Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Improving radiation-damage substructures for RIP

Max H. Nanao,[a] George M. Sheldrick[b] and Raimond B. G. Ravelli[a]*

[a]EMBL, 6 Rue Jules Horowitz, 38042 Grenoble, France, and [b]Georg-August Universität, Tammannstrasse 4, D-37077 Göttingen, Germany

Correspondence e-mail: ravelli@embl-grenoble.fr

Received 28 April 2005
Accepted 17 June 2005

PDB References: elastase, 2blo, r2blosf; 2blq, r2blosf; insulin, 2bn3, r2bn3sf; 2bn1, r2bn1sf; lysozyme, 2blx, r2blxsf; 2bly, r2blysf; ribonuclease A, 2blp, r2blpsf; 2blz, r2blzsf; thaumatin, 2blr, r2blrsf; 2blu, r2blusf; trypsin, 2blv, r2blvsf; 2blw, r2blwsf.

Specific radiation damage can be used to solve macromolecular structures using the radiation-damage-induced phasing (RIP) method. The method has been investigated for six disulfide-containing test structures (elastase, insulin, lysozyme, ribonuclease A, trypsin and thaumatin) using data sets that were collected on a third-generation synchrotron undulator beamline with a highly attenuated beam. Each crystal was exposed to the unattenuated X-ray beam between the collection of a 'before' and an 'after' data set. The X-ray 'burn'-induced intensity differences ranged from 5 to 15%, depending on the protein investigated. X-ray-susceptible substructures were determined using the integrated direct and Patterson methods in *SHELXD*. The best substructures were found by downscaling the 'after' data set in *SHELXC* by a scale factor K, with optimal values ranging from 0.96 to 0.99. The initial substructures were improved through iteration with *SHELXE* by the addition of negatively occupied sites as well as a large number of relatively weak sites. The final substructures ranged from 40 to more than 300 sites, with strongest peaks as high as 57σ. All structures except one could be solved: it was not possible to find the initial substructure for ribonuclease A, however, *SHELXE* iteration starting with the known five most susceptible sites gave excellent maps. Downscaling proved to be necessary for the solution of elastase, lysozyme and thaumatin and reduced the number of *SHELXE* iterations in the other cases. The combination of downscaling and substructure iteration provides important benefits for the phasing of macromolecular structures using radiation damage.

1. Introduction

With the advent of highly intense sources of X-rays from modern insertion-device beamlines on third-generation synchrotrons, understanding radiation damage to cryocooled crystals has assumed a prominent role in X-ray crystallography. Radiation damage can be a serious problem in the phasing of macromolecular crystal structures, as the heavy-atom substructure seems to be particularly prone to radiation damage, thereby reducing the anomalous signal (Rice *et al.*, 2000; Ravelli *et al.*, 2005). In addition, non-isomorphism introduced by changes in unit-cell parameters coupled with molecular rotations and translations (Ravelli & McSweeney, 2000) and specific changes such as breakage of disulfide bonds and loss of definition of carboxyl groups (Burmeister, 2000; Weik *et al.*, 2000) hamper the reliable measurements of small dispersive signals between the inflection point and remote wavelength during a multiple-wavelength anomalous dispersion (MAD) experiment.

It has recently been demonstrated that the specific effects of radiation damage can be used to provide phase information and indeed to phase structures *de novo* (Ravelli *et al.*, 2003). This method, named radiation-damage-induced phasing (RIP), has been applied to several cases, often in combination with anomalous scattering (RIPAS; Evans *et al.*, 2003; Banumathi *et al.*, 2004; Schiltz *et al.*, 2004; Weiss *et al.*, 2004; Zwart *et al.*, 2004; Ravelli *et al.*, 2005). The RIP method can loosely be thought of as similar to SIR (single isomorphous replacement), with concomitant issues of twofold ambiguity. Two or more data sets $R \ldots T$ are collected on the same crystal and the first data set R is compared with a later data set T. In the presence of radiation damage, significant intensity differences can be observed between these two data sets. Data set T is used as a 'pseudo-native' data set and data set R (the more 'electron-rich') as the derivative data set. Despite this general similarity to the SIR methodology, there are important differences between SIR and RIP. In contrast to RIP, successful heavy-atom (HA) derivatization is needed for SIR and data have to be collected from two different isomorphous crystals. HA derivatization is often time-consuming and even when successful can lead to changes in unit-cell parameters, causing non-isomorphism between the native crystal and the HA derivative. This disadvantage is offset by the large differences in the structure factors that can be obtained upon the addition of just a few electron-rich heavy atoms. In contrast, RIP data can be collected from a single native crystal. The origins of the differences in intensity upon radiation damage are twofold. Firstly, there are changes that occur at specific highly X-ray-susceptible chemical groups. Secondly, there is a background of non-specific changes, such as expansion of the unit cell, rotation and translation of molecules. The non-specific changes add to a general non-isomorphism between early and later data sets. Since the data are collected on one crystal, the non-isomorphism during a RIP experiment will in general be smaller than that of a typical SIR experiment. However, if the specific changes occur at light atoms, the resulting effects on the structure factors will be minor, which makes RIP inherently more difficult than SIR. The boundary between specific and non-specific is not clear-cut; rather, there is a continuum of susceptible sites. Disulfide breakage has become a well known marker of radiation damage (Burmeister, 2000; Ravelli & McSweeney, 2000; Weik *et al.*, 2000) since these sites are highly susceptible to radiation damage. Upon breakage of a disulfide bond, the direct environment of this bond will also be affected. This results in a large number of small changes in the immediate area and the resolution and data quality will determine whether these changes can be classified as specific changes or rather as factors adding to the global non-isomorphism.

This paper deals with the accurate determination of the position of radiation-sensitive sites and the use of these sites for phasing. We show that RIP can be seen as MIR (multiple-isomorphous replacement) that lacks a native data set and show that the Patterson maps can contain negative cross-vector peaks. Radiation damage will always cause an overall decrease in scattering power, an effect that is ignored by most scaling programs. We argue that this results in an over-scaling of the later data set T and propose a simple way to overcome this. Data have been collected on crystals of six different disulfide-containing proteins. These well diffracting model proteins could all be phased using RIP, although each of them had specific difficulties to overcome, either in substructure determination or in the subsequent phasing. We have used these model proteins to show that substructure iteration and scaling alone or in combination provide important benefits for the RIP method.

2. Theory

Assume that the total structure factor consists of two parts, one dose-independent, $F_{\rm const}$, and one that varies with exposure of the crystal to X-rays, $F_{\rm var}$. $F_{\rm var}^{\rm before}$ can be described as

$$F_{\rm var}^{\rm before} = \sum_{j=1}^{N_{\rm sus}} q_j(0) f_j \exp(2\pi i \mathbf{h} r_j), \qquad (1)$$

where $N_{\rm sus}$ is the number of susceptible sites located at position r_j and $q_j(0)$ is the occupancy of these sites at zero dose. Upon exposure of the crystal, the occupancy of the X-ray-susceptible atoms will change,

$$F_{\rm var1} = \sum_{j=1}^{N_{\rm sus}} q_j({\rm dose}) f_j \exp(2\pi i \mathbf{h} r_j). \qquad (2)$$

Banumathi *et al.* (2004) derived a model where all X-ray-susceptible sites lose definition upon X-ray exposure at an equal rate,

$$F_{\rm var1} = q({\rm dose}) \sum_{j=1}^{N_{\rm sus}} f_j \exp(2\pi i \mathbf{h} r_j). \qquad (3)$$

In this model, $F_{\rm var1}$ for a later data set has the same phase angle as $F_{\rm var}^{\rm before}$, but a reduced amplitude. The diffraction intensities will change as a quadratic function of dose if there are no unit-cell volume changes and q(dose) is a linear function of dose. This model might be of particular use when there is only one or a few identical highly susceptible sites in the structure, such as in the case of brominated DNA or RNA (Schiltz *et al.*, 2004; Ravelli *et al.*, 2003). However, previous studies have shown that an equal decay-rate model does not hold for disulfide breakage: different disulfide bonds have different susceptibilities (Weik *et al.*, 2000; Ravelli & McSweeney, 2000). This can cause small phase shifts in $F_{\rm var1}$ (Fig. 1). Additionally, for some bonds clear new positions can be found for the moving atoms (Ravelli & McSweeney, 2000). In the case of disulfide bridges, within each bond one S atom often forms an 'anchor', whereas the other moves more appreciably. This has also been observed upon breakage of a Br—G bond in RNA (Ravelli *et al.*, 2003). These new radiation-damage-induced positions will produce a structure factor $F_{\rm var2}$ with a phase angle unrelated to $F_{\rm var}^{\rm before}$,

$$F_{\rm var2} = \sum_{j=1}^{N_{\rm new}} q'_j({\rm dose}) f_j \exp(2\pi i \mathbf{h} r'_j), \qquad (4)$$

where N_{new} is the number of X-ray-induced sites located at position r'_j, with occupancies q'_j(dose) that start at zero at zero dose. In total, a new structure factor is created,

$$F_{var}^{after} = F_{var1} + F_{var2}. \quad (5)$$

The resulting Argand diagram (Fig. 1) is similar to that corresponding to an MIR experiment (Kendrew *et al.*, 1958), where F_{const} can be thought of as the structure factor belonging to the native structure, whereas F_{before} and F_{after} can be seen as equivalent to two different heavy-atom derivatives. Rossmann (1960) has shown that a difference Patterson synthesis with $(|F_1| - |F_2|)^2$ as coefficients is approximately equivalent to the self-Patterson of the heavy atoms in compound 1 plus the self-Patterson of the heavy atoms in compound 2, minus the cross-Patterson between the heavy atoms in compounds 1 and 2. Translated to RIP, a difference Patterson synthesis with $(|F_{before}| - |F_{after}|)^2$ will be approximately equivalent to the self-Patterson of the N_{sus} X-ray-susceptible atoms plus the self-Patterson of the N_{new} X-ray-induced sites, minus the cross-Patterson between the X-ray-susceptible and X-ray-induced sites.

Fig. 2 shows the $v = \frac{1}{2}$ Harker section of a calculated data series for turkey egg-white lysozyme (TEL) in space group $P2_1$, based on PDB code 1ljn. TEL has four disulfide bonds, three of which were subsequently altered. Fig. 2(*a*) shows the effect of an equal reduction of the occupancies of the Cys6–Cys127 S^{γ} atoms. In this case, F_{after} becomes equal to F_{const} and the difference Patterson synthesis will equal the self-Patterson of the N_{sus} X-ray-susceptible atoms. In reality, the S atoms will not just evaporate, but might appear in new positions. Fig. 2(*b*) shows the simulated $v = \frac{1}{2}$ Harker section when the Cys76–Cys94 bonds breaks and Cys94 S^{γ} moves, remaining fully occupied, to a new rotamer position. The Harker section now contains both positive (red) and negative (blue) peaks. Two positive peaks are found, one corresponding to the self-Patterson of the X-ray-susceptible site (marked as 'original') and one corresponding to the self-Patterson of the X-ray-induced site (marked as 'rotamer'). In between these peaks, a negative peak can be found corresponding to the cross-Patterson.

Fig. 2(*c*) shows the effect of disulfide-bond elongation (Weik *et al.*, 2002). Both S^{γ} atoms within the Cys30–Cys115 bond were modelled to move 0.4 Å, causing an overall disulfide-bond elongation from 2.0 to 2.8 Å. Given the limited resolution of the calculated data series (1.6 Å), the self- and cross-Pattersons overlap, thus reducing the overall absolute peak heights. The individual self-Patterson of the X-ray-susceptible sites, such as that shown in Fig. 2(*b*), cannot be identified owing to this overlap. Treating this situation as SIR will inevitably introduce errors.

In practice, X-ray-susceptible atoms are not likely to move to a single new position. Although not the case in Fig. 2(*b*), the positive peak corresponding to the new sulfur position can be expected to have a lower peak height than that of the original position, since multiple new rotamer positions might be available upon breakage of the disulfide bond. Breakage of a C^{β}—S^{γ} bond will result in even more alternative sulfur positions. The probability of an individual reaction path within all

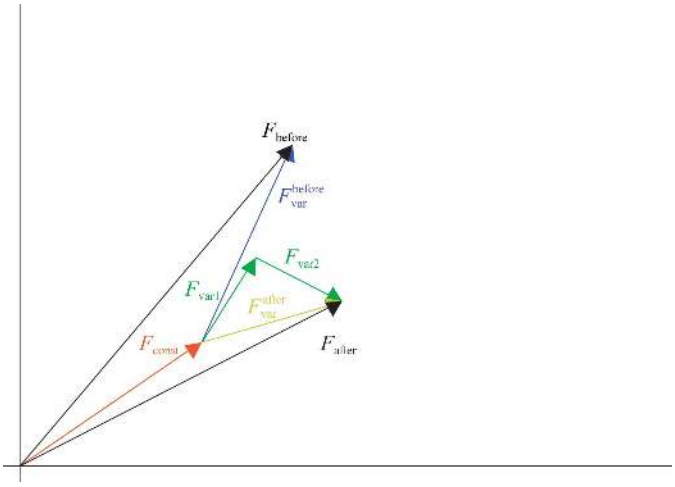

Figure 1
Argand diagram for RIP. A comparison with MIR is made, where the constant radiation-insensitive part of the structure factor (F_{const}) would correspond to a native structure. The 'before' data set could be seen as a first derivative and is represented as F_{before}, with $F_{before} = F_{const} + F_{var}^{before}$. The 'after' data set could be seen as a second derivative, as new radiation-damage-induced sites appear, $F_{after} = F_{const} + F_{var}^{after}$.

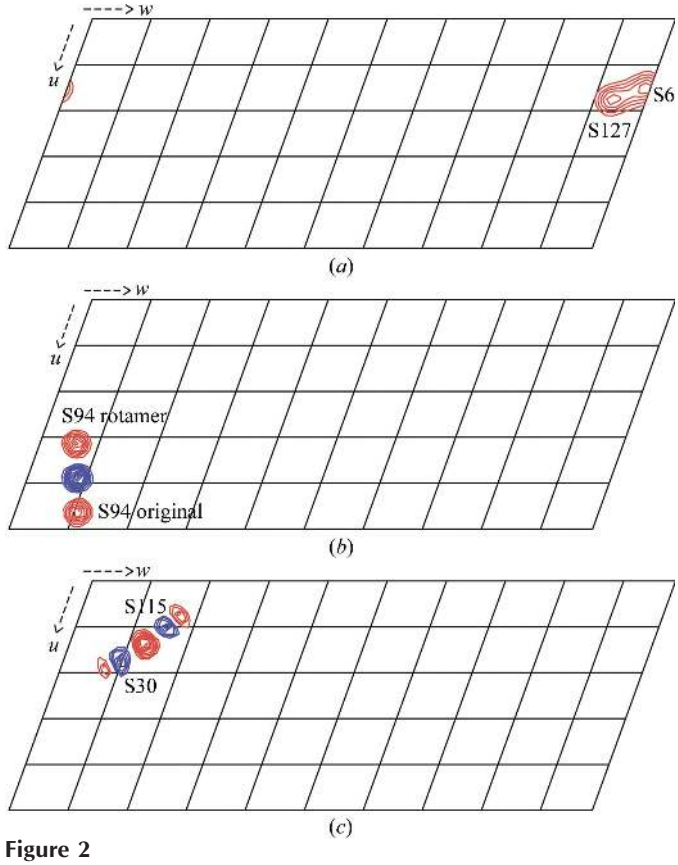

Figure 2
Simulated $v = 0.5$ Harker sections of the radiation-damage difference Patterson for $P2_1$ lysozyme. The effect of (*a*) loss of disulfide occupancy for Cys6–Cys127, (*b*) disulfide-bond breakage for Cys76–Cys94, where Cys94 S^{γ} moves to a new rotamer position, and (*c*) disulfide-bond elongation of Cys30–Cys115 is shown. Positive peaks are contoured in red and negative peaks in blue.

Table 1
Crystallization and cryoprotection conditions.

	Elastase	Insulin	Lysozyme	Ribonuclease A	Thaumatin	Trypsin
Crystallization conditions	0.25 *M* ammonium sulfate, 100 m*M* sodium acetate pH 5.0	0.4 *M* $NaPO_4$/Na_2HPO_4 pH 10.4, 1 m*M* EDTA	30% PEG 5000 MME, 1 *M* NaCl, 50 m*M* sodium acetate pH 4.5	1.75 *M* ammonium sulfate, 2 *M* NaCl, 100 m*M* sodium acetate pH 5.5	0.9 *M* sodium/potassium tartrate, 100 m*M* HEPES pH 7.3, 15% glycerol	25% PEG 8000, 0.2 *M* ammonium sulfate, 100 m*M* Tris–HCl pH 8.0
Cryoprotectant	30% glycerol	30% glycerol	30% PEG 400	—	30% glycerol	15% glycerol

Table 2
Data-collection statistics.

Values in parentheses are for the highest resolution shell.

	Elastase		Insulin		Lysozyme		Ribonuclease A		Thaumatin		Trypsin	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Space group	$P2_12_12_1$		$I2_13$		$P4_32_12$		$P3_221$		$P4_12_12$		$P2_12_12_1$	
Unit-cell parameters (Å)	a = 49.77, b = 57.62, c = 74.08	a = 49.79, b = 57.63, c = 74.09	a = 78.05	a = 78.08	a = 77.33, c = 38.16	a = 77.35, c = 38.19	a = 64.13, c = 63.64	a = 64.14, c = 63.74	a = 57.78, c = 150.08	a = 57.90, c = 150.36	a = 54.16, b = 58.25, c = 66.58	a = 54.18, b = 58.28, c = 66.60
Resolution (Å)	45.00–1.00 (1.06–1.00)	45.00–1.00 (1.06–1.00)	45.00–1.50 (1.59–1.50)	45.00–1.50 (1.59–1.50)	45.00–1.28 (1.36–1.28)	45.00–1.40 (1.49–1.40)	45.00–1.40 (1.49–1.40)	45.00–1.50 (1.59–1.50)	45.00–1.40 (1.49–1.40)	45.00–1.40 (1.49–1.40)	45.00–1.20 (1.30–1.20)	45.00–1.20 (1.30–1.20)
No. of reflections	490029 (75139)	490548 (75230)	92271 (14523)	92357 (14529)	223254 (24273)	181716 (28535)	106431 (16848)	87150 (13700)	235006 (38790)	236455 (39902)	261071 (54136)	261408 (54375)
No. of unique reflections	208762 (32156)	208883 (32208)	24081 (3757)	24081 (3756)	30260 (4762)	23305 (3684)	29829 (4695)	24403 (3835)	93386 (15483)	93941 (15870)	122171 (25449)	122363 (25566)
Completeness† (%)	93.9 (90.1)	93.9 (90.3)	97.9 (95.9)	97.9 (95.9)	99.7 (98.8)	99.8 (99.4)	99.2 (908.0)	99.3 (98.0)	97.6 (94.8)	98.2 (97.2)	96.0 (93.5)	96.1 (94.0)
$\langle I/\sigma(I)\rangle$	13.09 (5.4)	12.61 (3.7)	19.30 (5.0)	18.99 (4.1)	24.14 (4.9)	25.16 (6.4)	15.43 (3.9)	16.15 (3.7)	15.80 (3.8)	15.38 (3.4)	12.46 (6.5)	12.15 (5.8)
R factor‡ (%)	4.6 (20.7)	5.0 (24.5)	3.8 (26.9)	4.0 (33.2)	4.9 (29.8)	5.2 (29.2)	4.7 (27.8)	4.6 (30.0)	4.0 (26.4)	4.3 (30.3)	4.7 (12.2)	4.8 (14.1)
Redundancy	2.3	2.3	3.8	3.8	4.2	4.2	1.9	1.9	2.5	2.5	2.1	2.1
Dose/data set (Gy)	8×10^4	8×10^4	2×10^6	2×10^6	6×10^5	6×10^5	3×10^6	3×10^6	8×10^5	8×10^5	4×10^5	4×10^5
Crystal size (mm)	0.2 × 0.2 × 0.5		0.1 × 0.1 × 0.1		0.2 × 0.2 × 0.2		0.15 × 0.15 × 0.15		0.13 × 0.13 × 0.20		0.1 × 0.1 × 0.25	
Beam size (mm)	0.2 × 0.2		0.1 × 0.1		0.08 × 0.2		0.05 × 0.15		0.1 × 0.13		0.1 × 0.1	
Dose, X-ray burn (Gy)	2×10^6		2×10^6		2×10^6		4×10^6		2×10^6		2×10^6	
R_{int} compared with before (with local scaling)§ (%)	6.8 (6.6)		7.6 (7.2)		4.9 (4.9)		14.6 (12.9)		8.8 (8.6)		6.5 (6.5)	
$\langle \Delta F\rangle/\langle\sigma(\Delta F)\rangle$	1.32		1.90		1.69		2.13		1.87		1.21	

† Completeness is sometimes compromised solely for the higher resolution shells, as measurements in the corners of the square detector were included. ‡ R factor is $\sum |I(h,i) - I(h)| / \sum I(h)$, where $I(h)$ is the intensity of the reflection with index h and $I(h,i)$ is the intensity of the symmetry-related i reflection with index h. § From *XPREP*. R_{int} is $\sum |I - \langle I\rangle| / \sum I$.

possible paths, together with the resolution of the data and the distance between the old and new positions, determine which new positions can be observed. Overall, it is unlikely that every X-ray-susceptible atom will move to just one or two new well defined positions. This results in a net loss of the number of ordered atoms for later data sets, causing F_{after}, on average, to be smaller than F_{before}.

Most scaling programs, whether used for SAD, MAD, SIR or MIR, try to put different data sets on the same scale through Wilson or local scaling (Matthews & Czerwinski, 1975; Howell & Smith, 1992; Blessing, 1997). For weak derivatives, an accurate knowledge of the scale factor might not be required (Rossmann, 1960); however, it has been shown that adjustment of the scale factor to compensate for the additional scattering power of heavy atoms for strong derivatives can lead to significant improvements (Nikonov, 1983). Although individual structure-factor amplitudes might grow stronger with X-ray dose, the total scattering power will normally decrease during an RIP experiment. Ignoring this during scaling will result in the amplitudes of the later data set T being too large on average. In this paper, we will downscale the after data sets by a factor K in order to overcome this problem.

3. Materials and methods

3.1. Crystallization

All proteins were obtained as lyophilized powders from Sigma (except where noted otherwise) and could be used without further purification. Bovine insulin was dissolved in 20 m*M* phosphate buffer pH 10.4 to a concentration of 20 mg ml^{-1} and porcine elastase (Roche) was dissolved in water to a concentration of 20 mg ml^{-1}. A higher concentra-

Table 3
SHELXC, *SHELXD* and *SHELXE* phasing statistics.

	Elastase	Insulin	Lysozyme	Ribonuclease A	Thaumatin	Trypsin
FIND/DSUL	4/4	3/3	8/4	—	8/8	4/4
Range of scale factor K†	0.93343–0.99286	0.95200–1.00214	0.99100	—	0.95571–0.99471	0.98357–0.99657
Resolution for *SHELXD* (Å)	1.5	2.5	2.5	—	1.8	2.2
Expanded to (Å)	1.33	1.5	1.4	1.4	1.4	1.2
K for best CC	0.96500	0.96686	0.98060	—	0.97243	0.98880
CC_{best}‡ (%)	14.5	16.9	22.8	—	16.6	14.2
CC_{weak}‡ (%)	6.7	7.4	11.35	—	6.1	7.9
PATFOM‡	8.9	24.2	4.7	—	3.2	15.4
Solvent content (%)	39	64	35	50	50	41
K for best pseudo-free CC	0.99286	0.99657	0.99100	—	0.99471	0.99286
Contrast§	0.470	1.050	0.405	0.731	0.729	0.506
Connectivity§	0.868	0.943	0.904	0.959	0.926	0.913
Pseudo-free CC§ (%)	86.8	85.1	74.2	82.5	83.1	81.7
wMPE initial§ (°)	73.3	78.8	85.4	—	81.8	78.7
Final wMPE§ (°)	23.2	22.1	35.0	25.1	25.7	26.1
No. of *SHELXE* iterations	1	1	11	1	1	1

† Yielding a pseudo-free CC of >70%. ‡ Solution from *SHELXD* with highest CC_{best}. § Solution from *SHELXE* with highest pseudo-free CC.

tion of hen egg-white lysozyme was used: 100 mg ml^{-1} in 50 m*M* sodium acetate buffer pH 4.5. Bovine ribonuclease A was dissolved in 50 m*M* sodium acetate buffer pH 5.5 to a concentration of 10 mg ml^{-1}, whereas *Thaumatococcus daniellii* thaumatin was dissolved in 20 m*M* Na HEPES to a concentration of 35 mg ml^{-1}. Porcine trypsin was dissolved in 100 m*M* benzamidine to a concentration of 15 mg ml^{-1}. The crystallization and cryoprotection conditions are given in Table 1.

3.2. Data collection and processing

All 12 data sets were collected on the European Synchrotron Radiation Facility (ESRF) MAD beamline ID14-EH4 within a single 24 h period in order that conditions would be as similar as possible. Data were collected at a wavelength of 0.939 Å from crystals that were flash-frozen in a 100 K Oxford Cryosystem 600 nitrogen-gas stream. The data-collection scheme was similar to that used by Ravelli *et al.* (2003). For each crystal, a low-dose complete 'before' data set was collected using an attenuated beam. The attenuators were subsequently removed and the crystals were exposed to a high-dose X-ray 'burn' in order to induce radiation damage. A second complete low-dose 'after' data set was collected under identical conditions to those used for the first data set. Great care was taken to use the maximum amount of attenuation that prevented overloads at low resolution but still yielded high-resolution data for both the 'before' and 'after' data sets. The burns were performed by rotating the crystal 360° in the unattenuated beam (16-bunch mode) for 90 s, which corresponds to an average dose of 2×10^6 Gy, as calculated using the program *RADDOSE* (Murray *et al.*, 2004). The dose used for the X-ray burn was, in general, much greater than that used for the collection of the 'before' and 'after' data sets (Table 2).

All data were processed in *XDS*/*XSCALE* (Kabsch, 1988) and exported to merged and unmerged *SCALEPACK* files. The 'before' and 'after' data sets were put on an identical absolute scale using *XSCALE* and then combined in order to produce ΔF values using *SHELXC*. *SHELXC* scales the two data sets with a single scale factor for each of 21 resolution shells, followed by a 'local scaling' method as described by Matthews & Czerwinski (1975) using a sphere of radius $(100/V)^{1/3}$ Å^{-1} around each reflection, where V is the volume of the primitive unit cell. The latest version of *SHELXC* was used, which allowed the 'after' data set to be scaled down by a factor K (see §2).

3.3. RIP phasing

The resolution for phasing (Table 3) was chosen based on the resolution shell in which $\langle\Delta F\rangle/\langle\sigma(\Delta F)\rangle$ is greater than 1.5. The number of sites was set to be equal or slightly larger than the known number of disulfide bridges. ΔF values were calculated using an RIP option in *SHELXC* and used in *SHELXD* (Schneider & Sheldrick, 2002) to determine the most susceptible part of the specific radiation-damage substructure. All *SHELXD* processes were run for 500 cycles in Patterson seeding mode. Solutions from *SHELXD*, identified on the basis of the correlation coefficient CC_{best} between the observed and calculated radiation-damage substructure normalized structure factors, were subsequently submitted to *SHELXE* (Sheldrick, 2002) with 0 cycles without resolving the phase ambiguity. This allowed the identification of sites with negative occupancies, as well as the evaluation of the initial correlation coefficient to the full resolution of the data. The RIP substructures, including both positive and negative sites, were then recycled through *SHELXE* with 100 cycles of density modification. Difference Fourier analysis within *SHELXE* was used to update the RIP substructure, which was resubmitted to *SHELXE*. This last iteration was repeated until phases became interpretable (see also §4). The entire process of ΔF calculation (*SHELXC*) to phase and substructure improvement (*SHELXE*) was performed for up to 70 different scale factors K ranging from 0.90 to 1.03.

Table 4
Refinement statistics.

Values in parentheses are for the outer shell.

	Elastase		Insulin		Lysozyme		Ribonuclease A		Thaumatin		Trypsin	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Resolution (Å)	45.00–1.33 (1.40–1.33)	45.00–1.33 (1.40–1.33)	45.00–1.50 (1.58–1.50)	45.00–1.50 (1.58–1.50)	45.00–1.40 (1.48–1.40)	45.00–1.40 (1.48–1.40)	45.00–1.40 (1.48–1.40)	45.00–1.40 (1.48–1.40)	45.00–1.40 (1.48–1.40)	45.00–1.40 (1.48–1.40)	44.00–1.20 (1.27–1.20)	44.00–1.20 (1.27–1.20)
R_{cryst} (%)	11.4 (8.5)	11.3 (9.4)	12.9 (17.4)	13.1 (18.8)	13.0 (9.3)	12.9 (10.5)	14.5 (15.8)	14.9 (21.2)	12.6 (15.2)	12.5 (15.1)	10.6 (9.3)	10.7 (10.1)
R_{free} (%)	14.0 (13.4)	14.2 (14.6)	16.1 (21.8)	16.5 (21.4)	16.9 (15.1)	16.9 (16.4)	17.6 (22.0)	18.1 (26.7)	15.5 (19.6)	15.4 (19.6)	12.8 (11.9)	12.9 (13.3)
No. of non-solvent atoms	1852	1852	437	437	1021	1021	955	955	1614	1614	1803	1803
No. of solvent atoms	266	266	78	78	115	115	143	143	276	276	314	314
Mean B (Å^2)	7.0	7.1	14.6	14.4	10.2	10.1	11.1	10.7	9.9	9.9	6.5	6.4
Ramachandran plot												
Allowed (%)	85.4	86.4	93.0	93.0	87.6	87.6	91.3	89.6	91.1	91.1	88.8	88.3
Additional (%)	14.6	13.6	7.0	7.0	12.4	12.4	8.7	10.4	8.9	8.9	11.2	11.7
Disallowed (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R.m.s. deviations from ideality												
Bond lengths (Å)	0.019	0.020	0.018	0.021	0.022	0.022	0.017	0.019	0.018	0.017	0.013	0.013
Bond angles (°)	1.4	1.4	1.7	1.7	1.9	1.9	1.6	1.7	1.7	1.7	1.6	1.6
PDB code	2blo	2blq	2bn3	2bn1	2blx	2bly	2blp	2blz	2blr	2blu	2blv	2blw

3.4. Model building and refinement

PDB models (1gwa, 1b2f, 193l, 1kf5, 1lxz and 1h9j) were placed in the experimental density by *FFFEAR* and refined by iterative model-building and refinement cycles in *COOT* (Emsley & Cowtan, 2004) and *REFMAC* (Collaborative Computational Project, Number 4, 1994). The models were refined to convergence based on the 'before' data sets and then subjected to ten refinement cycles against the F_{after} amplitudes. Refinement statistics are given in Table 4. We did not try to manually model radiation-damage-induced new conformations for the 'after' data sets because of the lack of radiation-damage-specific refinement protocols (Banumathi *et al.*, 2004). The phases of the 'before' models were used to evaluate the efficacy of the methods introduced here. As phases are obviously not available during the structure determination of novel proteins, we also inspected the solution-prediction power of figures of merits such as the pseudo-free correlation coefficient as calculated by *SHELXE*.

4. Results and discussion

4.1. Elastase

The RIP signal present between the 'before' and 'after' data sets was quite low, with an R_{int} (defined as $\sum |I - \langle I\rangle|/\sum I$) of 6.8% and an overall $\langle\Delta F\rangle/\langle\sigma(\Delta F)\rangle$ of 1.32. Despite this fact, structure solution was easily achieved and a wide range of F_{after} scales ($0.93529 < K < 0.99286$) yielded interpretable phases without recycling in *SHELXE*. The downscaling of F_{after} was mandatory in this case, as random phases were obtained with a scale factor $K = 1$.

The size of the elastase RIP substructure is somewhat remarkable. After only one round of recycling in *SHELXE*, 316 sites were found at $K = 0.995$ using the *SHELXE* difference Fourier analysis. Manual inspection of the refined model, the RIP substructure and the $F_{before} - KF_{after}$ difference Fourier using model phases showed that many of the peaks did not relate to the specific radiation damage. Fig. 3 shows the peak list sorted on occupancy, as refined by *SHELXE*, *versus* normalized electron density in the weighted F_{before} map. Although the overall correlation is very good, numerous false sites are present at low σ levels. Using an absolute threshold of 4σ in the $F_{before} - KF_{after}$ difference Fourier, a more realistic RIP substructure can be obtained which consists of 161 sites. The spurious peaks were most likely falsely introduced in early *SHELXE* recycling owing to ripples near strong features in the map such as disulfides. Most of these peaks were refined to rather low occupancies, which probably explains why the density modification yielded excellent electron-density maps despite the presence of these false sites.

4.2. Insulin

Insulin can be solved with just the initial sites found by *SHELXD*. No recycling in *SHELXE* was needed to obtain interpretable phases, although the 0-cycle initial *SHELXE* run

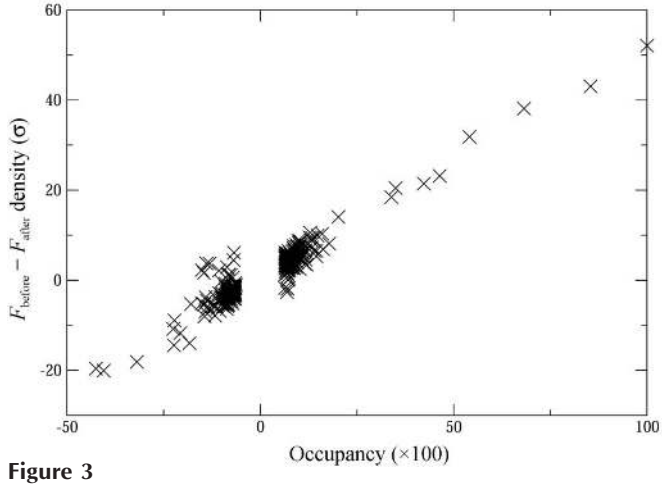

Figure 3
SHELXE-refined occupancies of the radiation-damage substructure of elastase *versus* normalized peak height in the final electron-density map.

did further improve the RIP substructure owing to the presence of strong negative sites. Scaling $F_{\rm after}$ is not necessary for the solution of the insulin structure, but it appears to be just at the boundary of the solvable region of $F_{\rm after}$ scales, which runs from $0.95200 < K < 1.00214$ (Fig. 4*a*). The total time needed to phase insulin using a native crystal was remarkably short. The data collection took 2 × 5 min, the burn 2 min (including resetting the attenuators), data reduction was performed in parallel to data collection and *SHELXC* ($k = 1$), *SHELXD* and *SHELXE* run within 2 min. Thus, the full structure determination took less than 15 min: this is to our knowledge one of the fastest protein structure determinations hitherto. This time could still be significantly reduced using faster-readout detectors.

Fig. 5 shows the $u = 0$ Harker section of the difference Patterson for insulin at different scale factors K. Many spurious peaks are found on the intersections of the $u = 0$ plane with other Harker sections. The height of these peaks is a function of K and reduces as K reduces. However, at the limit $K = 0$, a native Patterson will be obtained which, in the absence of heavy atoms or translation symmetry, would be more or less flat. An optimum is obtained for K values slightly smaller than 1 (Fig. 5), where the spurious peaks are reduced in peak height whereas peaks corresponding to highly susceptible sites are enhanced.

4.3. Lysozyme

Lysozyme has the smallest X-ray burn-induced intensity differences between the damaged and the undamaged data sets: the $R_{\rm int}$ is 4.9% (Table 2). This, together with the very low solvent content, makes lysozyme a difficult structure to solve using RIP. Despite clear improvements in the CCs on applying K (Fig. 4*b*), it was not straightforward to determine accurate phases. Indeed, 11 rounds of *SHELXE* recycling were necessary to obtain the best map (wMPE of 34.1°, pseudo-free CC of 75.2%). Even in this case, only one of the RIP substructures yielded interpretable phases, despite the fact that a range of K values gave non-random starting phases (Fig. 4*b*). This could perhaps have been improved by using more sophisticated substructure and density-modification iteration schemes. Fig. 6 shows the development of the phase errors and part of the

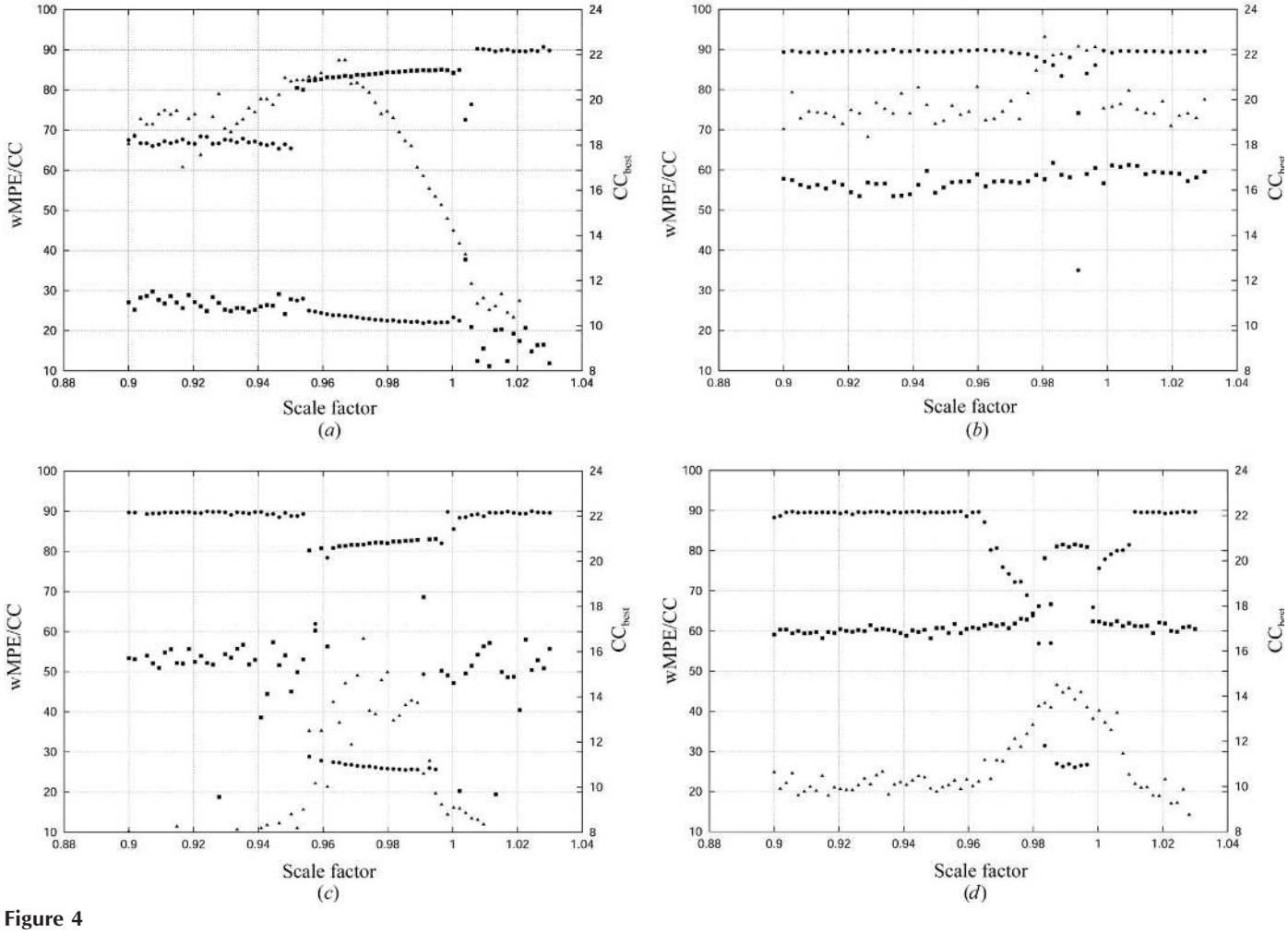

Figure 4
The effect of scale factor on the structure solution of (*a*) insulin, (*b*) lysozyme, (*c*) thaumatin and (*d*) trypsin. CC$_{\rm best}$ is represented by open squares, wMPE is represented by crosses and pseudo-free CC is represented as closed squares.

experimental electron-density map during the subsequent *SHELXE* iterations.

During the first six iterations, the size of the RIP substructure remained rather similar (varying between 13 and 18 atoms). The wMPE improved slightly but the maps remained virtually uninterpretable (Fig. 6). The RIP substructure increased to 25 sites for iteration No. 8 and remained similar up to iteration No. 11. Despite the moderate changes in the substructures, major phase improvements were obtained, especially after iteration No. 9 when the maps became clearly interpretable. Using the very good phase set that was produced after iteration No. 11, an extended RIP substructure was obtained containing more than 40 sites.

4.4. Ribonuclease A

Ribonuclease A had the highest overall $\langle \Delta F \rangle / \langle \sigma(\Delta F) \rangle$ and $R_{\rm int}$ between the 'before' and 'after' data sets, but we were not able to obtain interpretable phases of ribonuclease A *de novo*. Application of the scaling factor K did not succeed in phasing the structure as it did for lysozyme, thaumatin, elastase or

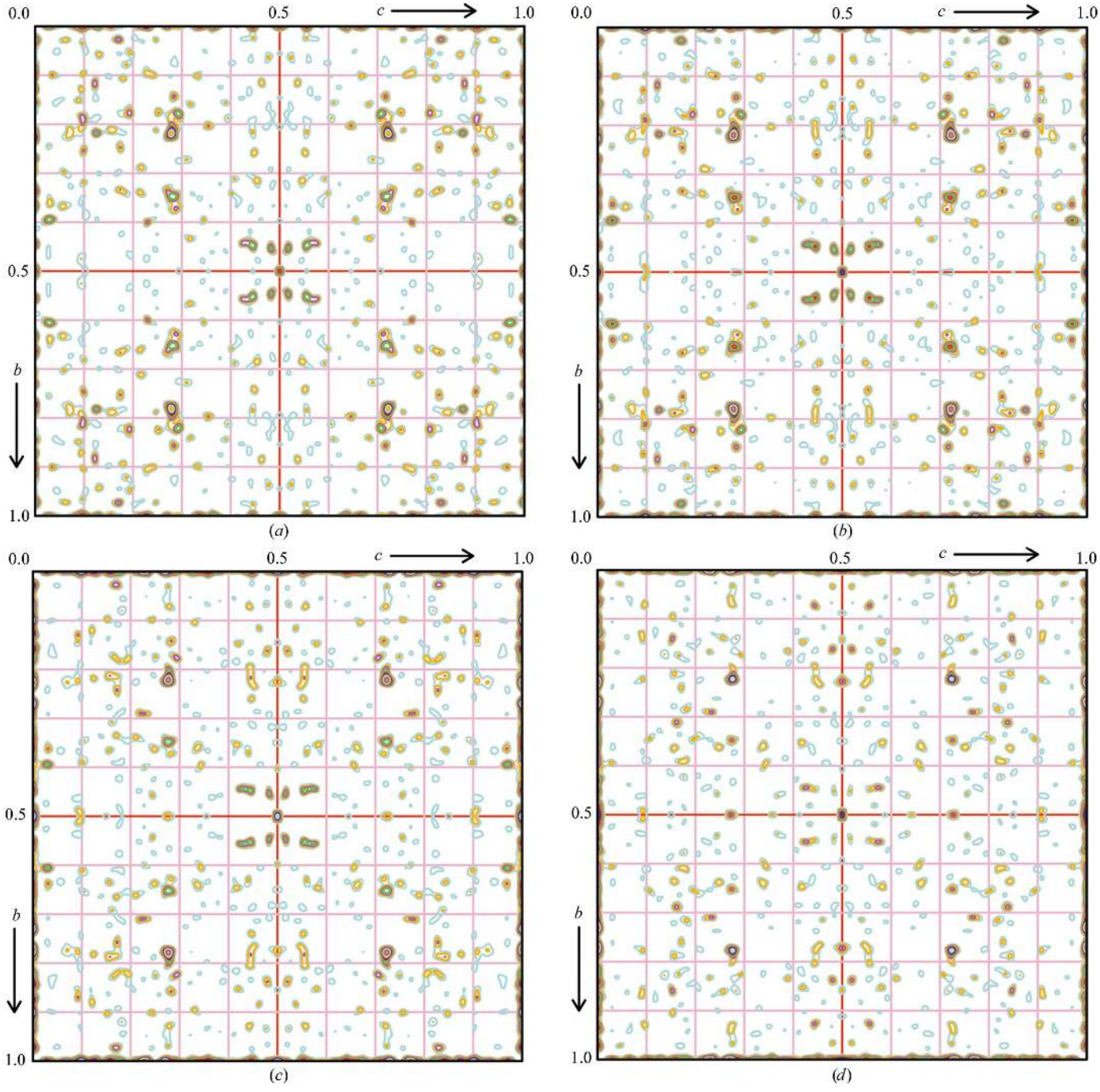

Figure 5
Experimental $u = 0$ Harker sections of radiation-damage difference Pattersons for insulin at different scale factors $K = 0.94$ (*a*), $K = 0.96$ (*b*), $K = 0.98$ (*c*) and $K = 1.00$ (*d*). The origin has been removed and peaks have been sharpened using *XPREP*.

trypsin. Upon examining $F_{\rm before} - KF_{\rm after}$ maps with refined model phases, it became clear that ribonuclease A is somewhat atypical in that despite an obviously sufficient X-ray 'burn', relatively little damage is incurred to the three disulfides. For example, peaks in the $F_{\rm before} - KF_{\rm after}$ model phased maps at the sulfur positions in ribonuclease A ($K = 1$) are maximally 11.9σ, compared with 31.1σ for insulin ($K = 0.99657$), 40.7σ for thaumatin ($K = 0.99471$), 57.4σ for trypsin ($K = 0.99286$) and 47.8σ for elastase ($K = 0.99286$).

The accuracy of the measured intensities differences between the 'before' and 'after' data sets was extremely good, raising the question why it was not possible to phase ribonuclease A *de novo*. Was the failure owing to difficulties in finding the initial RIP substructure or was it impossible to improve a partially correct *SHELXD* solution during the subsequent *SHELXE* iterations? Phase-error calculations showed that the substructures found were incorrect. To further investigate the second point, we used a known RIP substructure as input to *SHELXE*. The RIP substructure was determined by searching the $F_{\rm before} - KF_{\rm after}$ difference map (using model phases) for positive and negative peaks with an absolute value larger than 6σ. This 40-site RIP substructure yielded accurate phases (wMPE of 25.1°, pseudo-free CC 82.5%) after *SHELXE*. To determine whether such a large RIP substructure was necessary to obtain phases, we repeated the procedure with higher thresholds. Substructures with as few as five positions using a peak threshold of $|10\sigma|$ also yielded interpretable electron-density maps (wMPE of 25.2°, pseudo-free CC 82.5%). Interestingly, these sites were not from sulfurs, but from ordered waters and O atoms on the carboxyl group of the C-terminal Val124. Most of the radiation-susceptible sites were found on the surface and played a role in crystal packing. Ribonuclease A is the first example in our hands of a disulfide-containing protein in which none of the disulfide bonds correspond to the sites most susceptible to X-rays. In order to phase ribonuclease A, the five most susceptible sites, all O atoms, had to be given. It would yield promise if newer generation substructure-determination programs would be able to find these sites from strong and accurately measured intensity differences.

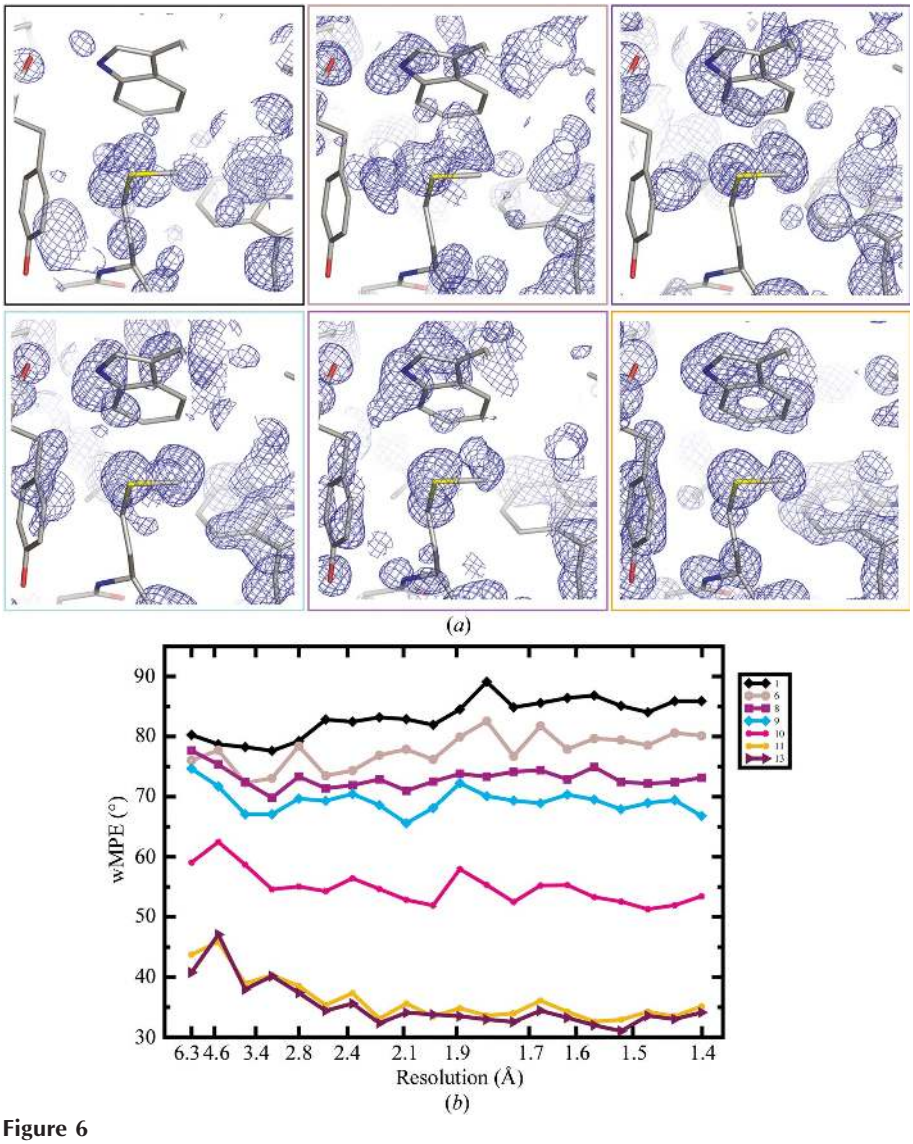

Figure 6
Experimental electron-density maps (*a*) and weighted mean phase errors (*b*) for lysozyme for different subsequent iterations in *SHELXE*. A total of 13 substructure iterations were made; maps were drawn after iteration Nos. 1, 6, 8, 9, 10 and 11.

4.5. Thaumatin

The overall $R_{\rm int}$ of 8.8% and $\langle\Delta F\rangle/\langle\sigma(\Delta F)\rangle$ of 1.7 (differences greater than 1.5σ extended to 1.8 Å) indicated the presence of a strong RIP signal. However, structure solution was not possible without the use of a scale factor K. By contrast, a wide range of $F_{\rm after}$ down-scalings ($0.9557 < K < 0.9947$) yielded interpretable maps with overall wMPEs of less than 30° (Fig. 4*c*). Correlation coefficients of RIP substructures from *SHELXD* were similarly improved from a baseline of 8% to as high as 17%. As for most of the examples given here, a very good inverse relationship was found between the pseudo-free correlation coefficient and the wMPE. Indeed, among all the indicators of correct *SHELXE* solutions (correlation coefficient $E^{\rm obs} - E^{\rm calc}$, contrast, connectivity, FOM), the pseudo-free correlation coefficient seems to be the most reliable for the studies presented here.

4.6. Trypsin

In terms of overall $\langle\Delta F\rangle/\langle\sigma(\Delta F)\rangle$, trypsin had the smallest RIP signal of all of the data series: 1.2. In spite of this, trypsin was readily solved even without *SHELXE* RIP-substructure iteration. Correlation coefficients of the substructures from *SHELXD* were improved from a baseline of 10% to a maximum of 14% (Fig. 4*d*). Using only one cycle of *SHELXE* density modification, interpretable maps were obtained in a very narrow range of $F_{\rm after}$ downscalings ($0.98357 < K < 0.99657$). This range was increased with subsequent recyclings in *SHELXE* to $0.97800 < K < 1.00029$, $0.97057 < K < 1.00400$ and $0.97057 < K < 1.00771$ in the second, third and fourth recycles, respectively. The trypsin data series demonstrates the benefits of both downscaling $F_{\rm after}$ as well as the substructure iteration with *SHELXE*.

5. Future perspectives and conclusions

We are just at the dawn of exploiting specific radiation damage for phasing, either as a primary or secondary source. Whereas it seems straightforward to measure accurate RIP signals, the interpretation is still in its infancy. In this work, we have addressed some of the difficulties in RIP data interpretation and presented ways to overcome them for a number of disulfide-containing model proteins.

The data-acquisition scheme we used consisted of the collection of two low-dose data sets *R* and *T*, interleaved with an X-ray 'burn' using an unattenuated beam. The 'burn' we used in this work was significantly less than that employed by Ravelli *et al.* (2003): 2 to 4×10^6 compared with 8 to 10×10^6 Gy. We feel that the dose used for the first data set *R* is of special importance for the success of RIP: this should be as low as possible. In contrast, successful RIP signals were obtained with a wide range of X-ray doses used for the burn. The 'burn' is perhaps a waste of the precious lifetime of the macromolecular crystal, since valuable high-resolution data could have been collected instead. Instead, it might seem better to collect highly redundant data and subsequently use extrapolation and interpolation schemes to reconstruct constant-dose data sets resembling data sets *R* and *T*. Diederichs *et al.* (2003) have presented a simple method based on curve-fitting individual reflections. However, this method is still purely empirical, as it does not take the decay behaviour of different X-ray-susceptible sites in the macromolecule into account. The most susceptible sites might very well be destroyed within a few percent of the total dose that a crystal can tolerate before its diffractive power is lost. These sites are very important for radiation-damage phasing. If, for example, five complete data sets are collected within the lifetime of a crystal, information about the very early states of radiation damage could easily be lost. Our method of collecting the first data set *R* with maximum attenuation obviate this difficulty, although future extrapolation and interpolation schemes will hopefully overcome this need.

The later data set *T* was collected on a crystal with macroscopic (*e.g.* higher mosaicity) and microscopic (altered molecular structure) damage. The overall crystalline diffraction will be less, an effect that is ignored by current scaling programs. We have shown that downscaling the 'after' data set *T* helps for all subsequent steps. For some cases (elastase, lysozyme and thaumatin) downscaling was mandatory for solution of the structure and in other cases (insulin and trypsin) it facilitated structure determination. We determined *K* in a brute-force manner by testing a wide range between 0.90 and 1.03. The behaviour of the Pattersons (Fig. 5) and $\mathrm{CC}_{\rm best}$ *versus* the scale factor *K* (Fig. 4) is, interestingly, an excellent predictor of the solvability of an RIP data set. That is to say, all data sets which were solvable by RIP showed a clear peak in the graph of $\mathrm{CC}_{\rm best}$ *versus K*. This peak generally coincided with the lowest phase errors in subsequent *SHELXE* iterations. Only ribonuclease A did not show such a peak and it was not possible to solve this structure starting from any *SHELXD* substructure. It would be useful if *K* could be determined by a method other than by brute force. Future programs might exploit unmerged data for this in order to derive a model of the intensity decay caused by radiation damage (Blake & Phillips, 1962), which is likely to have some relation to *K*, or perhaps maximize the peak heights of non-special position peaks in difference Pattersons.

RIP can be seen as MIR in which the native data set is missing. The difference Patterson can contain both positive and negative peaks, the latter arising from cross-vectors. Current Patterson-based substructure-determination programs ignore this fact, which might partly explain the low values of the different figures of merit used to identify a correct solution. In fact, for SAD, similar CC values have led to correct structure solutions (Debreczeni *et al.*, 2003), but in general much larger values are found with better discrimination from the wrong solutions. Future generations of substructure-determination programs that take negative cross-vectors into account might give a sharper contrast between correct and false solutions, as well as identify substructures that are currently being missed.

The density-modification algorithm *SHELXE* was of crucial importance to our work. It was primarily responsible for the high quality of the final maps, as well as for the ability to improve the poor initial phases (Table 3). In most cases, RIP structure determination starts with a rather incomplete substructure, thus compromising the quality of the initial maps. The substructure was improved after each *SHELXE* cycle by peak searching the $F_{\rm before} - KF_{\rm after}$ difference map while using phases obtained from the current protein model plus the estimated phase shift α (Sheldrick, 2002). This process was not improved by positional refinement, possibly because of the large number of weak sites which were often located close to each other. The calculation of difference Fouriers inevitably biased the substructures towards falsely introduced sites, probably explaining why a large number of false sites were found for elastase (§4.1). More sophisticated algorithms, such as the calculation of log-likelihood gradient maps (de La Fortelle & Bricogne, 1997), combined with robust filtering schemes to delete spurious sites might yield improvements, thus reducing the number of necessary *SHELXE* iterations.

The crystals we used as examples all diffracted to high resolution. The quality of the phase refinement in *SHELXE* is dependent on both the resolution and the solvent content. The latter was in general very low for the samples we tested, making us optimistic about the potential success of RIP on weaker diffracting samples that have higher solvent content. However, the resolution can also be of importance for locating the low-occupancy radiation-damage-induced new sites, which are often found in close proximity to the original sites.

It is tempting to compare the RIP method with long-wavelength sulfur SAD. These techniques are not mutually exclusive and have in fact been used in combination to phase crystals *de novo* (Weiss *et al.*, 2004). There are important similarities and differences to be discussed. Both techniques can at the very least provide additional phase information and in some cases *de novo* phases. Furthermore, both techniques have the attractive technical benefit of not requiring any modification to the protein or crystal. Often, only one crystal is needed. Sulfur SAD benefits from the presence of disulfide bonds, but can also work when only free cysteines and methionines are present, while for the moment RIP seems to require the presence of disulfide bonds. The Bijvoet differences are generally very small for sulfur SAD, a few percent or less, owing to the small values of f'' of sulfur at wavelengths typically used for sulfur SAD (0.56 e^- at 1.54 Å wavelength). By contrast, RIP signals can be large (5–15%; see Table 2) and are related to the atomic number, the fraction of atoms that are damaged and the significance of the movement of the X-ray-susceptible atom. Although RIP explores radiation damage, care must be taken to avoid introducing excessive non-isomorphism upon prolonged exposure. No SAD signal can be obtained for centric reflections, whereas it is still possible to measure RIP signals for these reflections. A further advantage of the RIP method is that it can be performed at shorter wavelengths, thus overcoming the need for absorption correction and facilitating the collection of higher resolution data. Nevertheless, there are still many improvements needed to make RIP as routine as sulfur SAD. In this paper, we have made the initial steps.

We gratefully acknowledge the staff of the EMBL Grenoble Outstation and the ESRF for support, maintenance and development of the JSBG beamlines and the ESRF for beamtime under MX347. Trypsin crystals were kindly provided by H.-K. Schrøder Leiros. The work was supported by FP6 EU BioXhit grant under contract No. LHSG-CT-2003-503420, as well as the FP6 EU I3 Project IA-SFS grant under contract No. RII3-CT-2004-506008. GMS thanks the Fonds der Chemischen Industrie for support.

References

Banumathi, S., Zwart, P. H., Ramagopal, U. A., Dauter, M. & Dauter, Z. (2004). *Acta Cryst.* D**60**, 1085–1093.
Blake, C. C. F. & Phillips, D. C. (1962). *Biological Effects of Ionizing Radiation at the Molecular Level*, pp. 183–191. Vienna: IAEA.
Blessing, R. (1997). *J. Appl. Cryst.* **30**, 176–177.
Burmeister, W. P. (2000). *Acta Cryst.* D**56**, 328–341.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Debreczeni, J. É., Bunkoczi, G., Ma, Q., Blaser, H. & Sheldrick, G. M. (2003). *Acta Cryst.* D**59**, 688–696.
Diederichs, K., McSweeney, S. & Ravelli, R. B. G. (2003). *Acta Cryst.* D**59**, 903–909.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Evans, G., Polentarutti, M., Djinovic Carugo, K. & Bricogne, G. (2003). *Acta Cryst.* D**59**, 1429–1434.
Howell, P. L. & Smith, G. D. (1992). *J. Appl. Cryst.* **25**, 81–86.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.
Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958). *Nature (London)*, **181**, 662–666.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472-494.
Matthews, B. W. & Czerwinski, E. W. (1975). *Acta Cryst.* A**31**, 480–487.
Murray, J. W., Garman, E. F. & Ravelli, R. B. G. (2004). *J. Appl. Cryst.* **37**, 513–522.
Nikonov, S. (1983). *Acta Cryst.* A**39**, 693–697.
Ravelli, R. B. G., Leiros, H. K., Pan, B., Caffrey, M. & McSweeney, S. (2003). *Structure*, **11**, 217–224.
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.
Rice, L. M., Earnest, T. N. & Brunger, A. T. (2000). *Acta Cryst.* D**56**, 1413–1420.
Rossmann, M. (1960). *Acta Cryst.* **13**, 221–226.
Schiltz, M., Dumas, P., Ennifar, E., Flensburg, C., Paciorek, W., Vonrhein, C. & Bricogne, G. (2004). *Acta Cryst.* D**60**, 1024–1031.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Sheldrick, G. M. (2002). *Z. Kristallogr.* **217**, 644–650.
Weik, M., Berges, J., Raves, M. L., Gros, P., McSweeney, S., Silman, I., Sussman, J. L., Houee-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.
Zwart, P. H., Banumathi, S., Dauter, M. & Dauter, Z. (2004). *Acta Cryst.* D**60**, 1958–1963.