaneously refine a dose-dependent internal scaling function. The importance of adjusting the scale factor between the *before* and *after* burn data sets in RIP was demonstrated by Nanao *et al.* (2005). Analytical expressions for the intensity decay as a function of X-ray irradiation have been proposed by Blake & Phillips (1962), Hendrickson (1976) and Sygusch & Allaire (1988). A more pragmatic approach would be to simply split the data into chunks of

small sequential batches, each consisting of data that have been recorded on a number of consecutive diffraction images, and to refine overall and resolution-dependent scale factors for each chunk. An improvement of such a simple strategy consists of parametrizing an empirical scale factor as a function of X-ray dose, *e.g.* by quadratic or cubic B-splines with a small number of refinable control points (placed at regular intervals of X-ray dose). Such upgrades are currently being implemented in *SHARP*.

2.5. Non-isomorphism parameters

The modelling and parametrization of non-isomorphism in the case of radiation-damage-corrupted data is significantly more complex than for standard data sets. The error model that is currently implemented in *SHARP* assumes that the effects of all sources of non-isomorphism are uncorrelated between different observations of a given reflection (La Fortelle & Bricogne, 1997; Bricogne *et al.*, 2003). In essence, a diagonal approximation is used for the non-isomorphism covariance matrix. Such an approximation may be inadequate since that part of non-isomorphism which is due to radiation damage is expected to be strongly correlated across observations that are closely spaced in time. The diagonal approximation is valid if the site-specific radiation-induced changes are adequately modelled (*i.e.* if the errors in the substructure model are small) and as long as overall radiation damage (which is always heavily correlated in time) is not too severe. For a more general treatment it is necessary to resort to multivariate likelihood functions which are capable of accommodating adequate patterns of covariances between the various observations (Bricogne, 2000; Pannu *et al.*, 2003). The implementation of these functions in *SHARP* is currently underway.

3. Example of a 'real life' selenomethionine case: RIP-aided SAD phasing of Ypa2, a PP2A phosphatase activator

The PTPA protein is an essential and specific activator of the Ser/Thr phosphatase 2A (PP2A) and functions as a peptidyl prolyl isomerase. The crystal structures of a yeast ortholog of PTPA, Ypa2, was solved by RIP-aided SAD phasing on a selenated protein. Details of the structure determination and a description of the molecular structure and biological implications are published elsewhere (Leulliot *et al.*, 2006). We will here focus on the aspects that are related to radiation-induced phasing.

Table 1
Data reduction statistics for Ypa2.

A total of 315 1°-rotation data frames were collected. Data reduction was carried out separately for the first 90 images, the first 180 images and all 315 images, respectively.

	Images 1–90		Images 1–180		Images 1–315 (all)	
	Overall	Last shell	Overall	Last shell	Overall	Last shell
Low resolution limit (Å)	53.45	2.85	53.45	2.85	53.45	2.85
High resolution limit (Å)	2.70	2.70	2.70	2.70	2.70	2.70
R_{merge}	0.125	0.786	0.135	0.582	0.190	0.582
R_{meas} (within $I+/I-$)	0.155	1.100†	0.158	0.746	0.209	0.746
R_{meas} (all $I+$ and $I-$)	0.169	0.879	0.169	0.758	0.213	0.757
R_{pim} (within $I+/I-$)	0.090	0.768	0.081	0.461	0.085	0.460
R_{pim} (all $I+$ and $I-$)	0.077	0.572	0.064	0.377	0.064	0.377
Fractional partial bias	−0.018	−0.042	−0.020	−0.021	−0.015	−0.019
Total number of observations	58399	2021	120443	7240	176223	7240
Total number unique	15273	1376	18827	2452	18828	2452
$\langle I/\sigma_I \rangle$	9.3	1.0	12.3	1.4	12.8	1.3
Completeness (%)	75.2	47.1	91.7	83.3	91.7	83.3
Multiplicity	3.8	1.5	6.4	3.0	9.4	3.0
Anomalous completeness (%)	59.2	13.8	86.9	55.7	86.9	55.7
Anomalous multiplicity	2.4	1.2	3.5	1.9	5.1	1.9
Δ_{ano} correlation between half-sets	0.430	−0.098	0.366	0.000	0.156	−0.058

† Since the multiplicity of the data in this last resolution shell is low, the quality indicators R_{meas} and R_{pim} are computed on a different (sub)set of symmetry-unique reflections than for the sets composed of images 1–180 and images 1–315. This explains why the values appear to be higher.

Ypa2 SeMet crystals were grown from a 1:1 μl mixture of protein (5 mg ml^{-1}) with 16% PEG4000, 0.4 *M* KCl, pH 5.6, 0.1 *M* tris-HCl, pH 8.5, giving thin needle crystals. Addition of guanidine to the mother liquor was found to increase the size of the crystals. X-ray diffraction data were collected on a cryocooled (at 100 K) thin rod-shaped crystal using the undulator beamline ID23-1 at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. In order to optimize the anomalous scattering signal strength, the wavelength was tuned to 0.979835 Å, corresponding to the peak (white line) position of the Se *K* edge, as determined from a fluorescence spectrum recorded on the sample, prior to the diffraction data collection. The crystals belong to space group $C222_1$ with cell parameters $a = 157.9$ Å, $b = 171.0$ Å, $c = 53.7$ Å. There are two molecules in the asymmetric unit, corresponding to 45% solvent content, with 2×10 selenomethionines (for a total of 2×304 residues). A total of 315 raw 1°-rotation data frames were collected in an uninterrupted fashion with the aim of recording a highly redundant data set. The maximum resolution at the start of the diffraction data collection is 2.7 Å. Relevant data reduction statistics are summarized in Table 1. The raw diffraction data were integrated using the *MOSFLM* package (Leslie, 1993). The data were scaled using the program *SCALA* (Evans, 1993) from the CCP4 (Collaborative Computational Project, Number 4, 1994) program suite in the usual way, *i.e.* by minimizing the disagreement between symmetry-equivalent reflections. Data reduction statistics are reported in Table 1. Attempts to locate the Se atoms were carried out simultaneously using the programs *SHELXD* (Schneider & Sheldrick, 2002) and *HYSS* (Grosse-Kunstleve & Adams, 2003). Substructure refinement and phasing was carried out using *SHARP* (La Fortelle & Bricogne, 1997; Bricogne *et al.*, 2003). Phase improvement by solvent flattening was carried out using the *SOLOMON*
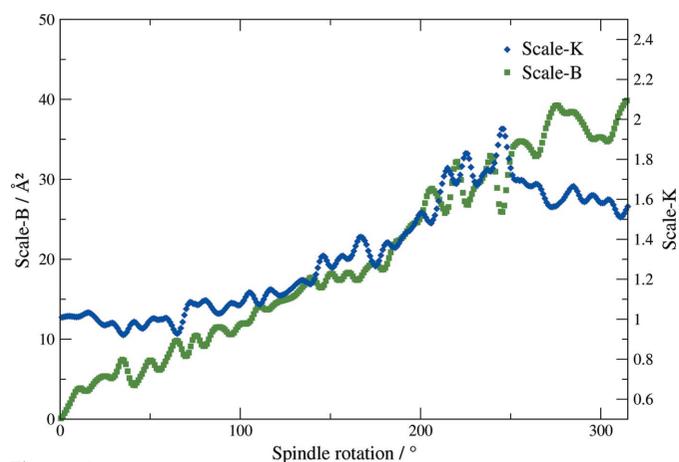
Figure 1
Ypa2: data reduction and scaling. The plot represents the overall (K) (in blue) and the resolution-dependent (B) (in green) scale factors computed by *SCALA* as a function of rotation axis (φ) position.

(Abrahams & Leslie, 1996) procedure as implemented in *autoSHARP* (Vonrhein & Bricogne, 2003).

Overall radiation damage was manifested in a continuous decrease of the resolution limit and mean I/σ_I of the diffraction data as a function of exposure time. Overall and resolution-dependent (pseudo B-factors) scale factors were refined at intervals of $5°$ in spindle rotation, with a smooth interpolation in between, in *SCALA* and are reported in Fig. 1. The steady increase of the scale-B factor is a clear sign of the unfolding of overall radiation damage (which affects the high-resolution reflections to a greater extent). Different data sets were created by excluding various amounts of the final data frames. Table 1 reports the data reduction statistics for data sets comprising data frames 1–90, data frames 1–180 and data frames 1–315 (*i.e.* all data). It is apparent that the inclusion of the final 135 data frames does not improve the data quality, but rather leads to some deterioration of the global quality indicators. Of greater concern is the fact that the anomalous signal strength (as gauged by, for example, the correlation in anomalous differences between half-sets or by the differences between the R_{meas} computed in the crystal point and Laue groups, respectively) also severely deteriorates if the final 135 data frames are included. This hints at the possibility of radiation damage which specifically affects the Se atoms. The detection of the Se atoms proved to be a strenuous task and was attempted simultaneously using the programs *SHELXD* and *HYSS*. A large number of attempts were run, applying different resolution cut-offs and using data sets including a varying number of data frames. Eventually, the program *HYSS* came up with a good solution using a data set that was obtained with data frames 1–180 only and applying a resolution cut-off at 4 Å. The initial refinement of the Se atoms was carried out against completely merged data sets (including various numbers of data frames). The data were then split up into three sequential parts, corresponding to data frames 1–90, 91–180 and 181–315. These chunks of data were declared as individual *Crystals* in the hierarchic organization of parameters in *SHARP* (La Fortelle & Bricogne, 1997), thus

Table 2
Ypa2: heavy-atom refinement in *SHARP*.

The values of the refined occupancy factors of the 20 Se atoms are reported. The occupancy factors were refined individually for the three data sets consisting of data frames 1–90, 91–180 and 181–315. Note that the data were not exactly on an absolute scale so that the absolute values have no physical meaning. All data sets were, however, on the same scale, so that all values can be compared with each other.

Site number	Images 1–90	Images 91–180	Images 181–315
1	0.46347	0.37085	0.32249
2	0.42322	0.36363	0.31286
3	0.49262	0.42141	0.32864
4	0.41715	0.32286	0.21827
5	0.39225	0.29345	0.24465
6	0.37068	0.31073	0.23729
7	0.40637	0.33640	0.27474
8	0.35734	0.27175	0.22183
9	0.50487	0.40258	0.35267
10	0.41929	0.34504	0.29237
11	0.43636	0.33418	0.24606
12	0.31231	0.24988	0.18412
13	0.31706	0.25897	0.22301
14	0.40403	0.31050	0.30399
15	0.43682	0.32955	0.28023
16	0.31491	0.22355	0.18202
17	0.42356	0.28764	0.25220
18	0.25299	0.19073	0.15841
19	0.29599	0.17050	0.10639
20	0.54296	0.49524	0.50047

allowing the refinement of individual occupancy factors for each Se atom in each of the three sets. For each atom the coordinates and B-factors were constrained to be identical in all three sets. The refined values for all Se sites displayed a noticeable trend of decreasing occupancies across the three sequential data sets (see Table 2), a feature which we ascribe to site-specific radiation damage. Finally, dose-dependent occupancy factors parametrized according to equations (5) and (6) were refined on an unmerged data set. Scale factors were also refined for chunks of 30 data frames. It turned out that the best phases were obtained by using data frames 1–180 only. Density modification using *SOLOMON* resulted in a map which allowed manual building of the structure model (see Fig. 2).

A closer analysis of the quality of phases (Fig. 3) reveals that a noticeable improvement is achieved if dose-dependent occupancy factors are modelled and refined on unmerged data, especially after density modification and for low-resolution reflections. Although the improvement in the quality of phases is apparent, it is by no means as pronounced as was the case in previous studies (Schiltz *et al.*, 2004). In the present case the phasing power largely stems from the anomalous (SAD) signal whereas the RIP part supplements useful additional phase information. The exact nature of the radiation damage to Se atoms remains unclear. In crystals of selenated nitroreductase, Ravelli *et al.* (2005) observed a pronounced radiation-induced loss of the anomalous signal from the Se atoms, but no clear indication of bond cleavages. During the refinement of the full Ypa2 structure against diffraction data merged from data frames 1–180, negative peaks showed up around the Se sites in difference Fourier $F_o − F_c$ maps and the
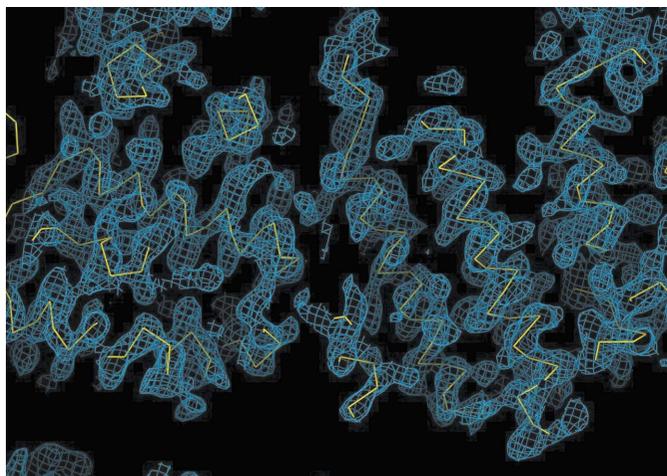
Figure 2
Ypa2: representative portion of the electron density map computed from RIP-aided SAD phases, after density modification. Superimposed on this map is the main-chain trace of the final refined molecule.
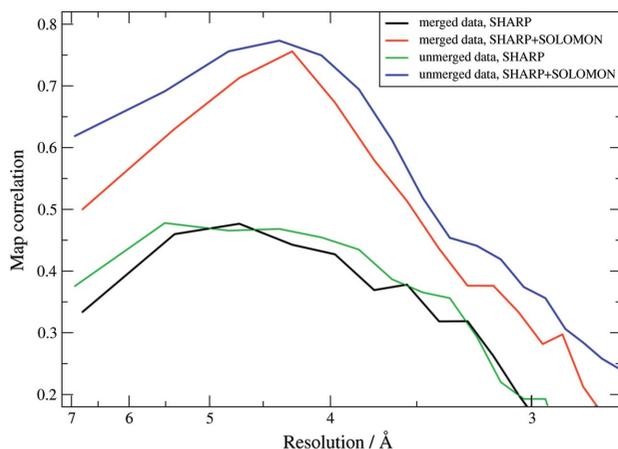

Figure 3
Ypa2: quality of the phases. The plots represent the correlation coefficients (as a function of resolution) of maps computed from experimental phases with respect to a map computed from the final refined structure. Black: SAD phases computed from merged data (images 1–180). Red: SAD phases computed from merged data (images 1–180), after density modification. Green: RIP-aided SAD phases computed from unmerged data (images 1–180). Blue: RIP-aided SAD phases computed from unmerged data (images 1–180), after density modification.

occupancy of these atoms had to be lowered (N. Leulliot, personal communication).

The difficulties involved in the detection of radiation-sensitive heavy atoms can turn out to be a limiting factor. As was mentioned above, many trials, using different programs and applying various cut-offs to the data, were necessary in order to solve the substructure. We did not manage to find the Se atoms using the data set which is composed of data frames 1–90 only. This is most likely due to the fact that the anomalous completeness is too low for this data set. It appears that the data set composed of data frames 1–180 has sufficient anomalous completeness, while still not being excessively radiation-damaged (at a resolution of 4 Å), for solving the substructure. Situations may, however, occur where it is

not possible to collect a sufficiently complete (anomalous completeness) but minimally radiation-damaged data set and in such cases direct methods and/or Patterson solvers are likely to fail in finding the substructure atoms.

This 'real life' case highlights some of the problems that may arise in SAD and/or RIPAS phasing and which have been pointed out by Ravelli *et al.* (2005). The anomalous signal of Se atoms appears to be highly susceptible to radiation damage, though the exact nature of the radiochemical modifications remains uncertain. The common prescription that a highly redundant data set increases the probability of success in SAD phasing (Usón *et al.*, 2003) has to be mitigated somewhat in cases where overall and/or site-specific radiation damage is important. The deleterious effects of site-specific radiation damage can be partly overcome if the time- or dose-evolution of structural changes are adequately modelled and if the data are kept unmerged. This approach is prone to generate additional RIP phase information. However, the quality of the phases is ultimately limited by overall radiation damage (which *cannot* generate phase information). In the present case, both site-specific and overall radiation damage evolve on similar time scales, in sharp contrast to the majority of model RIP case studies reported in the past. Even though the site-specific radiation damage was explicitly modelled here, the data contained in the final 135 frames did not generate additional phase information. The quality of the data in this final chunk of diffraction images was probably already too much impaired by overall radiation damage and this ultimately limits their usefulness for phasing.

4. Conclusion

Modelling and refining site-specific radiation damage of heavy atoms and/or anomalously scattering atoms has the potential to generate additional RIP phase information in SAD or MAD methods. The data need to be kept unmerged, down to the Harker construction where phasing power is generated through the intensity differences of symmetry-related reflections or repeated measurements of the same reflection recorded at different X-ray doses. This approach does not critically depend on high data redundancy, but a structural model describing the evolution of the damaged sites as a function of X-ray irradiation needs to be available. In the present communication we have described several models that have been implemented, or are currently in the process of being implemented, in the heavy-atom refinement and phasing program *SHARP* (La Fortelle & Bricogne, 1997; Bricogne *et al.*, 2003). In 'real life' situations, anomalous scattering experiments on selenated protein crystals may be corrupted by both site-specific and overall radiation damage. If properly modelled, site-specific radiation damage can be used to yield RIP phase information that is complementary to the phase information arising from the anomalous signal. However, the situation is more difficult in cases where overall radiation damage unfolds at a rate that is not significantly different in comparison with the evolution of site-specific radiation damage or in comparison with the total time that is required to

record a complete data set. In such cases the quality of the phases will ultimately be limited by the effects of overall radiation damage.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

Banumathi, S., Zwart, P. H., Ramagopal, U. A., Dauter, M. & Dauter, Z. (2004). *Acta Cryst.* D**60**, 1085–1093.

Berglund, G. I., Carlsson, G. H., Smith, A. T., Szöke, H., Henriksen, A. & Hajdu, J. (2002). *Nature (London)*, **417**, 463–468.

Blake, C. C. F. & Phillips, D. C. (1962). *Proceedings of the Symposium on the Biologiocal Effects of Ionizing Radiation at the Molecular Level*, pp. 183–191. IAEA, Vienna, Austria.

Bricogne, G. (2000). *Proceedings of the Workshop on Advanced Special Functions and Applications*, edited by D. Cocolicchio, G. Dattoli and H. M. Srivastava, pp. 315–321, Melfi (PZ), Italy, 9–12 May 1999. Rome: Arcane Editrice.

Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M. & Paciorek, W. (2003). *Acta Cryst.* D**59**, 2023–2030.

Burmeister, W. P. (2000). *Acta Cryst.* D**56**, 328–341.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Diederichs, K. (2006). *Acta Cryst.* D**62**, 96–101.

Diederichs, K., McSweeney, S. & Ravelli, R. G. B. (2003). *Acta Cryst.* D**59**, 903–909.

Donder, T. de (1920). *Leçons de Thermodynamique et de Chimie-Physique.* Paris: Gauthier-Villars.

Donder, T. de & Van Rysselberghe, P. (1936). *Affinity.* Menlo Park, CA: Stanford University Press.

Ennifar, E., Carpentier, P., Ferrer, J.-L., Walter, P. & Dumas, P. (2002). *Acta Cryst.* D**58**, 1262–1268.

Ennifar, E., Meyer, J. E. W., Buchholz, F., Stewart, A. F. & Suck, D. (2003). *Nucl. Acids Res.* **31**, 5449–5460.

Evans, G., Polentarutti, M., Djinovic Carugo, K. & Bricogne, G. (2003). *Acta Cryst.* D**59**, 1429–1434.

Evans, P. R. (1993). *Proceedings of CCP4 Study Weekend on Data Collection and Processing*, pp. 114–122. Warrington: SERC Daresbury Laboratory.

Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Acta Cryst.* D**59**, 1966–1973.

Harker, D. (1956). *Acta Cryst.* **9**, 1–9.

Hendrickson, W. A. (1976). *J. Mol. Biol.* **106**, 889–893.

La Fortelle, E. de & Bricogne, G. (1997). *Methods in Enzymology: Macromolecular Crystallography*, edited by R. M. Sweet and C. W. Carter Jr, Vol. 276, pp. 472–494. New York: Academic Press.

Leslie, A. G. W. (1993). *Proceedings of CCP4 Study Weekend on Data Collection and Processing*, pp. 44–51. Warrington: SERC Daresbury Laboratory.

Leulliot, N., Vicentini, G., Jordene, J., Cheruel, S., Liger, D., Schiltz, M., Van Tilbeurgh, H., Barford, D. & Goris, J. (2006). *Mol. Cell*, **23**, 413–424.

Nanao, M. & Ravelli, R. B. G. (2006). *Structure*, **14**, 791–800.

Nanao, M. H., Sheldrick, G. M. & Ravelli, R. B. G. (2005). *Acta Cryst.* D**61**, 1227–1237.

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

Pannu, N. S., McCoy, A. J. & Read, R. J. (2003). *Acta Cryst.* D**59**, 1801–1808.

Penner-Hahn, J. E., Murata, M., Hodgson, K. O. & Freeman, H. C. (1989). *Inorg. Chem.* **28**, 1826–1832.

Ramagopal, U. A., Dauter, Z., Thirumuruhan, R., Fedorov, E. & Almo, S. C. (2005). *Acta Cryst.* D**61**, 1289–1298.

Ravelli, R. B. G. & McSweeney, S. M. (2000). *Structure*, **8**, 315–328.

Ravelli, R. B. G., Nanao, M. H., Lovering, A., White, S. & McSweeney, S. (2005). *J. Synchrotron Rad.* **12**, 276–284.

Ravelli, R. B. G., Schrøder Leiros, H.-K., Pan, B., Caffrey, M. & McSweeney, S. M. (2003). *Structure*, **11**, 217–224.

Rice, L. M., Earnest, T. N. & Brunger, A. T. (2000). *Acta Cryst.* D**56**, 1413–1420.

Schiltz, M., Dumas, P., Ennifar, E., Flensburg, C., Paciorek, W., Vonrhein, C. & Bricogne, G. (2004). *Acta Cryst.* D**60**, 1024–1031.

Schlichting, I., Berendzen, J., Chu, K., Stock, A. M., Maves, S. A., Benson, D. E., Sweet, R. M., Ringe, D., Petsko, G. A. & Sligar, S. G. (2000). *Science*, **287**, 1615–1622.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.

Sygusch, J. & Allaire, M. (1988). *Acta Cryst.* A**44**, 443–448.

Usón, I., Schmidt, B., von Bülow, R., Grimme, S., von Figura, K., Dauter, M., Rajashankar, K. R., Dauter, Z. & Sheldrick, G. M. (2003). *Acta Cryst.* D**59**, 57–66.

Vonrhein, C. & Bricogne, G. (2003). *autoSHARP. An Automated Structure Determination System.* Version 3.0.15. Global Phasing, Cambridge, UK.

Weik, M., Bergès, J., Raves, M. L., Gros, P., McSweeney, S., Silman, I., Sussman, J. L., Houée-Levin, C. & Ravelli, R. B. G. (2002). *J. Synchrotron Rad.* **9**, 342–346.

Weik, M., Ravelli, R. B. G., 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.

Weiss, M. S., Mander, G., Hedderich, R., Diederichs, K., Ermler, U. & Warkentin, E. (2004). *Acta Cryst.* D**60**, 686–695.

Yano, J., Kern, J., Irrgang, K.-D., Latimer, M. J., Bergmann, U., Glatzel, P., Pushkar, Y., Biesiadka, J., Loll, B., Sauer, K., Messinger, J., Zouni, A. & Yachandra, V. K. (2005). *Proc. Natl. Acad. Sci. USA*, **102**, 12047–12052.

Zwart, P. H., Banumathi, S., Dauter, M. & Dauter, Z. (2004). *Acta Cryst.* D**60**, 1958–1963.