Journal of
Synchrotron
Radiation

ISSN 0909-0495

Received 18 September 2012
Accepted 27 November 2012

Identification of patterns in diffraction intensities affected by radiation exposure

Dominika Borek,[a,b] Zbigniew Dauter[c] and Zbyszek Otwinowski[a,b]*

[a]Department of Biophysics, UT Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd, Dallas, TX 75390, USA, [b]Department of Biochemistry, UT Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd, Dallas, TX 75390, USA, and [c]Macromolecular Crystallography Laboratory, Synchrotron Radiation Research Section, National Cancer Institute, Argonne National Laboratory, Bioscience Division, 9700 South Cass Avenue, Argonne, IL 60439, USA.
E-mail: zbyszek@work.swmed.edu

In an X-ray diffraction experiment, the structure of molecules and the crystal lattice changes owing to chemical reactions and physical processes induced by the absorption of X-ray photons. These structural changes alter structure factors, affecting the scaling and merging of data collected at different absorbed doses. Many crystallographic procedures rely on the analysis of consistency between symmetry-equivalent reflections, so failure to account for the drift of their intensities hinders the structure solution and the interpretation of structural results. The building of a conceptual model of radiation-induced changes in macromolecular crystals is the first step in the process of correcting for radiation-induced inconsistencies in diffraction data. Here the complexity of radiation-induced changes in real and reciprocal space is analysed using matrix singular value decomposition applied to multiple complete datasets obtained from single crystals. The model consists of a resolution-dependent decay correction and a uniform-per-unique-reflection term modelling specific radiation-induced changes. This model is typically sufficient to explain radiation-induced effects observed in diffraction intensities. This analysis will guide the parameterization of the model, enabling its use in subsequent crystallographic calculations.

Keywords: radiation damage; matrix singular value decomposition; experimental phasing; radiolysis.

1. Introduction

Chemical reactions originating from the absorption of X-ray photons in a crystal influence the diffraction intensities. Chemical and physical properties of the crystallized molecules (Ravelli & McSweeney, 2000; Borek *et al.*, 2007, 2010; Burmeister, 2000), the type and concentration of chemical compounds in a particular crystallization solution (Borek *et al.*, 2010; Paithankar & Garman, 2010; Paithankar *et al.*, 2009; Garman & Nave, 2009; Holton, 2007; Murray *et al.*, 2005) and, potentially, crystal packing (Warkentin *et al.*, 2012*a*) modulate the outcomes of these interactions.

These processes start with the absorption of X-ray photons. Every absorbed X-ray photon results in hundreds of secondary ionizations producing excited states that can migrate by a combination of diffusion, temperature-dependent hopping or temperature-independent tunnelling (O'Neill *et al.*, 2002; Gray & Winkler, 1996, 2009). Their migration may end either in a recombination, including a geminate recombination, in radical-induced reactions altering the protein structure, or in water radiolysis (Terryn *et al.*, 2005). Therefore, measured diffraction intensities represent states that have already been altered by radiation-induced processes, and they need to be corrected during data analysis to recover the signal representing the original structure.

Here, a physical model of radiation-induced changes in diffraction is discussed as the first and essential step in developing computational corrections for radiation-induced changes at the level of diffraction data. Historically, a large part of radiation-induced effects has been corrected by scaling procedures during the data reduction step (Fox & Holmes, 1966; Arnott & Wonacott, 1966; Otwinowski *et al.*, 2003; Otwinowski & Minor, 1997; Evans, 2006, 2011). The rationale for them and what type of further corrections are needed (Diederichs, 2006; Diederichs *et al.*, 2003; Borek *et al.*, 2010) will be explained. An approach based on singular value decomposition of multiple datasets allows the validation of assumptions about the data model and the identification of the additional complexity that can be considered in computational analysis.

doi:10.1107/S0909049512048807

1.1. Radiation-induced effects in diffraction data and their models

All macromolecular crystals are built from molecules of similar composition, *i.e.* containing predominantly oxygen, carbon, nitrogen and hydrogen atoms, with a frequent presence of a smaller number of sulfur or phosphorus atoms, and occasionally atoms of heavier elements. In terms of radiation physics, the crystallization and cryo-stabilization solutions are of as equal importance as the macromolecules being studied (Itikawa & Mason, 2005; Paithankar & Garman, 2010; Paithankar *et al.*, 2009; Fütterer *et al.*, 2008; Borek *et al.*, 2007). X-ray photons interact with atoms in a crystal lattice randomly, with the probability of an interaction defined by the absorption cross section of the particular atom type. Each primary interaction between a photon and an atom generates hundreds of secondary ionizations, within a typical radius of ~10000 Å (~1 µm) (Timneanu *et al.*, 2004; Sanishvili *et al.*, 2011). Within the crystal lattice, sites of secondary ionizations are not correlated with the location of the primary event, so changes in atom positions due to these events are also randomly distributed. The practical limit for the radiation dose to a cryo-cooled crystal of about 20 to 30 MGy (Murray *et al.*, 2005; Owen *et al.*, 2006; Henderson, 1995) corresponds to about 2 eV per atom (for a dose of 20 MGy). For comparison, the ionization energy of water is 12.6 eV (Dean, 1992). Therefore, at a dose of 20 MGy, a dense network of localized changes in the molecules building the crystal lattice would be expected to be observed. At room temperature, covalent changes may unfold the protein and create long-range effects with complex kinetics (Hendrickson, 1976). For this reason, we focus here on currently more typical cryo-cooled conditions, where the vitrified state of the crystal immobilizes products of radiolysis, preventing long-range reorganizations by thermal diffusion (Terryn *et al.*, 2005). However, at cryo-temperatures, radicals can migrate for substantial distances, up to tens of Ångströms, so the diffusion of charged radicals will be influenced by the atomic electric fields (Gray & Winkler, 2005, 2010; Tezcan *et al.*, 2001). Rearrangements of the covalent structure of macromolecules or water radiolysis will necessarily shift the positions of the atoms involved. Also, each covalent change is expected to be accompanied by smaller shifts of multiple atoms in the neighbourhood.

The first component of the model presented here considers the impact of an average positional displacement of atoms on the diffraction intensity of Bragg peaks. It is assumed that atomic displacements resulting from interactions with X-ray photons are frequent, have small magnitudes in the range of fractions of Ångströms, and are randomly distributed, *i.e.* typically there is no correlation between the displacements in different unit cells of the crystal lattice. With these assumptions, by the central limit theorem, the radiation-induced repositioning of atoms in the unit cell can be approximated as a convolution of their initial positions with a Gaussian function.

As individual displacements accumulate at a rate proportional to the dose, by the same central limit theorem, it is expected that the second moment, σ^2, of the displacement will linearly increase with the accumulation of changes,

$$\begin{aligned} P\left(x_{i,0}, x_{i,d}\right) &\simeq \exp\left[-\left(x_{i,0} - x_{i,d}\right)^2/2cd\right], \\ \rho(d) &\simeq \exp\left(-\Delta x^2/2cd\right) * \rho(0), \end{aligned} \tag{1}$$

where the probability of displacement P from the initial position $x_{i,0}$ to the position $x_{i,d}$ at dose d is a Gaussian function with width $\sigma = (cd)^{1/2}$, where the dose d is scaled by a constant c. For an initial electron density $\rho(0)$, when convolved (*) with this Gaussian function, the electron density $\rho(d)$ at dose d is obtained.

Together, all these random displacements affect diffraction intensities. By applying a Fourier transform to the electron density $\rho(d)$ and squaring the amplitude $\mathbf{F}_h$, with the diffraction vector $\mathbf{S}_h$, we obtain the intensity decay formula, which quantifies the first component of our model. In this formula we use units of displacement squared, scaled up by $8\pi^2$, so that the displacement is expressed in units of the atomic displacement parameter B, customary in macromolecular crystallography,

$$\begin{aligned} \mathcal{F}[\rho(d)] &= \mathcal{F}[\rho(0)] \exp\left[-(2\pi)^2 cd\left|\mathbf{S}_h\right|^2/2\right], \\ I_h(d) &= I_h(0) \exp\left[-2(2\pi)^2 cd\left|\mathbf{S}_h\right|^2/2\right], \\ I_h(d) &= I_h(0) \exp\left(-B_d\left|\mathbf{S}_h\right|^2/2\right). \end{aligned} \tag{2}$$

The dose d is frequently not well known due to the rotation of larger-than-the-beam crystals, due to non-uniformity of the beam profile or due to lack of an absolute beam calibration at the crystal position. However, the dose can be determined from the scaling B factor, B_d, using the relationship $B_d = 8\pi^2 cd$, with the constant $8\pi^2 c$ estimated to be about 1 Å^2 MGy^{-1} at 100 K (Kmetko *et al.*, 2006; Borek *et al.*, 2007; Krojer & von Delft, 2011). This parameterization of the decay correction has a very long history in macromolecular crystallography (Arnott & Wonacott, 1966; Fox & Holmes, 1966; Otwinowski & Minor, 1997; Otwinowski & Minor, 2000; Evans, 2006). In practice, the decay correction is often modelled together with other multiplicative factors during scaling. Depending on the specific conditions and software used, the separation of the decay of diffraction intensities from other multiplicative effects may be difficult. The second contributor to the radiation damage model presented here results from changes in electron density localized at specific points of the structure, including the ordered solvent. The real-space modelling of these effects (Schiltz & Bricogne, 2007; Schiltz *et al.*, 2004; Warkentin *et al.*, 2012*b*) does not address the issue of how to account for them in the process of merging multiple observations. For the data space approach of so-called zero-dose extrapolation, a specific data model of the radiation-induced changes was chosen (Diederichs *et al.*, 2003; Diederichs, 2006). However, this was without testing the optimality of the model. Here, consecutive datasets are evaluated, with each series collected from a single crystal, to identify patterns of specific changes in data during the exposure, and these patterns are analyzed by singular value decomposition to investigate whether or not the physical two-component model of radiation-induced processes described above can explain the differences observed in diffraction intensities. The impact of localized radiation effects can only be interpreted using data already corrected for overall decay. Therefore, the second contributor will be called ‘specific radiation changes’ when talking about data corrected for decay. This name emphasizes that this component represents discrete covalent rearrangements, distinguished from decay of diffraction intensities resulting from random small displacements distributed uniformly within the structure.

1.2. Singular value decomposition

A generic approach to deduce relationships among multiple datasets is to combine them into a large matrix and then subject it to a decomposition analysis. These calculations serve two main functions: (i) they can simplify calculations on large matrices to a set of reduced operations that are calculated in a faster and/or more stable manner, and (ii) they help to reveal characteristic patterns in data, which may be difficult to notice otherwise due to the size of the matrices or because of the presence of experimental errors.

Matrix singular value decomposition (SVD) is one of the most general and useful methods of decomposition and is used widely in data analysis and data mining (Stewart, 1993; Alter *et al.*, 2000). In SVD, an $m \times n$ real or complex matrix $\mathbf{A}$ is decomposed into $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\rm T}$, where $\mathbf{U}$ is an $m \times m$ real or complex unitary matrix whose columns are called left singular vectors of $\mathbf{A}$, $\boldsymbol{\Sigma}$ is an $m \times n$ diagonal matrix with non-negative entries σ_{ii}, called singular values of $\mathbf{A}$, ordered on the diagonal with decreasing magnitude, and $\mathbf{V}^{\rm T}$ is a conjugate transpose of $\mathbf{V}$, which is an $n \times n$ real or complex unitary matrix whose columns are called right singular vectors of $\mathbf{A}$.

When SVD is applied to a matrix, the result can be expressed as a weighted sum of separable matrices. A separable matrix is an outer product of right and left singular vectors, and a corresponding singular value σ_{ii}. Singular values weigh contributions of separable matrices to the matrix being decomposed. Separable matrices frequently have a physical interpretation and the goal in this work was to associate them with the physical processes that are part of our model of radiation-induced changes.

SVD is related conceptually to principal component analysis (PCA), and in some specific cases, for instance in the calculations discussed here later, it is equivalent to it (Wall *et al.*, 2003). In PCA, the eigenvectors and eigenvalues of the covariance matrix, which by definition is square and symmetric, are analysed. To apply PCA, the covariance matrix is calculated by squaring the matrix that is to be decomposed: $\mathbf{C} = \mathbf{A}\mathbf{A}^{\rm T}$. PCA performed on $\mathbf{C}$ will yield the principal components (eigenvectors), which are the same as the right singular vectors of SVD. The eigenvalues of $\mathbf{C}$ are equal to $(\sigma_{ii})^2$ of SVD, and are proportional to variances of the principal components.

SVD was previously used in crystallography to analyze time-resolved series of diffraction data (Rajagopal *et al.*, 2004*a*,*b*; Schmidt *et al.*, 2003; Romo *et al.*, 1995). The similarities and differences between the method used in these studies and that reported in this work are discussed in §2.2.

2. Methods

2.1. Crystals and diffraction data

All proteins were crystallized by a vapour diffusion method.

Thaumatin solution of 36 mg ml^{-1} in 29 m*M* HEPES, pH 7.0, and 10 m*M* $\rm CaCl_2$ was mixed 1:1 with the well solution containing 0.75 *M* KNa tartrate, 0.1 *M* citrate buffer, pH 6.5, and 10% (*v*/*v*) glycerol. A bipyramidal crystal of dimensions 0.15 mm × 0.15 mm × 0.25 mm was cryo-protected by dipping it for a few seconds in the well solution supplemented by 27% (*v*/*v*) of glycerol.

Crystals of thermolysin were obtained from a 90 mg ml^{-1} solution of protein in 45% DMSO after mixing it 2:1 with the well solution of 1.4 *M* Ca acetate in 0.1 *M* Tris buffer, pH 7.2. The cryoprotecting solution was made up of well solution supplemented with 25% (*v*/*v*) glycerol. Crystals were elongated, with dimensions of about 0.07 mm × 0.06 mm × 0.3 mm.

Crystals of elastase grew from 15 mg ml^{-1} protein solution in water mixed 1:1 with the well solution of 0.1 *M* Na acetate buffer, pH 5.0, containing 0.2 *M* Na citrate and 0.05 *M* $\rm CaCl_2$. They were derivatized by soaking in the well solution containing 5 m*M* of $\rm KAu(CN)_2$ or $\rm K_2PtCl_6$ for about 4 h and cryo-preserved by dipping shortly in the same solution supplemented with glycerol to 25% (*v*/*v*). The brick-shaped crystals had dimensions of approximately 0.1 mm × 0.12 mm × 0.15 mm.

Crystals of Lon protease as a SeMet variant were obtained as described earlier (Botos *et al.*, 2004). The data were measured from a crystal of dimensions 0.15 mm × 0.15 mm × 0.35 mm.

All diffraction data were measured at beamline X9B at the National Synchrotron Light Source, Brookhaven National Laboratory, USA, using the ADSC Quantum 4CCD detector and crystals cooled at 100 K with the Oxford Cryosystems device.

The datasets were measured successively from selected crystals of the above proteins by repeating identical collection protocols and conditions. In the following, the number of datasets used in the analysis is reported in parentheses after the number actually collected. 22 (22) successive complete datasets were collected from one crystal of thaumatin, 11 (9) sets from one crystal of thermolysin, 20 (19) for Lon protease, 20 (18) from the platinum derivative of elastase, and 26 (25) sets from the gold derivative of the same protein. The statistics of the obtained datasets are summarized in Table 1.

2.2. Validation of the data model by SVD

The goal was to reveal patterns of differences due to radiation damage by data mining multiple datasets collected from a single crystal. Each individual dataset was integrated and scaled with *HKL2000* (Otwinowski & Minor, 1997). Then,

Table 1
Data-related statistical indicators for the first and the last diffraction datasets scaled separately.

There are no signs of decreased data quality, even though the radiation-induced changes for each group of the datasets are quite significant at the end of data collection. $B_{\rm rel}$ describes the increase of the scaling B-factor across all datasets scaled together as described in the text.

	Thaumatin	Thermolysin	Lon protease	Elastase-Au	Elastase-Pt
Space group	$P4_12_12$	$P6_122$	$P3_1$	$P2_12_12_1$	$P2_12_12_1$
Unit-cell parameters (Å)	$a = b = 57.75$, $c = 150.11$	$a = b = 92.63$, $c = 128.43$	$a = b = 86.70$, $c = 128.53$	$a = 49.91$, $b = 57.83$, $c = 74.41$	$a = 50.12$, $b = 57.70$, $c = 74.37$
Resolution (Å)	1.45	1.45	2.30†	1.21	1.3
Number of datasets collected/used	22/22	9/9	20/19	26/25	20/18
Completeness (%) first/last dataset	99.0/99.0	99.4/99.2	98.4/99.5	96.2/95.6	97.6/97.8
$\langle I\rangle/\langle\sigma_I\rangle$ for unmerged data, first/last dataset‡	40.4/31.9	33.1/32.0	22.8/21.4	25.5/24.8	25.8/18.9
$R_{\rm merge}$					
Range between consecutive datasets (%)	0.8–1.2	1.3–2.8	1.0–2.0	2.0–3.9	2.9–4.3
First/last dataset (%)	4.3/6.3	4.0/4.5	2.0/2.5	3.4/3.5	3.1/4.2
Between first and last dataset (%)	9.1	7.6	13.3	11.1	7.9
$B_{\rm rel}$ (Å^2)	2.9	3.3	16.8	2.9	1.7

† Data for Lon protease initially extended to 1.75 Å, and to about 2.00 Å by the end of data collection. However, owing to the presence of ice rings, datasets were scaled to a resolution of 2.30 Å. The first dataset collected was omitted in the analysis due to beam instability issues during the collection of this dataset, and the second dataset had some missed frames, so its completeness is lower than the subsequent ones. ‡ The intensity of the X-ray beam fluctuated during data collection. Therefore, the changes in $\langle I\rangle/\langle\sigma_I\rangle$ between the first and the last dataset in each series do not represent the impact of radiation decay. $B_{\rm rel}$ values are resistant to these fluctuations and they should be used as an indicator of the overall decay due to X-ray exposure.

for each protein, multiple datasets were scaled together to correct for beam intensity differences and overall crystal decay. This decay combines resolution-dependent and resolution-independent components as implemented in *Scalepack* (Otwinowski *et al.*, 2003; Otwinowski & Minor, 1997, 2000; Fox & Holmes, 1966). The correction is the first part of our data model for radiation-induced changes in diffraction intensities, and, after applying it, further analysis of the data can be carried out to determine the presence of other effects.

SVD can be performed on any uniform data series. In crystallography, SVD can be applied either in the results space, *i.e.* electron densities, or in the data space, *i.e.* diffraction intensities. The main interest here is to differentiate random experimental errors arising in the data space from systematic structural effects, in particular those resulting from specific radiation damage. The availability of multiple datasets from the same crystal, or, alternatively, from nearly isomorphous crystals, allows for a model-free approach to the interpretation of the sources of variations in the data series. SVD can be performed on the original data or on the data transformed in such a way that variations in SVD matrix elements are more uniform. More uniform variations of the matrix elements provide benefits in SVD-based analyses. SVD performed on data with uniform variations will be more informative, because the results will be more uniformly affected by each matrix element. Therefore, to achieve more uniform variations, a conceptual analysis of the effects of radiation damage in real space is first carried out in the context of multiple diffraction datasets. Since in the approach presented here SVD is equivalent to PCA (§1.2), the covariance matrix between electron densities corresponding to diffraction data acquired at different dose points is first defined as:

$$\mathbf{C}_{i,j} = \int_V \rho_i\,\rho_j, \tag{3}$$

where indices i and j correspond to particular dose points, and ρ_i, ρ_j are electron densities calculated over the whole volume of the unit cell V assuming that $\mathbf{F}_{000} = 0$. This assumption results in the average electron densities being equal to zero, which applies centering to data in a SVD/PCA sense. The elements of the covariance matrix are scalar products of electron densities, so, by the unitary property of the Fourier transform, they can be calculated in reciprocal space as scalar products of the corresponding structure factor sets,

$$\mathbf{C}_{i,j} = \sum_h \mathbf{F}_i\,\mathbf{F}_j^*, \tag{4}$$

where h is a unique index.

An additional refinement to this approach is to make contributors to the matrix elements more uniform by applying resolution normalization to the structure factors, *i.e.* replace structure factors with normalized structure factors. The matrix can also be scaled so that the diagonal elements are equal to 1; it then becomes a correlation matrix. For nearly isomorphous datasets, most of the entries of the correlation matrix will be close to 1 and non-isomorphism between datasets will be indicated by departure from 1 of the corresponding non-diagonal element. The amplitude of structure factors can be calculated directly from the scaled observed intensities by taking the square root of them. While formally the phases should be included in SVD, an equivalent result can be obtained without doing this, as explained below. If needed, after solving the structure, just one set of reference phases can be used to interpret the results in real space.

In nearly isomorphous datasets, the main component of the signal in real space corresponds to the average structure, for which its Fourier representation will be referred to as a set of parent structure factors $\mathbf{F}_h^{\rm P}$. The differences between sets of scaled structure factors, $\mathbf{F}_h^i - \mathbf{F}_h^j$, will be orthogonal to this main component $\mathbf{F}_h^{\rm P}$. It can generally be assumed that the phases of $\mathbf{F}_h^i - \mathbf{F}_h^j$ differences are uncorrelated with the phases of the parent structure factors $\mathbf{F}_h^{\rm P}$. However, this assumption may not be valid for rare cases, for instance if the diffraction power is dominated by a simple heavy-atom substructure, such as a single heavy-atom cluster.

To avoid potential problems arising from phase bias, it is preferable to perform the calculations without including knowledge of the phases of the $\mathbf{F}_h^i - \mathbf{F}_h^j$ differences. Therefore, the $\mathbf{F}_h^i - \mathbf{F}_h^j$ differences are conceptually separated into the part that affects only the phase of $\mathbf{F}_h^{\rm P}$ when added to it, and the part that affects only the amplitude of $\mathbf{F}_h^{\rm P}$ when added to it. To achieve this, the difference vector $\mathbf{F}_h^i - \mathbf{F}_h^j$ in the complex plane is projected onto the parent structure factor $\mathbf{F}_h^{\rm P}$ for each h. Two vectors are obtained for each structure factor: one vector $(\mathbf{F}_h^i - \mathbf{F}_h^j)_{\parallel}$ that is parallel to the $\mathbf{F}_h^{\rm P}$ and a second $(\mathbf{F}_h^i - \mathbf{F}_h^j)_{\perp}$ that is perpendicular to the $\mathbf{F}_h^{\rm P}$. Owing to the lack of correlation between the phases of the parent structure factors $\mathbf{F}_h^{\rm P}$ and those of the differences, the norms of the projected parallel and perpendicular vectors summed over all h will have about the same value, $(\mathbf{F}_h^i - \mathbf{F}_h^j)_{\parallel} = |\mathbf{F}_h^i - \mathbf{F}_h^j|\cos(2\pi(\varphi_h^{\rm P} - \varphi_h^{i-j})] = |\mathbf{F}_h^i| - |\mathbf{F}_h^j|$, so these values can be calculated directly from measured intensities. The sum of squared norms of perpendicular vectors is about the same as the sum of squared norms of parallel vectors so, if the contribution from the perpendicular vectors is ignored, the value calculated directly from intensities will underestimate the true number by a factor of two. This underestimation does not affect the eigenvector structure in the analysis. It has a very regular impact on the correlation matrix and affects only its non-diagonal elements. The correlation coefficient calculated between two sets of perfectly isomorphous complex structure factors is equal to 1. When differences between nearly isomorphous structures are calculated and the $(\mathbf{F}_h^i - \mathbf{F}_h^j)_{\perp}$ component is ignored, non-diagonal elements of the correlation matrix represent the arithmetic average between the true value of the correlation coefficient and the number 1, which represents no difference, the perfect correlation between structure factors. Such decomposition of the covariance matrix scales down all eigenvalues other than the largest one by a factor of two, while preserving the eigenvector structure. Measurement errors increase the difference between non-diagonal correlation matrix elements and the perfect correlation coefficient. Therefore, errors may result in additional eigenvalue(s) of a small magnitude. The sum of small eigenvalues that cannot be interpreted as systematic effects describes the level of random error.

Generally, the number of datasets is expected to be smaller than the number of electron density grid points or the number of unique reflections. Therefore, the shorter dimension of the matrix being decomposed will correspond to the dataset index and right singular vector. The longer dimension and left singular vector will correspond to the electron density, or its Fourier representation: a set of structure factors. For a left singular vector, its elements can be combined with model phases $\exp(i2\pi\varphi_h^{\rm P})$ to produce a set of structure factors, referred to from now on as a left singular vector with phases, from which the electron density map can be calculated. For the dominant singular value, the left singular vector with phases represents the parent structure $\mathbf{F}_h^{\rm P}$. For nearly isomorphous data, elements of the corresponding right singular vector will all have the same value. Consequently, due to the orthogonality of singular vectors, for all other singular values, right singular vectors will have the average value of their elements equal to zero. As a result, the corresponding left singular vectors with phases will represent coefficients of difference Fourier maps. This is a simple generalization of the concept of difference Fourier maps calculated from two datasets, now extended to multiple datasets and representing various types of structural variation identified by SVD decomposition. To interpret SVD, these difference maps were calculated and the coefficients of right singular vectors plotted.

The right singular vector describing the specific radiation-induced changes is expected to have a characteristic dependence on the dose, which as discussed before is correlated with the scaling B-factor. However, when presenting the results, adjustments must be made to compensate for singular vectors being orthogonal, normalized and having an arbitrary sign. The need for such compensation arises from the fact that the right singular vector describing specific radiation damage is defined with respect to the average dose, and on an arbitrary scale that can be negative. Therefore, when comparing this singular vector(s) with the scaling B-factor, which is defined with respect to the first dataset, the singular vector elements are offset so that the first element equals zero, and, if needed, the sign is also reversed.

Another component of the variations in the measured intensities may arise from anomalous scattering. To identify and estimate it using SVD, only the acentric reflections are analysed. The Friedel pairs were split into two separate vectors of the matrix to be analyzed by SVD, separately for each scan in a multiwavelength dataset. For a single type of anomalous scatterer, the Bijvoet differences at different wavelengths are fully correlated, *i.e.* their values are related by a scaling factor proportional to f''. Therefore, in this typical case, it was expected that only one singular value would be obtained corresponding to Bijvoet differences and defining the extent of the anomalous signal. For a multi-wavelength dataset, with anomalous signal arising from different types of anomalous scatterers, SVD analysis may produce an additional singular value describing Bijvoet differences, since different datasets are no longer fully correlated. In a multi-wavelength dataset, SVD can also identify a component corresponding to dispersive differences. Its absence in the data may indicate that the preferred phasing strategy would be to use the Bijvoet differences component across multiple datasets to create a pseudo-SAD dataset. To achieve it, average datasets can be acquired at different wavelengths, while applying the wavelength-dependent scaling factor to Bijvoet differences.

Multi-dataset SVD analysis has been performed previously on time-resolved data series collected using the Laue method (Romo *et al.*, 1995; Schmidt *et al.*, 2003; Rajagopal *et al.*, 2004*a*,*b*). As such datasets are incomplete, the authors applied real-space SVD analysis, restricted to a mask where the electron-density changes were the largest. Such restriction is a form of data filtering, and so reduces the effect of noise or incomplete data on the analysis. A future extension of the

SVD method proposed here may apply a similar filtering in real space, for instance to determine the decay components in the anomalous signal by restricting the SVD matrix to the neighbourhood of heavy-atom sites.

3. Results and discussion

Our model of radiation-induced effects relies on decomposition of signals in data space which corresponds to reciprocal space in crystallography. The model consists of two components: the radiation-induced decay of diffraction intensities and the specific radiation-induced changes. The decay is modelled by the scaling B-factor with a very small number of parameters for all the data. In principle, if the dose is accurately known, just one overall parameter scaling it to the B-factor would be sufficient. In practice, *Scalepack* uses two parameters per dataset to characterize decay (Otwinowski *et al.*, 2003; Otwinowski & Minor, 1997, 2000). The specific radiation-induced change requires at least one parameter per unique h index, which complicates the analysis by introducing a large number of parameters. In reciprocal space, the specific radiation-induced change can be considered as drift from the scaled, thus decay-corrected, values of initial intensities. Traditionally, the decay is modelled in reciprocal space, whereas the specific radiation changes, while calculated in reciprocal space, are interpreted in real space. However, it is important to remember that in a crystal these effects happen together and their decomposition into two components is performed to simplify the data analysis and the interpretation of results.

Recently, a different decomposition was proposed, based on structure refinement of datasets collected with incremental radiation dose (Warkentin *et al.*, 2012*b*). The authors observed very regular patterns of temperature factor increase with dose for individual residues. This regularity is consistent with the results of our SVD analysis, where one or two singular values are sufficient to describe the radiation-induced changes. They also observed that different residues had different rates of temperature factor increase. This variation is not surprising, as the migration of free radicals resulting in radiolysis and its structural consequences are likely to be affected by electrostatics, and differ somewhat between the inside and the surface of the protein. The authors defined uniform radiation damage by the slowest rate of B-factor increase, to constitute what they call a non-uniform component, to have only a positive rate of change. While their definition is self-consistent, it is of little relevance to the issues encountered during data reduction. From the data analysis perspective, the goal of splitting the diffraction consequences of radiation dose is to minimize the norm (or root mean square) of specific radiation damage. This is accomplished by factoring out the decay, a direct consequence of the average value of B-factor increase in the structure rather than its minimal value.

The properties of macromolecular crystals such as unit-cell size, crystal size, microscopic order and solvent content, in combination with unavoidable intensity decay due to X-ray exposure, define the limit of the achievable number of diffracted photons [reviewed by Holton & Frankel (2010)]. Intense synchrotron sources allow for data collection up to this limit. However, the decay of intensities is accompanied by the simultaneous presence of specific radiation-induced changes, which have variable magnitude for different macromolecules even at the same dose. This variability at the level of radiation-induced specific changes frequently leads to overcautious data collection strategies, which focus on minimizing the X-ray exposure, and result in a signal-to-noise ratio that is too low to achieve successful experimental phasing. To fully utilize crystal scattering power, there is a need to improve computational approaches based on understanding of the complexity and magnitude of the radiation-induced specific changes.

3.1. Data reduction and SVD

The simplest data model for multiple observations per unique reflection h uses one structure factor amplitude and its uncertainty. This can be built upon to describe more effects. The first extension separates Friedel mates in analysis, resulting in two experimental signals and their uncertainties. More contributors to changes of a structure factor during an experiment can be included, for instance those due to radiation damage, the changes resulting from using different wavelengths, intentionally or unintentionally introduced non-isomorphisms between crystals, and violations of rotational symmetry, *etc.* However, Ockham's razor should be applied when deciding how many parameters per unique reflection h are needed to model the results.

Frequently, scaled and unmerged data are considered more informative than scaled and merged observations. This is only true if a significant contributor to data variability was omitted during the merging analysis. SVD guides the selection of significant parameters by objectively identifying the most important types of variability in the data. As in many other applications of SVD to data filtering, the analysis should only include those components that are significantly above the noise level. In such a case, merged data will contain information equivalent to the unmerged data with respect to the signals; however, to fully describe their uncertainties, a matrix describing correlations between errors is required.

3.2. Linearity of radiation-induced changes of diffraction intensities

Here groups of datasets from five crystals were tested, one for thaumatin, one for thermolysin, one for Lon protease and two for elastase. The thaumatin and thermolysin data series were collected from native crystals, Lon protease data were acquired from a SeMet derivative at the selenium absorption edge, while elastase was separately derivatized with two heavy-atom compounds, $KAu(CN)_2$ and $KPtCl_6$. In the elastase crystals, two gold atom sites had a high level of substitution and a larger number of platinum sites had only partial occupancies. Datasets for each crystal were collected at one wavelength. Individual datasets within each group were exposed with substantial variation of dose per dataset; however, this has no impact on the SVD analysis.

To answer the question of how complex the radiation-induced changes in diffraction intensities are, singular values and singular vectors obtained by SVD were analysed. For SVD performed with Friedel pairs merged, the second singular value always corresponded to specific radiation changes. All subsequent singular values are not significantly above the noise level, with the exception of thaumatin and Lon protease, where the much lower third singular values also described specific radiation changes (Fig. 1). This behaviour can be interpreted as specific radiation changes occurring at a proportional rate across the crystal. Owing to the very high quality of the thaumatin data, a small departure from this simplifying assumption was observed, presumably due to non-linear kinetics of damage to disulfide bridges.

How can the magnitude of the structural effects associated with a particular singular value be interpreted? The magnitude of the native signal defined as 100% and represented by the first singular value was used as a reference point. The subsequent singular values, which may represent structural effects, were multiplied by a factor of two. This factor results from disregarding the phase-dependent/amplitude-independent component in SVD (§2.2). The singular value represents the variance, and thus it is necessary to take the square root of it to obtain the root mean square (RMS). Additionally, it may be appropriate to describe the effect as representing the full range of variation, rather than the RMS within the datasets. For instance, in the case of radiation-induced specific effects, which change linearly with dose, the question can be asked as to how much the structure factors have changed from the start to the end of data collection rather than how much they changed on average relative to the midpoint of the data collection. The relationship of RMS deviations in an evenly sampled range of values from 0 to d is such that the RMS is equal to one-third of d. Singular values obtained in SVD analysis represent RMS deviations, *i.e.* they can be interpreted as the change of structure factors expressed with respect to the structure factors at average dose d with RMS describing their variability within one-third of the dose range. Therefore, if a full range of the structure factors change from dose zero to d is required, it can be derived from a singular value by multiplying it by a factor of three. However, for the Bijvoet signal, the crystallographic convention is to represent it with respect to the average, so in this specific case a multiplication factor is not applied.

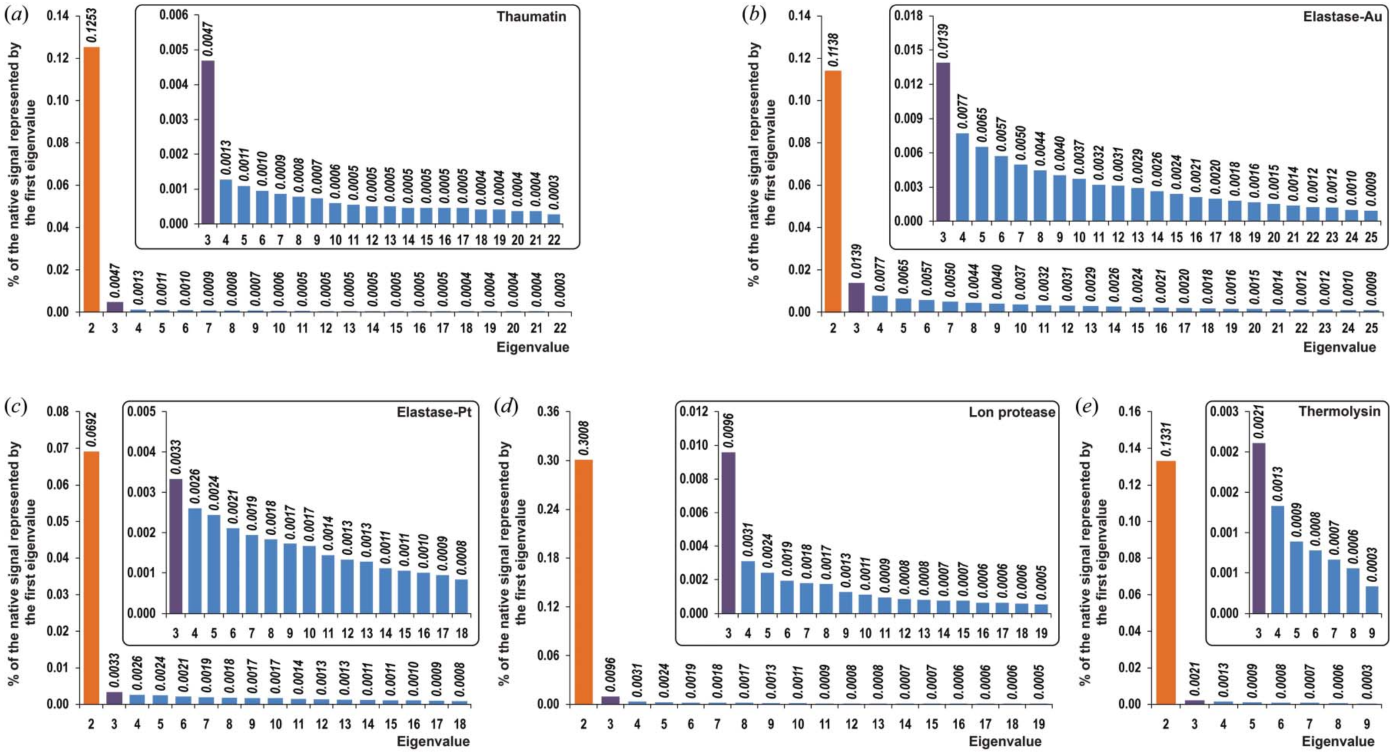

Figure 1
Eigenvalues of the PCA matrix for five series of datasets: (*a*) thaumatin, (*b*) Au-derivatized elastase, (*c*) Pt-derivatized elastase, (*d*) Lon protease, (*e*) thermolysin. Centric and non-centric reflections were used in the data analysis with the Friedel pairs merged. The native signal represents 100% and corresponds to the first eigenvalue, which was removed from the figures to show clearly the contributions from other components. The orange bars represent the second eigenvalue, the purple bars represent the third eigenvalue, and the blue bars represent the remaining eigenvalues. The insert shows the same plot after removing the second eigenvalue, so it can be seen how the third eigenvalue compares with the subsequent ones. The plots show that in the case of thaumatin and Lon protease there are two significant (above the noise level) components of radiation-induced specific changes, whereas for thermolysin and elastase the second component of radiation-induced specific changes has borderline statistical significance. The noise components were defined by analyzing ratios between consecutive eigenvalues, *i.e.* 2:3, 3:4 and so on. If these ratios are only slightly above 1, these eigenvalues are interpreted as noise components. The components preceding such a group have a combination of noise and signals. If the preceding component has a ratio to the next one that is less than 2.0, the signal in it is interpreted to be smaller than the noise; the signal can still be statistically significant, but is not very informative.

The right singular vector or, in the thaumatin case, a linear combination of two right singular vectors, was plotted against the scaling *B*-factor increase. The latter is a proxy of dose and the very good match between the right singular value(s) and increase in scaling *B*-factor supports the interpretation of singular vectors as corresponding to specific radiation changes (Fig. 2). The analysis of data variance indicates that the model is complete within the experimental error level (Fig. 1).

When Friedel mates were used separately in SVD, another significant singular value represented the anomalous signal in all datasets. A right singular vector shows that for this component there is a small variability of the anomalous signal between datasets. SVD uses only a linear combination of datasets, thus the changes in values of elements of the right singular vector represent a proportional increase or decrease in anomalous signal between datasets. This SVD component describes only proportional changes in occupancies of heavy atoms or proportional changes in wavelength-dependent scattering factors. The change of scale of anomalous signal between datasets does not generate an additional phasing source. An additional phasing source would require significant non-proportional changes in heavy-atom occupancies or changes in their positions. For instance, a crystal containing two different anomalous scatters, *e.g.* Zn and Se, will have very different anomalous scattering ratios for the pair of wavelengths below and above the absorption edge of one of them. An experiment performed at these two wavelengths may potentially generate two phasing components. In the case reported here, the presence of significant non-proportional changes in heavy-atom occupancies would create another singular value in the anomalous signal category, but this was not observed (Table 2).

When only one singular component is observed, modelling heavy-atom changes using unmerged data cannot generate better phasing than using properly weighted merged data. In the case of thaumatin, anomalous scatterers were affected by specific radiation changes; however, the additional component in the electron density changes was quite low, so its impact on the anomalous scattering was below the noise level, even though the total level of anomalous signal was sufficient to solve the structure. This directly answers the question of under what circumstances there may be an advantage of using unmerged data in experimental phasing calculations. In the cases presented here, using unmerged data does not provide any advantage because no additional components in the anomalous signal were observed, and the data were properly down-weighted for decay of the anomalous signal (which is typically not yet employed by standard procedures). However,

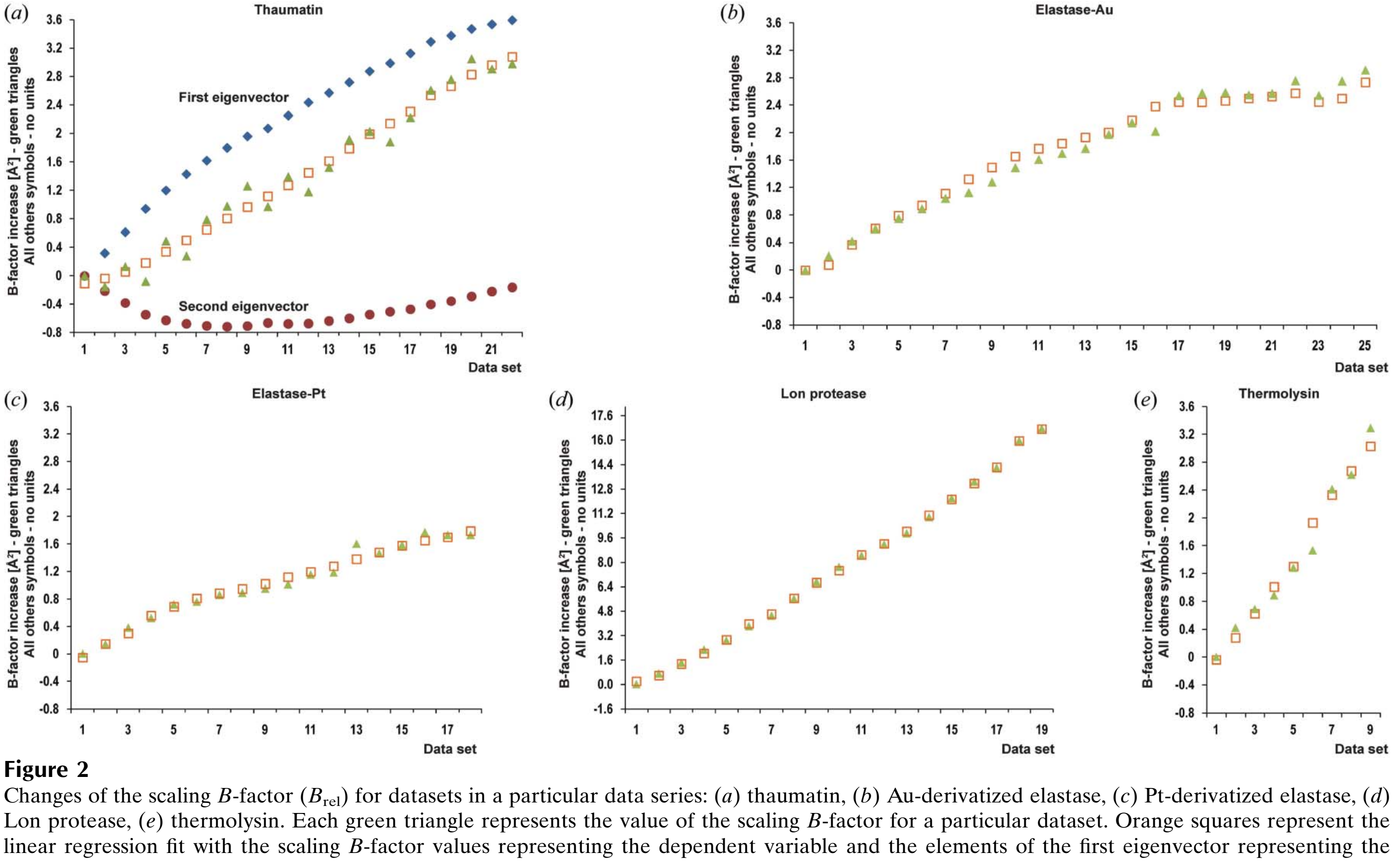

Figure 2

Changes of the scaling *B*-factor ($B_{\rm rel}$) for datasets in a particular data series: (*a*) thaumatin, (*b*) Au-derivatized elastase, (*c*) Pt-derivatized elastase, (*d*) Lon protease, (*e*) thermolysin. Each green triangle represents the value of the scaling *B*-factor for a particular dataset. Orange squares represent the linear regression fit with the scaling *B*-factor values representing the dependent variable and the elements of the first eigenvector representing the independent variable set. The linear regression was performed for thermolysin ($R = 0.986$) and both versions of elastase crystals ($R = 0.988$ for Pt-derivatized elastase and $R = 0.984$ for Au-derivatized elastase). In the case of thaumatin and Lon protease, elements of the first two eigenvectors were used for multiple linear regression with $R = 0.987$ for thaumatin and $R = 1.000$ for Lon protease. The excellent fit in all cases shows that the physical interpretation of the first and sometimes the first and second eigenvectors as representing effects arising from specific radiation changes agrees well with the increase of the scaling *B*-factor, a physical quantity, which is directly related to the radiation-induced decay. In (*a*) the brown diamonds and the blue circles represent the elements of the first (brown) and the second (blue) eigenvectors scaled by their singular values.

Table 2
The most significant components identified by SVD were based on the calculation performed on the acentric reflections only; the components are represented as a percentage of the native signal (100%).

Disregarding the phases during SVD calculations resulted in the components having values two times lower than those calculated from true electron densities. Therefore, each eigenvalue was multiplied by two before taking its square root. Only for Lon protease did we observe an additional eigenvector corresponding to the anomalous signal. For all datasets the second component of specific radiation changes was observed, but only for thaumatin and Lon protease did this component have values above the level of the components corresponding to noise.

SVD components' contribution	Thaumatin	Thermolysin	Lon protease	Elastase-Au	Elastase-Pt
Anomalous signal I (%)	0.7	1.4	4.0	6.8	1.9
Anomalous signal II (%)	N/A	N/A	1.1	N/A	N/A
Radiation changes I (%)	4.8	4.9	7.7	4.5	3.6
Radiation changes II (%)	0.9	0.6	1.3	0.9	0.7

in experiments where heavy-atom occupancies change rapidly, for instance in crystals of mercury derivatized proteins (Ramagopal *et al.*, 2005) or halogenated nucleic acid (Ennifar *et al.*, 2002), and where sampling of unique reflections is poor on the timescale of occupancies decay, there may be an advantage in modelling the variability of heavy-atom occupancies in unmerged data (Schiltz *et al.*, 2004).

In the map calculated from the left singular vector for the gold derivative of elastase, the peaks at the gold site are of the same magnitude as the peaks at the disulfide bridges. Gold has five times more electrons than sulfur, so this observation implies an approximately five times slower change in occupancy at gold atoms compared with the sulfur atoms in disulfides. This is consistent with an observed 11% decrease of the anomalous signal on top of the overall decay, as defined by the right singular vector calculated from SVD with Friedel mates separated. The occupancy changes at the gold sites are remarkably small considering that, owing to their much larger atomic cross section for X-rays, gold atoms absorb about 2000 times more photons than light atoms of the protein, and 50 times more photons than the sulfur atoms. A conclusion, therefore, is that the impact of radiation on the phasing signal generated by heavy atoms is defined by their redox chemistry and the radiolysis of coordinating side chains rather than by direct X-ray absorption. The limited impact of specific radiation changes on heavy atoms used in phasing is a typical occurrence in the authors' experience; however, such a rule has many exceptions. Radiation-induced phasing (RIP) from X-ray-induced damage (Ravelli *et al.*, 2003) will work only occasionally, since specific damage to heavier atoms will be accompanied by many other rearrangements of the covalent structure that can be difficult to model for the purpose of such phasing. The RIP method has better chances of working if the damage is induced by UV radiation (Nanao & Ravelli, 2006), because absorption of UV produces fewer types of excited states than ionization resulting from high-energy photoelectrons. UV radiation used in such experiments had insufficient energy to effectively generate water radiolysis and therefore its impact on the protein structure was highly localized.

Experimental phasing based on SeMet derivatives is one of the most popular approaches to structure solution. Selenium atoms have a large atomic cross section and can dominate the overall X-ray absorption at or above the Se absorption peak, and correspondingly increase the decay rate. Despite the extensive use of the SeMet-based phasing methods, there are few published results reporting selenium atoms being differentially affected by the dose (Borek *et al.*, 2010; Schiltz & Bricogne, 2007). For Lon protease (Dauter *et al.*, 2005; Botos *et al.*, 2004), 19 consecutive full datasets were analysed that were exposed to a much higher dose than in other experiments to date with these crystals as indicated by the scaling *B*-factor increasing to 17 Å^2. This decay is equivalent to a decrease of diffraction intensity by more than 60% at a resolution of 3.0 Å. For experimental phasing, this exposure is close to the upper advisable limit, as defined by a diminishing return from adding even higher decayed diffraction (Schiltz & Bricogne, 2007). The real and the anomalous signals were analysed to investigate how they were affected when the *B*-factor increased by 17 Å^2. SVD analysis identified, as in the previous cases, two major components: specific changes induced by radiation and anomalous scattering (Table 2). Analysis of the anomalous scattering component indicated that Bijvoet differences decreased by 25% after correcting for the overall decay of intensities. This decrease is consistent with changes in occupancies or increased disorder of anomalous scatterers' positions (Figs. 3 and 4) observed in the difference maps based on the radiation-induced specific changes component. However, even though changes at Se positions are significant, they are not a dominating contributor to radiation-induced specific changes. Changes in anomalous scattering are, as in the case of the gold atoms described above, proportional between datasets, so they do not constitute an additional phasing component. Therefore, there is no advantage in modelling them at the level of individual anomalous scatterers, and it is sufficient to use the information about specific radiation damage during merging to down-weight the contribution from the observations acquired at high-exposure levels. In this case, down-weighting had no impact on the structure solution, due to the overall very high data quality. In practice, such an improvement would not be expected to be crucial for structure solution; however, in some borderline cases, the impact could be noticeable. In this case, the presence of weak additional components was also noticed for both the anomalous signal and the specific radiation-induced changes (Table 2). However, in both cases these second components, even though statistically significant, were below the threshold of useful signals for phasing or for correcting data.

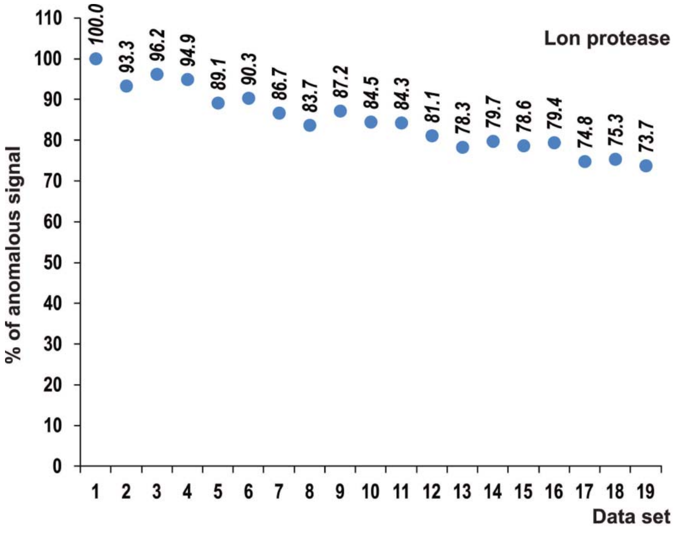

Figure 3
Plot of the elements of the eigenvector corresponding to the first component of the anomalous signal for Lon protease. The plot shows how the anomalous signal decreases during data collection, even after the correction for overall decay. It shows that Se positions decay faster than the rest of the structure, and, at the end of data collection, when the scaling B-factor increases by 16.8 Å^2, the anomalous signal is weaker by about 26%.

3.3. *A*-factor *versus B*-factor

One of the assumptions made in modelling the overall decay is that the atomic displacements induced by radiation obey a Gaussian distribution. This assumption was tested by implementing an alternative decay model proposed for X-ray crystallography by Holton (Holton & Frankel, 2010), and comparing it with the Gaussian model. Based on previous electron microscopy work, Holton proposed the use of a Lorentz distribution of atomic displacements, which after being Fourier transformed generates a somewhat different resolution dependence of decay,

$$I_h(d) = I_h(0)\exp\big(-A_d|\mathbf{S}_h|\big), \qquad (5)$$

where A_d is, as is B_d, proportional to dose d. However, A_d is expressed in Å whereas B_d is expressed in Å^2. Holton also produced a synthetic dataset with the A-factor used to simulate decay; in the publication, the letter H was used for the inverse of A.

In our analysis, incorrectly modelled resolution dependence of the decay produces a residual, which contributes to the second component of the model of radiation damage, the one which describes specific changes induced by radiation. To test which of the decay corrections is more appropriate, the decay correction was estimated using both models, and then a calculation was performed to identify which of them produced a lower magnitude of a component describing specific changes induced by radiation as defined by its norm or a singular value.

As expected, for the simulated dataset, the A-factor was the preferred result while, for the real experimental data, the B-factor model produced a very small advantage, about 0.01% in the component magnitude. Unlike in the work of Holton & Frankel, the resolution dependence of both models, B and A/H, was tested in scaling. Holton & Frankel compared the decay rate of different crystals, diffracting to different resolutions, and, based on that, drew conclusions on the resolution dependence of the decay. However, there are many possibilities as to why the dose-to-decay rate may vary or be incorrectly estimated between experiments. For this reason, we interpret Fig. 3 of Holton & Frankel as testing the dose-to-decay rate calibration rather than the decay resolution dependence.

In our analyses, no indication of a need to readjust the traditional modelling of decay was found; however, it is easy to check which model is better for a particular dataset. Regardless of this, when using a correction for specific changes, the scaling and merging results produced by these two models are very similar, as inaccuracy of the scaling model is incorporated

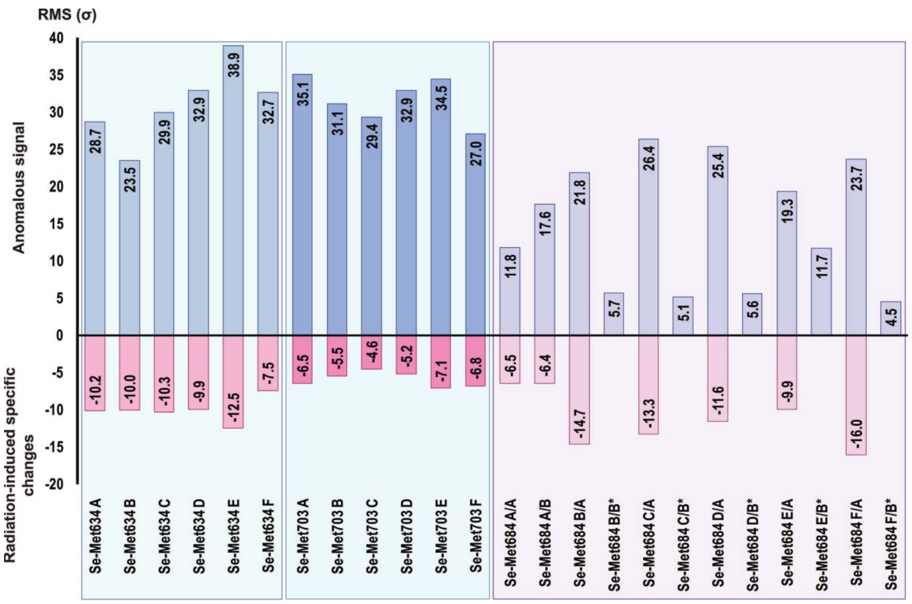

Figure 4
Plot of anomalous signal levels at different selenium positions (blue bars) and the levels of specific radiation changes for the same positions (red bars) for Lon protease. The signals are presented in RMS (σ) units of corresponding difference maps. Lon protease has six molecules in the asymmetric unit, so each selenium position in the protomer is represented six times. Additionally, one of the positions has a consistently double conformation in all monomers. The plot clearly shows a different level of order for specific Se positions, which results in a higher anomalous signal for better ordered side chains. The specific radiation changes are also different for different positions, and are not exactly proportional to the anomalous signal levels. The pattern of anomalous signal and specific radiation changes is mostly preserved in equivalent positions of different monomers. However, differences in equivalent positions are larger than expected from random variations (1 RMS level, 1σ), and therefore crystal packing also has some impact on specific changes at Se positions.

into the specific change component without affecting the downstream calculations.

3.4. The chemistry of specific radiation damage

A well known type of specific damage is decarboxylation of the side chains of glutamic and aspartic acids (Banumathi *et al.*, 2004; Burmeister, 2000; Ravelli & McSweeney, 2000; Weik *et al.*, 2000). These reactions happen at different rates for particular amino acids, possibly indicating that the electric field within the crystal is a major modulator of the migration of positive holes that can then cause the creation of hydronium ions. A decarboxylation reaction can change the local electrostatic interactions in the protein, potentially inducing a small but significant, in terms of its impact on the structure factors, motion of domains. Such effects strongly depend on the strength of the electric field on the surface of a particular protein. Since the presence of strong electric fields is a feature of many catalytic sites, specific radiation damage may be enhanced in these regions. On the other hand, it is difficult to predict *a priori*, particularly for a new protein structure, what the ratio of the magnitude of the specific change component to the dose will be. This observation is confirmed by the large spread of ratios of the magnitudes of specific changes to dose in various experiments (Borek *et al.*, 2007, 2010).

A large component of chemical bond rearrangements is radiolysis of water and proteins resulting in presumable accumulation of molecular hydrogen (Meents *et al.*, 2009; Reimann *et al.*, 1984). Since H_2 does not contribute much to X-ray scattering, its main impact is through changing neighbouring atom positions. This is an explanation for the chemical origin of the decay of diffraction intensities in cryo-cooled crystals. To the extent that the radiolysis rate per unit of the dose is constant in different crystals, the decay rate, as estimated by Kmetko *et al.* (2006), should be the same for different experiments.

4. Summary

There are several reasons why it is important to include corrections for specific radiation-induced changes in data processing. First, lack of these corrections affects the estimation of differences between symmetry-equivalent reflections. These differences are an important part of data processing and phasing analysis. They are used: (i) to determine crystallographic symmetry, (ii) to assess the level and quality of the phasing signal, and (iii) to accurately estimate the level of signal and noise in diffraction experiments. Discrepancies in the intensities of symmetry-equivalent reflections exceeding the random error level are an important hallmark of undetected systematic effects, for instance instrumental problems. In the oscillation/rotation methods, symmetry-equivalent observations are frequently acquired at different points in time, and also at different levels of X-ray dose, with a potential for diffraction measurements originating from different structural states. Therefore, every crystallographic procedure that relies on merging symmetry-equivalent observations may be negatively affected by radiation-induced changes.

The National Institutes of Health supported this work with grant GM053163. We are indebted to Raquel Bromberg and Halszka Czarnocka for help in editing the manuscript.

References

Alter, O., Brown, P. O. & Botstein, D. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.

Arnott, S. & Wonacott, A. J. (1966). *Polymer*, **7**, 157–166.

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

Borek, D., Cymborowski, M., Machius, M., Minor, W. & Otwinowski, Z. (2010). *Acta Cryst.* D**66**, 426–436.

Borek, D., Ginell, S. L., Cymborowski, M., Minor, W. & Otwinowski, Z. (2007). *J. Synchrotron Rad.* **14**, 24–33.

Botos, I., Melnikov, E. E., Cherry, S., Tropea, J. E., Khalatova, A. G., Rasulova, F., Dauter, Z., Maurizi, M. R., Rotanova, T. V., Wlodawer, A. & Gustchina, A. (2004). *J. Biol. Chem.* **279**, 8140–8148.

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

Dauter, Z., Botos, I., LaRonde-LeBlanc, N. & Wlodawer, A. (2005). *Acta Cryst.* D**61**, 967–975.

Dean, J. A. (1992). *Lange's Handbook of Chemistry*, 14th ed. New York: McGraw-Hill.

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

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

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

Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.

Evans, P. R. (2011). *Acta Cryst.* D**67**, 282–292.

Fox, G. C. & Holmes, K. C. (1966). *Acta Cryst.* **20**, 886–891.

Fütterer, K., Ravelli, R. B. G., White, S. A., Nicoll, A. J. & Allemann, R. K. (2008). *Acta Cryst.* D**64**, 264–272.

Garman, E. F. & Nave, C. (2009). *J. Synchrotron Rad.* **16**, 129–132.

Gray, H. B. & Winkler, J. R. (1996). *Annu. Rev. Biochem.* **65**, 537–561.

Gray, H. B. & Winkler, J. R. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 3534–3539.

Gray, H. B. & Winkler, J. R. (2009). *Chem. Phys. Lett.* **483**, 1–9.

Gray, H. B. & Winkler, J. R. (2010). *Biochim. Biophys. Acta*, **1797**, 1563–1572.

Henderson, R. (1995). *Q. Rev. Biophys.* **28**, 171–193.

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

Holton, J. M. (2007). *J. Synchrotron Rad.* **14**, 51–72.

Holton, J. M. & Frankel, K. A. (2010). *Acta Cryst.* D**66**, 393–408.

Itikawa, Y. & Mason, N. (2005). *J. Phys. Chem. Ref. Data*, **34**, 1–22.

Kmetko, J., Husseini, N. S., Naides, M., Kalinin, Y. & Thorne, R. E. (2006). *Acta Cryst.* D**62**, 1030–1038.

Krojer, T. & von Delft, F. (2011). *J. Synchrotron Rad.* **18**, 387–397.

Meents, A., Dittrich, B. & Gutmann, S. (2009). *J. Synchrotron Rad.* **16**, 183–190.

Murray, J. W., Rudiño-Piñera, E., Owen, R. L., Grininger, M., Ravelli, R. B. G. & Garman, E. F. (2005). *J. Synchrotron Rad.* **12**, 268–275.

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

O'Neill, P., Stevens, D. L. & Garman, E. (2002). *J. Synchrotron Rad.* **9**, 329–332.

Otwinowski, Z., Borek, D., Majewski, W. & Minor, W. (2003). *Acta Cryst.* A**59**, 228–234.

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

Otwinowski, Z. & Minor, W. (2000). *International Tables for Crystallography*, Vol. F, edited by M. G. Rossmann, pp. 226–235. Dordrecht: Kluwer Academic Publishers.

Owen, R. L., Rudiño-Piñera, E. & Garman, E. F. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 4912–4917.

Paithankar, K. S. & Garman, E. F. (2010). *Acta Cryst.* D**66**, 381–388.

Paithankar, K. S., Owen, R. L. & Garman, E. F. (2009). *J. Synchrotron Rad.* **16**, 152–162.

Rajagopal, S., Kostov, K. S. & Moffat, K. (2004*a*). *J. Struct. Biol.* **147**, 211–222.

Rajagopal, S., Schmidt, M., Anderson, S., Ihee, H. & Moffat, K. (2004*b*). *Acta Cryst.* D**60**, 860–871.

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

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

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

Reimann, C. T., Boring, J. W., Johnson, R. E., Garrett, J. W., Farmer, K. R., Brown, W. L., Marcantonio, K. J. & Augustyniak, W. M. (1984). *Surf. Sci.* **147**, 227–240.

Romo, T. D., Clarage, J. B., Sorensen, D. C. & Phillips, G. N. (1995). *Proteins Struct. Funct. Genet.* **22**, 311–321.

Sanishvili, R., Yoder, D. W., Pothineni, S. B., Rosenbaum, G., Xu, S., Vogt, S., Stepanov, S., Makarov, O. A., Corcoran, S., Benn, R., Nagarajan, V., Smith, J. L. & Fischetti, R. F. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 6127–6132.

Schiltz, M. & Bricogne, G. (2007). *J. Synchrotron Rad.* **14**, 34–42.

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

Schmidt, M., Rajagopal, S., Ren, Z. & Moffat, K. (2003). *Biophys. J.* **84**, 2112–2129.

Stewart, G. W. (1993). *Siam Rev.* **35**, 551–566.

Terryn, H., Deridder, V., Sicard-Roselli, C., Tilquin, B. & Houée-Levin, C. (2005). *J. Synchrotron Rad.* **12**, 292–298.

Tezcan, F. A., Crane, B. R., Winkler, J. R. & Gray, H. B. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 5002–5006.

Timneanu, N., Caleman, C., Hajdu, J. & van der Spoel, D. (2004). *Chem. Phys.* **299**, 277–283.

Wall, M. E., Rechtsteiner, A. & Rocha, L. M. (2003). *A Practical Approach to Microarray Data Analysis.* Norwell: Kluwer.

Warkentin, M., Badeau, R., Hopkins, J. B., Mulichak, A. M., Keefe, L. J. & Thorne, R. E. (2012*a*). *Acta Cryst.* D**68**, 124–133.

Warkentin, M., Badeau, R., Hopkins, J. B. & Thorne, R. E. (2012*b*). *Acta Cryst.* D**68**, 1108–1117.

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.