Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Segmenting data sets for RIP

Daniele de Sanctis[a] and Max H. Nanao[b,c]*

[a]Structural Biology Group, European Synchrotron Radiation Facility, 6 Rue Jules Horowitz, 38043 Grenoble, France, [b]European Molecular Biology Laboratory, 6 Rue Jules Horowitz, BP 181, 38042 Grenoble, France, and [c]Unit of Virus Host Cell Interactions, UJF–EMBL–CNRS, UMI 3265, 6 Rue Jules Horowitz, BP 181, 38042 Grenoble, France

Correspondence e-mail: mnanao@embl.fr

Received 11 April 2012
Accepted 22 May 2012

Specific radiation damage can be used for the phasing of macromolecular crystal structures. In practice, however, the optimization of the X-ray dose used to 'burn' the crystal to induce specific damage can be difficult. Here, a method is presented in which a single large data set that has not been optimized in any way for radiation-damage-induced phasing (RIP) is segmented into multiple sub-data sets, which can then be used for RIP. The efficacy of this method is demonstrated using two model systems and two test systems. A method to improve the success of this type of phasing experiment by varying the composition of the two sub-data sets with respect to their separation by image number, and hence by absorbed dose, as well as their individual completeness is illustrated.

1. Introduction

Although practitioners of macromolecular crystallography (MX) expend significant effort in minimizing radiation damage, specific radiation damage can be useful for the calculation of phases. Radiation-damage-induced phasing (RIP) is a method by which these specific changes are exploited in order to determine phases (Evans *et al.*, 2003; Ravelli *et al.*, 2003, 2005; Banumathi *et al.*, 2004; Schiltz *et al.*, 2004; Weiss *et al.*, 2004; Nanao *et al.*, 2005; Nanao & Ravelli, 2006; Schiltz & Bricogne, 2007; Rudiño-Piñera *et al.*, 2007; Schönfeld *et al.*, 2008; Fütterer *et al.*, 2008; de Sanctis *et al.*, 2011; Panjikar *et al.*, 2011). A typical RIP experiment consists of collecting a complete low-dose data set, followed by an X-ray 'burn' and the collection of a second complete low-dose data set. The first data set (the 'before' data set) and the second data set (the 'after' data set) are then used in a manner analogous to the single isomorphous replacement (SIR) method. One common problem encountered by practitioners of this method is the determination of the optimal dose used to 'burn' the crystal. Too great a dose will introduce noise that will swamp the signal from specific radiation-damage sites, while too low a dose will not result in sufficient damage to specific sites (Ramagopal *et al.*, 2005). In practice, even with calibrated photodiodes that provide estimates of the X-ray flux (and from which the absorbed dose can be derived) and specialized software for dose calculation (Murray *et al.*, 2005), it is not straightforward to calculate an optimal dose for RIP, as it may vary from protein to protein (Leal *et al.*, 2011). Once a dose has been settled upon, even under optimal circumstances, determination of the radiation-damage substructure is the most failure-prone step of the RIP process (Nanao *et al.*, 2005). We have previously described a method for improving

Table 1
Overall data-collection statistics.

Values in parentheses are for the outermost resolution shell.

	C3a	Conkunitzin-S1	Thaumatin	Trypsin
Space group	$P6_322$	$P6_3$	$P4_12_12$	$P2_12_12_1$
Unit-cell parameters (Å)	$a = b = 63.48$, $c = 105.5$	$a = b = 50.89$, $c = 42.43$	$a = b = 58.01$, $c = 150.4$	$a = 54.17$, $b = 57.07$, $c = 65.97$
Resolution (Å)	60.0–2.29 (2.43–2.29)	60.0–1.05 (1.11–1.05)	60.0–1.25 (1.33–1.25)	60.0–1.03 (1.09–1.03)
Wavelength (Å)	1.0723	0.9537	0.9199	0.8726
Completeness (%)	100 (100)	98.2 (98.4)	99.7 (97.9)	96.8 (88.3)
No. of observed reflections	121786 (19809)	292336 (36747)	712431 (38337)	3244992 (355285)
No. of unique reflections	6119 (952)	28704 (4387)	65389 (8038)	98295 (13844)
Multiplicity	19.9 (20.8)	10.2 (8.4)	10.9 (4.7)	9.13 (7.10)
$\langle I/\sigma(I)\rangle$	28.48 (4.19)	23.39 (6.01)	21.56 (3.06)	32.25 (2.17)
$R_{\rm merge}$† (%)	8.8 (84.6)	6.2 (45.6)	5.9 (41.8)	8.1 (168.5)
$R_{\rm r.i.m.}$‡ (%)	9.1 (86.8)	6.6 (48.6)	6.2 (47.0)	7.9 (74.7)
Mosaicity (°)	0.22	0.16	0.13	0.30

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$. ‡ $R_{\rm r.i.m.} = \sum_{hkl}\{N(hkl)/[N(hkl)-1]\}^{1/2}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$.

the success rate of RIP substructure determination by altering the scaling between RIP data sets as well as ‘recycling’ substructures: revising RIP substructures upon successive rounds of density improvement (Nanao *et al.*, 2005). Nevertheless, there still remains a need for experimental and computational methods development that can improve the success rate of RIP substructure determination. Here, we describe a simple method in which a single large oscillation data set is segmented into sub-data sets for RIP substructure determination and phasing. The method is well adapted to the trend towards faster detectors that result in data sets containing extremely large numbers of images and has been shown to be effective on diverse sample types and beamlines. Finally, we demonstrate that the RIP signal and structure solution can be optimized by altering the image ranges of the sub-data sets.

2. Methodology

Data are collected with no specific effort to optimize for segmented RIP. The only difference from typical data-collection strategies is that a sufficient number of images are collected in order to obtain a highly redundant data set. The data set is first processed with *XDS* (Kabsch, 2010*b*) and then segmented into two sub-data sets with different image ranges using the DATA_RANGE keyword and CORRECT job. The segmentation scheme is shown in Fig. 1. The ‘before’ sub-data set is defined as the set of images beginning at image 1 and including as many images as is necessary to attain a certain completeness threshold (90% unless noted otherwise). The ‘after’ sub-data set is defined as the set of images beginning with the last image in the data set and extending as many images towards the beginning of the data set to attain the same completeness as the ‘before’ data set. Once these image ranges have been defined, the ‘before’ and ‘after’ data sets are scaled together with *XSCALE* (Kabsch, 2010*a*) and then converted to *SCALEPACK* format using *XDS2SCA* (R. B. G. Ravelli, unpublished work). The resolution limits for the before and after data sets are set based on the resolution of the full data set, so in some cases (particularly for the after data sets) outer shell statistics can be of reduced quality. Structure determination then proceeds as has been described previously (Nanao *et al.*, 2005). Briefly, isomorphous signal strength is assessed with *SHELXC* (Sheldrick, 2010) using multiple scale factors defined by the DSCA keyword. This step uses the scaled before and after *SCALEPACK* files to produce input files and $F_{\rm A}$ values. At each scale factor, a substructure determination is performed in *SHELXD* (Sheldrick, 2010), with a resolution cutoff generally set to the resolution at which the $\langle d'/{\rm sig}\rangle$ for isomorphous differences drops below 1.3. For each *SHELXD* run, phase calculation in *SHELXE* (Sheldrick, 2010) with no density modification (-m0) is then performed. A revised RIP substructure is then recycled back into *SHELXE* for subsequent rounds of density modification with >0 cycles of density modification (usually 500). At each stage of density modification, weighted mean phase errors (wMPEs) are calculated against a refined reference data set using *CPHASEMATCH* (Winn *et al.*, 2011). To assess the success of substructure determination for a particular scale factor k, the best substructure from each *SHELXD* run is compared against a reference substructure using *phenix.emma* (Adams *et al.*, 2010). The program *ANODE* (Thorn & Sheldrick, 2011) is used to assess RIP peak height at damaged sites and to determine the reference substructure by peak-searching a model-phased $k(F_{\rm before} - F_{\rm after})$ RIP difference map at a threshold of 8σ. This analysis was intended to provide a rough estimate of the general success of substructure determination, but it should be noted that the reference substructure does not contain negative peaks. However, we feel that this is

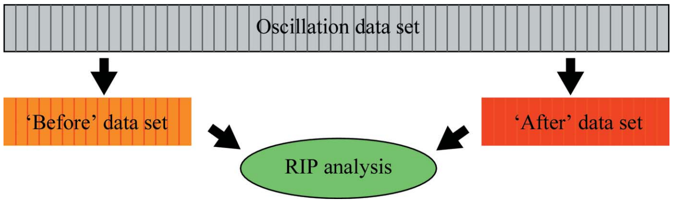


Figure 1
Schematic diagram of the segmented RIP method. Images from a normally collected oscillation data set are used to create two sub-data sets, the ‘before’ (orange) and ‘after’ (red) data sets, which are then subjected to RIP analysis (green).

acceptable since *SHELXD* cannot identify negatively occupied sites and therefore this reference substructure represents the best substructure that *SHELXD* could produce.

3. Model systems

3.1. Thaumatin

3.1.1. Experimental details. Thaumatin crystals in space group $P4_12_12$ were obtained by mixing 1 *M* potassium tartrate, 0.1 *M* HEPES pH 7, 15% glycerol with commercially available thaumatin (from *Thaumatococcus daniellii*; Sigma–Aldrich; 22.2 kDa) which had previously been incubated with 10 m*M* 5-amino-2,4,6-tribromoisophthalic acid (B3C; Beck *et al.*, 2010) in order to derivatize the protein. Crystals grew as tetragonal bipyramids of approximate dimensions 30 × 80 × 80 µm with one protomer in the asymmetric unit. The crystals were transferred into mother liquor supplemented with 25% glycerol prior to flash-cooling. Diffraction data were collected on beamline ID29 at the ESRF (de Sanctis *et al.*, 2012) equipped with an MD2 microdiffractometer and a Pilatus 6M detector at a wavelength of 0.9199 Å in shutterless mode and with a Gaussian beam size of 50 × 30 µm FWHM (full-width at half-maximum). A total of 1800 frames were collected with an acquisition time of 80 ms for each 0.1° angle increment. The synchrotron mode was four-bunch, with a ring current of 34 mA. Data were collected with 8% transmission to give a photon flux at the sample position of 6.5×10^{10} photons s^{-1} at the maximum resolution of 1.25 Å (Table 1). The total absorbed dose as calculated using *RADDOSE* (Murray *et al.*, 2005; Paithankar *et al.*, 2009; Paithankar & Garman, 2010) was 3.4 MGy.

3.1.2. Structure determination. The structure of thaumatin was readily determined using the segmented RIP method. The isomorphous signal was 2.37 for 60–1.03 Å at $k = 1$. The parameters used for structure determination were 16 values of k ($0.88 > k > 1.00$), a resolution cutoff of 1.35 Å for substructure solution and 500 cycles of *SHELXE*. The maximum substructure correctness varied between 15 and 73% (Fig. 2*a*). Interestingly, there were some values of k for which correct substructures could be determined but no interpretable maps could be calculated. Conversely, there were other values of k for which the substructures were relatively inaccurate but could be bootstrapped to produce interpretable maps. These situations happened at low and high k, respectively. Pseudo-free correlation coefficients after the final cycle were as high as 81.0%. Excellent maps with low wMPEs (down to 21.4°) were produced across a broad range of k values ($0.92800 < k < 0.9920$). Indeed, even after the initial *SHELXE* run which used 0 cycles the wMPE was as low as 75.0°. Only a single round of *SHELXE* with 500 cycles was necessary to solve the structure in all cases. *ANODE* analysis revealed the damage sites to occur at bromines from the B3C compound (peak heights of 30.1σ, 28.7σ and 25.5σ for the three Br atoms) as well as the S^{γ} positions of disulfide-bonded cysteines (peak heights of 11.9–20.8σ). New positions for S^{γ} positions also appeared, but were less significant at −8.9σ and −8.1σ. All peak heights are at the optimal value of k (0.944). While RIP was successful, all attempts to solve the thaumatin structure using the anomalous signal from Br atoms failed. This is not surprising considering that the anomalous signal recorded is extremely weak, with $\langle d''/\mathrm{sig}\rangle$ below 1 at 6 Å. We presume that this is a consequence of both low occupancy of the B3C compound and rapidly progressing radiation damage.

3.1.3. Phasing optimization by altering sub-data-set boundaries. If one varies the number of unused images between the before and after data sets, it is possible that a 'sweet spot' could occur which minimizes global radiation damage and maximizes the specific radiation damage. In other words, although our first attempt at segmenting data sets for RIP made use of sub-data sets that were spaced with a maximum number of intervening images, since no explicit attempts were made to provide an optimum dose per image it is possible that such a sweet spot could occur earlier in the data collection (*i.e.* with fewer intervening images between the before and after data sets). To study this possibility, the same phasing process as described previously was used but the composition of the after data set was varied. Specifically, the last image of the after data set was shifted 100 images at a time towards the before data set (images 1–776, φ range 0–77.6°, approximately 1.46 MGy) and sufficient images were then added to the beginning of this new after data set to achieve 90% completeness overall. The same k range was used as for the initial structure solution, with two recyclings in *SHELXE*. Several metrics of signal strength, substructure solution and phase quality were used. These included the height of the strongest peaks in RIP difference maps, the overall $\langle I/\sigma(I)\rangle$ on RIP differences calculated in *XPREP* (Bruker AXS), the best wMPE from each run, the best *SHELXD* CFOM from each run, the mean solution completeness across all values of k for a particular image-range run and finally the percentage of k values that produced interpretable phases. This last metric used a wMPE cutoff of 50.0°, which was chosen as the threshold of map interpretability in all test cases. Using these analyses, we were able to make several conclusions about the effects of sliding the after data set closer to the before data set for this B3C-derivatized thaumatin data set. The first is that there was an increase in the RIP signal as the after data set was moved away from the before data set [from 1.39 to 2.37 in overall $\langle I/\sigma(I)\rangle$], with a concomitant slight rise in the RIP difference-map peak heights (29.8σ to 34.2σ) as well as an increase in the mean solution completeness from *SHELXD* (Fig. 3*a*). Similarly, the percentage of k values that produced interpretable phases also increased dramatically from 27% to 67%. Interestingly, in spite of the improvement in the signal and the success rate of substructure solution observed with larger gaps, interpretable maps (wMPEs of 21.0–22.0°) were observed even with very small gaps between the before and after data sets. Indeed, in two of the three cases where there was overlap between the before and after data sets (after data-set ranges of 659–1100 and 738–1200 with 117 and 38 images of overlap, respectively) the structure could still be solved. These two cases had overlap doses of 0.22 and 0.07 MGy and overall doses (before, gap and after) of 2.08 and 2.27 MGy. Thus, the overall conclusions of this analysis are that while the facility of structure solution can be improved by increasing the gap between the before and after data sets, total doses as small as 2.08 MGy are sufficient to introduce sufficient signal for substructure determination and phasing. Because the regions of overlap between the before and after data sets use some of the same reflections, it is likely that they contribute very little to substructure determination. Therefore, another way of looking at the overlap case is that phasing can succeed even with reduced completeness for the before and after data sets. We therefore analyzed the limits of completeness necessary for this method.

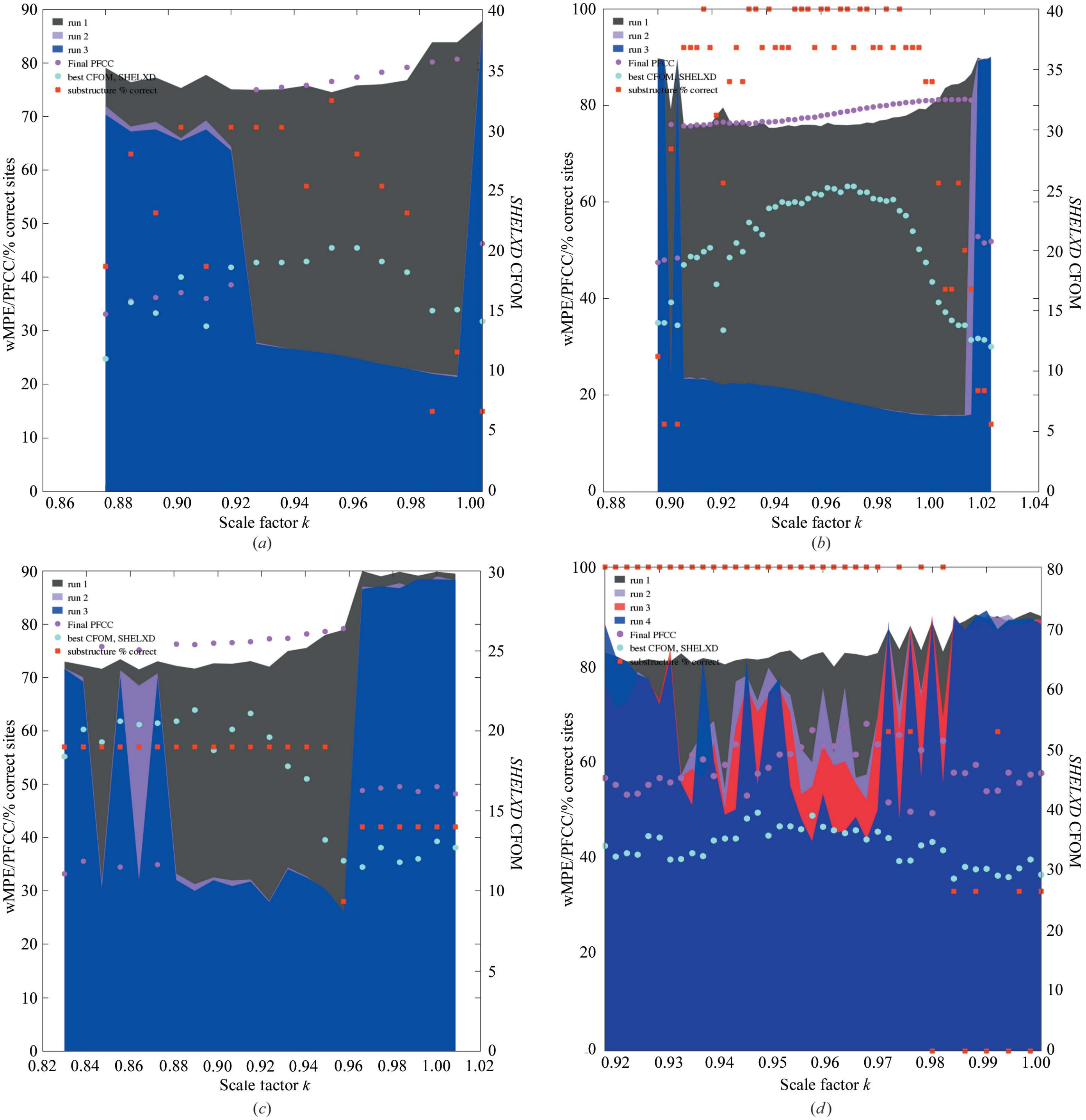


Figure 2
RIP using segmented sub-data sets for thaumatin (*a*), trypsin (*b*), conkunitzin-S1 (*c*) and C3a (*d*). Phase errors for individual density-modification runs are represented as semi-transparent area plots. Run 1 is the initial *SHELXE* run with 0 cycles. Subsequent runs are with 500 *SHELXE* cycles (1000 in the case of C3A) and use revised sites from difference RIP maps. The final pseudo-free correlation coefficient is shown in pink circles and the best CFOM is shown in cyan circles. The correctness (%) of the best substructure from a run with a given *k* is shown in red boxes.

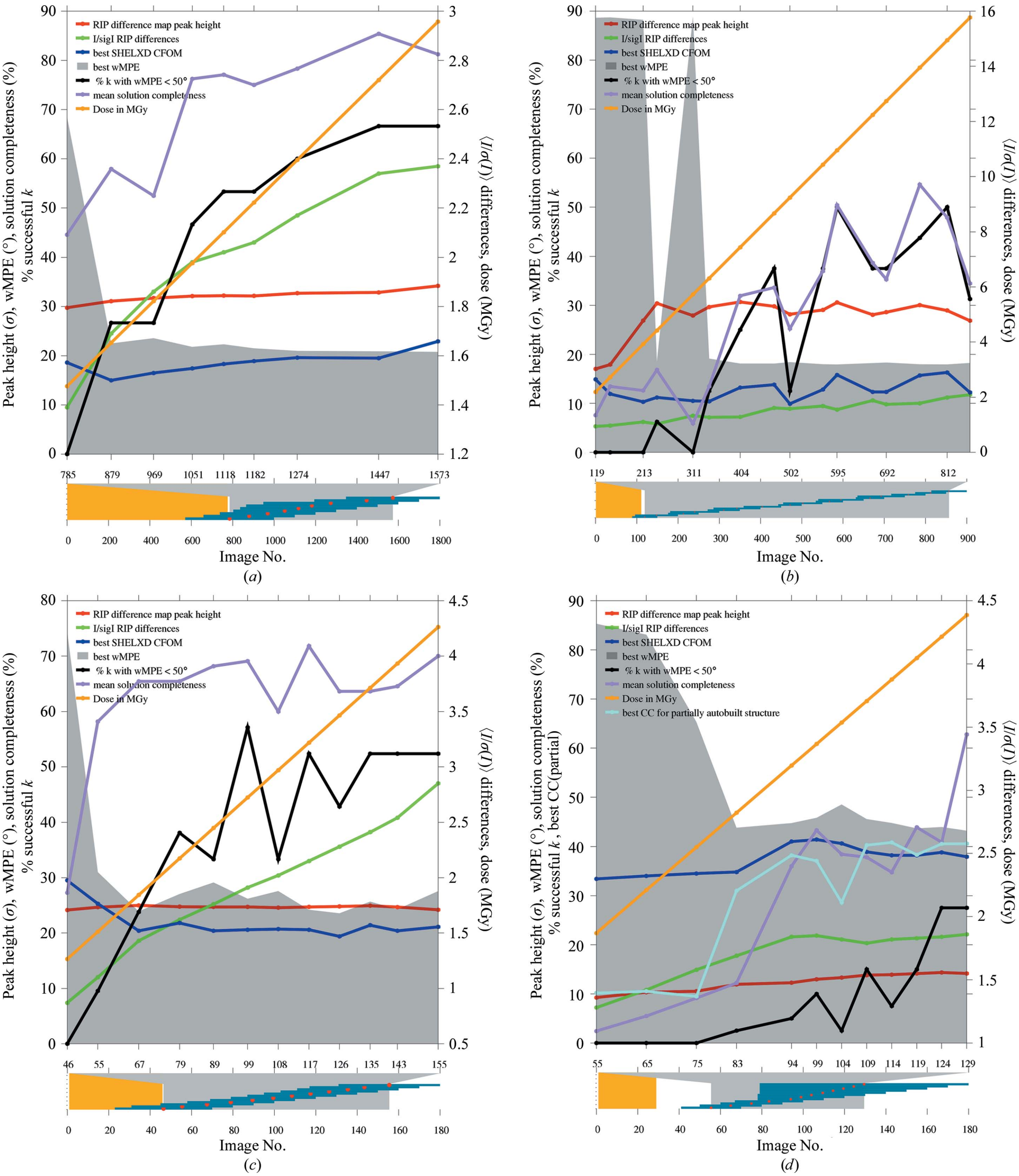


Figure 3

The effect of varying the position of the after data set for thaumatin (*a*), trypsin (*b*), conkunitzin-S1 (*c*) and C3a (*d*). The upper panel shows metrics of signal strength [RIP difference-map peak height in red, $\langle I/\sigma(I)\rangle$ of RIP differences in green], substructure quality (best *SHELXD* CFOM in blue, percentage of k runs that produced phases sets with overall wMPEs of less than 50° in black and mean substructure solution completeness in purple), phase errors (wMPE compared with phases from a refined reference model in grey) and dose (orange). For C3a (Fig. 3*d*) an additional metric is shown, which is the best correlation coefficient (CC) of the partial model that was automatically built by *SHELXE*. The lower panel maps the image range of the upper panel to the complete unsegmented data set. The extent of the before data set is shown in yellow and those of the various after data sets as blue horizontal bars with a red dot indicating the middle of each data set.

3.1.4. Completeness analysis. Because we were able to obtain interpretable phases with overlapping data sets and because it might not always be practical to collect data sets of adequate redundancy to segment into two complete sub-data sets, we investigated the possibility of reducing the sub-data-set completeness. The effects of completeness upon successful structure determination were studied by varying the target percentage completeness of the before and after data sets and by using the same basic RIP strategy as described earlier but with one important modification. Because extremely incomplete data sets could cause problems for density modification, we introduced the concept of a three-data-set RIP experiment. In this analysis, two sub-data sets are used for substructure determination as usual and a third highly redundant complete

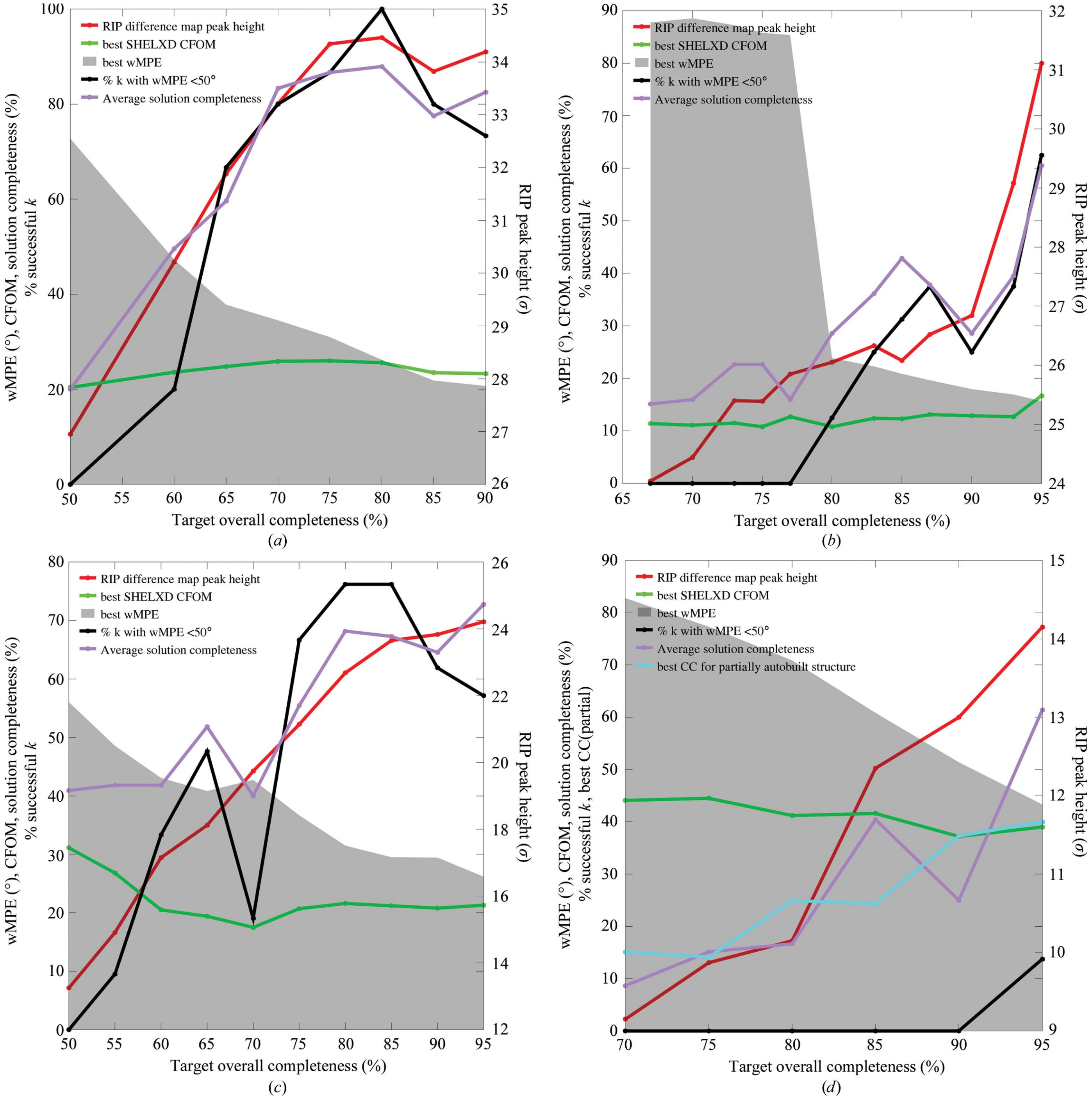


Figure 4

Completeness analysis for thaumatin (*a*), trypsin (*b*), conkunitzin-S1 (*c*) and C3a (*d*). wMPE is shown in grey. The strongest RIP difference-map peak heights are shown in red, the best *SHELXD* CFOM in green and the percentage of *k* runs that produced overall wMPEs of less than 50° in black. Average solution completeness for all values of *k* for a given completeness is shown in purple. For C3a (*d*) an additional metric is shown in cyan, which is the best CC for the structure partially automatically built by *SHELXE*.

Table 2
Segmented data-collection statistics.

Values in parentheses are for the outermost resolution shell.

	C3a		Conkunitzin-S1		Thaumatin		Trypsin	
	After	Before	After	Before	After	Before	After	Before
Frame range	79–180	1–29	131–180	1–46	1347–1800	1–776	814–900	1–110
Angular range (°)	101	29	49	46	45.3	77.6	86	110
Space group	$P6_322$	$P6_322$	$P6_3$	$P6_3$	$P4_12_12$	$P4_12_12$	$P2_12_12_1$	$P2_12_12_1$
Unit-cell parameters (Å)	$a = b = 63.65$, $c = 105.43$	$a = b = 63.65$, $c = 105.44$	$a = b = 50.96$, $c = 42.49$	$a = b = 50.89$, $c = 42.41$	$a = b = 57.95$, $c = 150.23$	$a = b = 57.92$, $c = 150.15$	$a = 54.21$, $b = 57.09$, $c = 65.99$	$a = 54.15$, $b = 57.05$, $c = 65.98$
Resolution (Å)	60.0–2.28 (2.42–2.28)	60.0–2.28 (2.42–2.28)	60.0–1.05 (1.11–1.05)	60.0–1.05 (1.11–1.05)	60.0–1.25 (1.33–1.25)	60.0–1.25 (1.33–1.25)	60.0–1.03 (1.09–1.03)	60.0–1.03 (1.09–1.03)
Completeness (%)	95.7 (97.8)	94.8 (93.9)	95.3 (89.3)	94.9 (92.7)	89.8 (63.1)	89.8 (59.0)	98.2 (96.7)	97.8 (96.7)
No. of observed reflections	70890 (11552)	19847 (3144)	82198 (27871)	73866 (27747)	186084 (64604)	316220 (64602)	328271 (63409)	360710 (63193)
No. of unique reflections	5940 (953)	5880 (915)	10330 (3982)	9326 (4133)	10883 (7567)	18469 (7078)	52155 (10826)	59262 (10829)
Multiplicity	11.9 (12.1)	3.4 (3.4)	7.9 (7.0)	7.9 (6.7)	17.1 (8.5)	17.1 (9.1)	6.3 (5.8)	6.1 (5.8)
$\langle I/\sigma(I)\rangle$	20.73 (1.72)	15.83 (2.61)	15.60 (1.65)	18.59 (6.44)	11.73 (1.21)	18.64 (2.51)	18.74 (2.65)	23.87 (4.96)
$R_{\rm merge}$ (%)	10.4 (147.2)	5.3 (48.2)	5.1 (63.6)	3.5 (13.3)	4.7 (51.2)	4.1 (34.2)	5.8 (56.8)	4.3 (33.8)
$R_{\rm r.i.m.}$ (%)	10.9 (154.0)	6.2 (56.7)	6.3 (79.6)	4.4 (16.8)	5.6 (68.7)	4.6 (41.2)	6.5 (63.6)	4.8 (37.2)

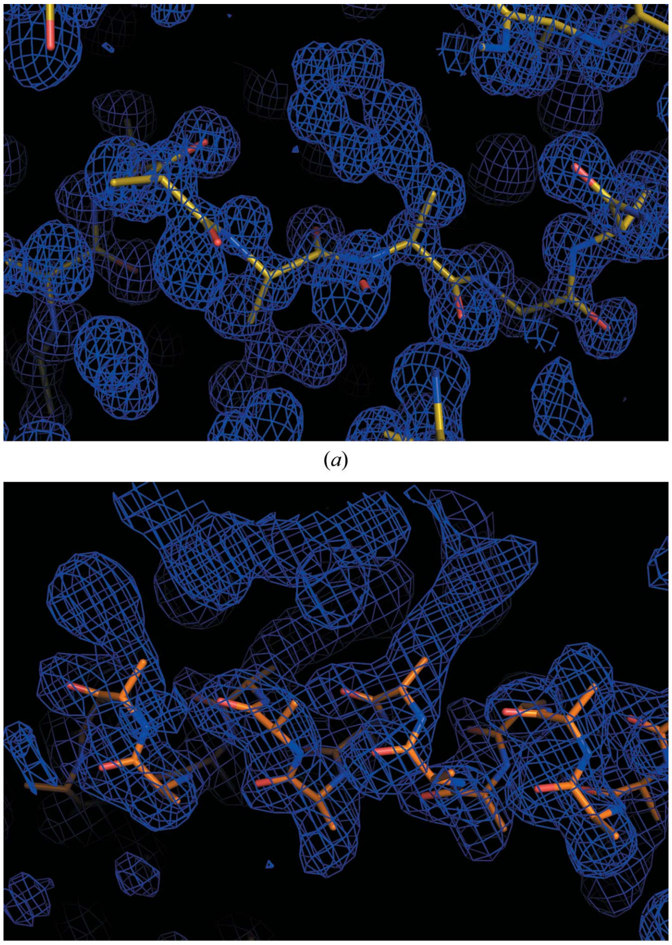
sub-data set is used for phasing and density modification. This third sub-data set consists of the images required for 95% complete data. In producing sub-data sets with decreasing completeness, it should be noted that the *overall* completeness was studied rather than attempting to pick images that provided a uniform completeness across all resolution shells. Owing to a combination of the crystal orientation and space group and the anisotropy of the sample, adding or subtracting consecutive images from sub-data sets caused a non-uniform completeness to accumulate as a function of resolution. Specifically, low-resolution shells typically had a lower completeness than the overall target value, while high-resolution shells had a higher completeness compared with the target overall value for this crystal. Several metrics were analyzed for this study: RIP difference-map peak height, best *SHELXD* CFOM, best wMPE, the percentage of k values that produced interpretable maps and the mean solution completeness. The conclusions of this analysis were that RIP peak heights, the percentage of k runs that produced interpretable maps, the lowest wMPEs and the average solution completeness all improved dramatically with higher completeness (Fig. 4*a*). Interestingly, completenesses down to 60% could still produce interpretable maps. It therefore appears that in high-resolution, high-signal cases, relatively incomplete data sets can still be successfully used for substructure solution and phasing.

3.2. Trypsin

3.2.1. Experimental details. Primitive orthorhombic crystals of bovine pancreatic trypsin (23.8 kDa) were grown in 100 m*M* benzamidine, 3 m*M* $CaCl_2$, 2 *M* ammonium sulfate, 0.1 *M* Tris–HCl pH 8.5 and were cryoprotected in the same solution supplemented with 20% glycerol. This crystal form contains one protomer in the asymmetric unit and crystallizes in space group $P2_12_12_1$. A data set was collected from a rod-shaped crystal with dimensions of 50 × 50 × 80 µm on the ESRF microfocus beamline ID23-2 (Flot *et al.*, 2010) at a wavelength of 0.8726 Å (14.209 keV) and a Gaussian beam size of approximately 8 × 11 µm FWHM. The data collection was performed with an MD2M mini-diffractometer (Maatel, Voreppe, France) and a MAR 225 3 × 3 CCD detector with 1 s exposures and a 1° oscillation range. The ring mode was 16 bunch, ring current 74 mA. No attenuation was used and the measured photon flux was 2.28×10^{10} photons s^{-1} using a calibrated diode. Because the crystal was significantly larger than the beam, dose calculation was not straightforward and indeed the definition of 'dose' in this context is somewhat different from the case in which the beam size is well matched to the crystal size. This is because the absorbed dose becomes more dependent on the region of interest in the crystal: some regions receive much higher doses than others. Because of this, the dose calculations for trypsin must be treated with caution. Nevertheless, a pre-release version of *RADDOSE* optimized to deal with such cases (from Elspeth Garman and Oliver Zeldin) was used to calculate the approximate dose per image (0.018 MGy per image). The trypsin crystal diffracted to 1.03 Å resolution and 900 images were collected from a single position (Table 1). The before sub-data set was comprised of images 1–110 (Table 1). The after sub-data set was comprised of images 814–900 (Table 2). Diffraction was somewhat anisotropic, with the poorer region appearing around $\varphi = 66°$ and at roughly 180° intervals subsequently and observed as a periodic increase in the R value and decrease in $\langle I/\sigma(I)\rangle$ per frame.

3.2.2. Structure determination. The parameters used for structure determination were 52 values of k, $0.90 > k > 1.02$, with a substructure-solution resolution cutoff in *SHELXD* of 1.3 Å. The isomorphous signal ($\langle d'/{\rm sig}\rangle$) was 1.95 for 50–1.03 Å at $k = 1$. Inspection of model-phased RIP $k(F_{\rm before} - F_{\rm after})$ maps revealed positive peaks with heights of 17.2–31.1σ at cysteine S^{γ} positions, negative peak heights at new S^{γ} positions of up to -11σ and much lower intensity but still strong peaks of 6–8σ at O atoms ($O^{\delta1}$ and $O^{\delta2}$ of Asp189 and $O^{\varepsilon1}$ of Gln192, for example). Good RIP substructures occurred in the range $0.9048 > k > 1.0152$ and the wMPEs converged to 15.7–23.2° after a single cycle of density modification. Without any density modification, the best overall wMPEs were already as low as 75.4° after substructure determination (Fig. 2*b*). The majority of the structure could be built automatically by *SHELXE*. The pseudo-free correlation coefficient followed the same variation with k and ranged from 76.0 to 81.1% in the range $0.9048 > k > 1.0152$. Substructure correctness also followed this trend and there were no boundary regions in which the substructure could be determined but which resulted in uninterpretable electron-density maps.

3.2.3. Completeness analysis. We were interested in determining whether the extremely low requirements for completeness of thaumatin would be borne out by a second test case and we therefore applied the same analysis to trypsin. We again observed a steady increase in the RIP difference-map peak heights up to a maximum of 31σ (Fig. 4*b*). We also observed that the percentage of k runs that yielded interpretable phases increased, as did the mean solution completeness when the after data set was moved further downstream. Where trypsin differed strikingly from thaumatin was in the required completeness level. When the plot of the best wMPE as a function of completeness was examined, we found that substructures were unsuccessful unless the before and after data sets were >80% complete (compared with 60% for thaumatin).


(*a*)

(*b*)

Figure 5
(*a*) Experimental electron-density map contoured at 1σ of conkunitzin-S1 with 50% complete before and after data sets after chain tracing in *SHELXE* (shown in orange sticks). (*b*) Representative electron density of C3a after partial automatic building in *SHELXE*, contoured at 1σ. Side-chain density is clearly visible. This figure was generated in *PyMOL* v.1.5.0.1 (Schrödinger LLC).

3.2.4. Sub-data-set boundary analysis. As with thaumatin, we wished to determine what effect the gap between the before and after data sets might have upon substructure solution and phase determination. The after data set was shifted in 50-image increments (Fig. 3*b*) towards the before data set, which was 90% complete and composed of images 1–111. On examining the results of this analysis, we found that, as with thaumatin, the percentage of k trials that yielded interpretable phases, the mean substructure completeness, the RIP difference-map peak height and the RIP difference $\langle I/\sigma(I)\rangle$ all increased as the gap between before and after data sets widened. However, there were two important distinctions between these two test cases. The first was that the RIP difference-map peak height increased rapidly for the first four runs (89–150, 96–200, 176–250 and 181–300) and then levelled off. This is in contrast to thaumatin, for which the peak heights steadily increased. The second key difference was that no overlap between the before and after data sets was tolerated: neither of the two overlapping after data sets (which overlapped by 22 and 15 images) produced interpretable maps. Indeed, interpretable maps did not consistently appear until there were 175 images (~3.15 MGy) between the before and after data sets, with a total absorbed dose of before, ‘gap’ and after of 7.2 MGy (for 400 images). After this point, phase errors in the 17.9–19° range were obtained.

4. Test cases

Having demonstrated the efficacy of this method for well studied model systems, we applied the same analysis to two well diffracting ‘real-world’ test cases.

4.1. Conkunitzin-S1

4.1.1. Experimental details. A 100 × 100 × 100 µm crystal of conkunitzin-S1, a cone-snail-derived 6.7 kDa peptide inhibitor of voltage-gated potassium channels, was subjected to the segmented RIP method. Conkunitzin-S1 crystallizes in 0.2 *M* ammonium sulfate, 0.1 *M* sodium acetate trihydrate pH 4.6, 25% polyethylene glycol 4000. The crystallization condition does not require additional components for cryoprotection. Conkunitzin-S1 crystallizes in space group $P6_3$, with one protomer in the asymmetric unit and 47% solvent content. Crystals diffracted to 1.05 Å resolution on ESRF beamline ID23-1. 180 images were collected, with 1° oscillation, 0.2 s exposure and 5% transmission in 7/8 mode with 188 mA ring current, resulting in a flux of 2×10^{11} photons s^{-1} and a total absorbed dose of 4.9 MGy. The beam has a Gaussian profile of size 45 × 30 µm at FWHM.

4.1.2. Structure solution of conkunitzin-S1. The parameters used for structure determination were 0.83000 > k > 1.0085, 22 values of k and a substructure-solution resolution cutoff of 1.2 Å. The overall isomorphous signal ($\langle d'/\mathrm{sig}\rangle$) was 2.79 for 50–1.05 Å at k = 1. Inspection of model-phased RIP $k(F_{\rm before} - F_{\rm after})$ maps revealed positive peaks with heights of 18.8–24.2σ at cysteine S^{γ} positions, very weak negative peak heights at new S^{γ} positions of up to -7σ and, intriguingly, a 14.6σ peak at the P atom of a bound phosphate ion. The electron-density maps were of excellent quality (Fig. 5*a*). Good RIP substructures were obtained in the range 0.83000 > k > 0.94900 (Fig. 2*c*). wMPEs generally converged to 26.2–34.1° after a single cycle of density modification. At one lower value of k (0.86400) a second cycle greatly improved the wMPE (from 68.5 to 32.0°). Without any density modification, the best overall wMPEs ranged from 71.6 to 73.0° after substructure determination for values of k that yielded good substructures (0.83000 > k > 0.92350). The pseudo-free correlation coefficient followed the same variation with k and ranged from 76.3 to 79.2% in the range 0.88100 > k > 0.95750. At very low values of k accurate substructures could be determined but interpretable electron-density maps could not be obtained. Taken together, these metrics demonstrate that phasing was straightforward for conkunitzin-S1 using the segmented RIP method.

4.1.3. Completeness analysis of conkunitzin-S1. Minimum completeness analysis revealed a profile that was more similar to that of thaumatin than that of trypsin: quite low completeness was tolerated (Fig. 4*c*). At very low completeness (50%), although the map quality was quite poor (wMPE of 56.0°) and at the limit of interpretability, the addition of several rounds of automatic building in *SHELXE* produced excellent maps with a wMPE of 26° overall, correlation coefficients of the partial model of 31.3% and pseudo-free correlation coefficients of 80.3%. Thus, with some modification to include automatic chain tracing, the structure could be determined even with very low before and after data-set completeness. Additionally, as in both previous cases, increasing the completeness of the sub-data sets had a profound effect on the average solution completeness, the RIP difference-map peak height and the percentage of k runs that produced phase sets of <50° overall.

4.1.4. Conkunitzin-S1 sub-data-set boundary analysis. The after data set was shifted in 10° increments to study the effects of changing the position of the after data set (Fig. 3*c*). The before data set was composed of images 1–46. One can see a similar pattern to both thaumatin and trypsin in that there is an accumulation of overall $\langle I/\sigma(I)\rangle$ RIP difference strength, percentage of successful k runs and mean solution completeness as a larger gap is opened between the before and after data sets. Similarly, very high quality maps can be obtained after only a small number of images and the wMPEs of these maps do not substantially improve as the gap is widened. Some overlap between the before and after data sets is tolerated, with runs 2 (images 31–80) and 3 (images 45–90) producing phase sets with wMPEs of 30.9 and 23.9°, respectively. Run 2 had an overlap with the before data set of 15 images (total dose of 2.16 MGy, overlap dose of 0.41 MGy) and run 3 had an overlap of one image (total dose of 2.43 MGy, gap dose of 0.027 MGy). What is striking here and in the other cases is not necessarily the number of tolerated overlapping images, but the fact that such a small gap between the before and after data sets is necessary for successful phasing. This is consistent with the completeness analysis, which showed that quite incomplete data sets could still result in interpretable maps.

4.2. C3a

4.2.1. Experimental details. A 260 × 20 × 100 µm crystal of C3a, a 9.1 kDa component of the complement pathway, was used as a second test case for the segmented RIP method. C3a crystallizes in space group $P6_322$ with a solvent content of 66% and one protomer in the asymmetric unit (Bajic *et al.*, manuscript in preparation). In the course of refinement, Bajic and coworkers observed that a lower $R_{\rm free}$ was obtained when the data were treated as space group $P6_3$ with two protomers in the asymmetric unit. This is likely to be a consequence of small conformational differences at the carboxy-terminus. For RIP structure determination, we instead treated the data in the higher symmetry space group because the structure was much more readily solved in $P6_322$ compared with $P6_3$. Crystals diffracted to 2.2 Å resolution on ESRF beamline ID23-1 using 10% beam transmission and a photon flux of 1.52 × 10^{11} photons s^{-1}. The ring mode was 7/8 multibunch with a ring current of 191 mA and the beam size was 45 × 30 µm FWHM. The total absorbed dose was estimated to be 6.3 MGy, although no effort was made to model the fact that the crystal size exceeded the beam size. Data were collected using a wavelength of 1.0723 Å to a maximum resolution of 2.28 Å. A model that had been partially refined against the before data set was used as a reference for the computation of wMPEs. It is worth mentioning that the crystal was oriented with the *b* axis nearly parallel to the spindle axis, which resulted in very sparse improvements in the completeness from images 80 to 180.

4.2.2. C3a structure solution. The parameters used for structure determination were 5000 *SHELXD* cycles for substructure determination, 1000 cycles per *SHELXE* run, 0.92 > k > 1.00 for 41 values of k and a substructure-solution resolution cutoff of 2.7 Å. The isomorphous signal was 1.87 for 50–2.28 Å at k = 1. Inspection of model-phased RIP $k(F_{\rm before} - F_{\rm after})$ maps revealed positive peaks with heights of 14.2σ at cysteine S^{γ} positions and negative peak heights at new S^{γ} positions of up to -6σ. In this case, interpretable maps could not be obtained even with seven or more rounds of *SHELXE* recycling. Instead, *SHELXE* with auto-tracing (200 cycles of density modification and five rounds of auto-tracing) was used. After only two non-autobuilding and one autobuilding round of *SHELXE* excellent maps could be obtained (Fig. 5*b*). The best correlation coefficient for a partial structure against native data was 39.41%. In contrast, the wMPE values were relatively high, most likely owing to the fact that the model was not fully refined (Fig. 2*d*). Good RIP substructures could be found in a broad peak from 0.92000 > k > 0.97200, but the peaks for other metrics, most notably wMPE, did not have the clean inverted top-hat relationship seen in the other systems (Figs. 2*a*, 2*b* and 2*c*). Indeed, there was significant noisiness in the various metrics rather than an uninterrupted region of good statistics. Nevertheless, there was a central region which consistently yielded interpretable maps within the range 0.93400 > k > 0.98200. Pseudo-free correlation coefficients were also quite poor in the last *SHELXE* run before auto-tracing, ranging from 49.2% to a maximum of 67.6%. It is interesting to note that at very low values of k, substructures could be determined correctly but interpretable phases could not be produced.

4.2.3. C3a completeness analysis. Completeness analysis revealed that this test case did not permit very much truncation of data (Fig. 4*d*). A phase set with wMPE < 50° was produced but was barely interpretable. Thus, the width of k values that produced interpretable maps was zero in most cases. In order to produce easily interpretable maps, cycles of auto-tracing were required and even then the most incomplete data that produced interpretable maps was 80%, which produced maps with a 24.9% correlation coefficient to the partial model. Otherwise, as with the other systems, average solution completeness, RIP difference-map peak height and the percentage of k runs that produced interpretable maps all benefitted from higher completeness.

4.2.4. C3a sub-data-set boundary analysis. As mentioned earlier, the data set suffered from poor orientation of the crystallographic b axis relative to the axis of rotation. Because of the marginal isomorphous signal, extremely complete before and after data sets were required for successful phasing. In spite of this, we performed an analysis of the effect of the after data-set position on phasing. The target completeness was 95% (rather than 90% as used in other systems) and the window was moved in ten-image (0.35 MGy dose) increments. The before data set consisted of images 1–29. The overall RIP difference $\langle I/\sigma(I)\rangle$, the RIP difference-map peak height, the mean substructure completeness and the percentage of successful k runs all improved with a larger gap between the before and after data sets (Fig. 3*d*). This is consistent with thaumatin, trypsin and conkunitzin-S1. However, C3a differed in that a much larger gap was required before any successful phasing occurred. The first run with an interpretable map was run 4: images 67–110, with a 38-image gap between the before and after data sets. This corresponds to a dose of 1.33 MGy for the gap and to 3.85 MGy absorbed dose for both data sets and gap.

5. Conclusions and future perspectives

We have presented a method for the segmentation of one large data set into multiple sub-data sets which can then be used for RIP phasing. As with standard 'induced-burn' RIP, this method is applicable to samples with heavy atoms and/or disulfide bonds (Evans *et al.*, 2003; Nanao *et al.*, 2005; Ramagopal *et al.*, 2005; Fütterer *et al.*, 2008), although the presence of relatively strong (14σ and higher) peaks around other groups such as phosphates suggests that it may also be possible to determine phases from damage at other atom types. We have applied this method to four crystals for which a large but otherwise normal data collection was performed with no attempt to optimize the data collection for RIP. In order to determine the practical limits of segmented RIP, we analyzed the effects of the spacing between the before and after data sets, as well as the effects of truncating both data sets. We found in all cases that the proportion of successful substructure determinations and the calculation of interpretable maps improved dramatically when the inter-data-set gap and/or the completeness were increased. Additionally, while major improvements in the quality of the phases could not be achieved by the relative positioning of the before and after data sets, phase errors could be significantly improved in all cases by increasing the completeness of the sub-data sets. It is also interesting to note that in some cases very low completeness (down to 50%) could still be successfully used both to determine the RIP substructure and to calculate interpretable maps. This is of particular interest in cases where it is not possible to collect data sets with a sufficiently high redundancy to produce two highly complete sub-data sets. Since we have focused on the practical aspects of this method in this study, we have not yet explored the effects of high redundancy or matching oscillation ranges for the before and after data sets. In studying the effect of the inter-data-set gap, we found that in some cases not only is no gap required between the before and after data sets, but the data sets can also even overlap.

Because radiation damage occurs throughout data collection, it is not obvious which dose should be taken as the most relevant for maximizing the success rate of segmented RIP: the dose of the gap between the data sets, the total dose of the before data set and the gap, the dose of the before data set, the gap and the after data set, or some other possibility. In these systems, the gap doses ranged from 'negative', in the sense that there was overlap between the data sets (of 0.41 MGy), to a positive dose of 3.15 MGy. Nevertheless, the minimum total dose of the before data set, gap and after data set that was required for successful phasing ranged from 2.08 to 7.20 MGy. However, it is likely that this dose is highly dependent on the contents of the crystal.

We believe that segmented RIP is particularly well adapted to synchrotron beamlines equipped with the latest generation of fast-readout detectors [for example, PILATUS (Dectris) or MX-HS (Rayonix) detectors], which can collect very large high-redundancy data sets in minutes. At such a beamline, users could simply collect data until the Garman limit (Owen *et al.*, 2006) is attained and subsequently utilize the segmented RIP method to determine whether an RIP signal exists and, if so, to optimize and exploit it for phasing. This approach is complementary to the now common approach of collecting highly redundant SAD data sets, and the phases from each analysis can be combined. In addition to the utility of this method for *de novo* phasing, if external sources of phases exist (for example from molecular replacement), they could be used to compute $F_{\rm before} - F_{\rm after}$ difference maps to identify sites of radiation damage or to validate weak molecular-replacement solutions.

Several improvements to this method are envisioned. At the substructure-determination level, while *SHELXC* and *SHELXD* represent the state of the art for RIP substructure determination, it is likely that improved software could greatly boost the success rate. We have shown that sufficient RIP signal can be found in large data sets that have not been optimized for RIP using a simple segmenting approach. It is therefore likely that structure-solution software which explicitly considers specific radiation damage could greatly improve the success rate of RIP substructure determination and possibly obviate the need for sub-data-set boundary optimization. For example, a more sophisticated approach to substructure determination than segmentation could be to model specific radiation damage on a per-frame or a per-dose basis (for example, in a similar fashion to the modelling of radiation damage by *SHARP*). In addition to improving phasing for segmented RIP experiments, it would also greatly benefit structure solution using radiation-sensitive heavy atoms such as selenium. Similarly, for phase calculation and heavy-atom-site refinement our analysis relied on *SHELXE*, but it is likely that in more marginal cases the use of *SHARP* (Schiltz *et al.*, 2004; Schiltz & Bricogne, 2007) could further reduce the signal requirements once a substructure has been correctly determined. Even in the absence of these developments, however, segmented RIP represents a potentially valuable tool in the experimental determination of phases in MX and requires very little if any modification to typical data-collection strategies.

The authors would like to thank Dr Elspeth Garman and Oliver Zeldin for assistance with *RADDOSE* to determine dose in cases where the beam size exceeds the crystal size, Dr Prashantha Karunakar for the cocrystallization of thaumatin, Dr Andrew McCarthy for critical reading of the manuscript, Dr Felix Frolow for allowing us to discuss the structure determination of conkunitzin-S1, and Dr Gregers Andersen, Dr Laure Yatime and Goran Bajic for allowing us to discuss the structure determination of C3a. We thank Ulrike Kapp and Dr Chloe Zubieta for trypsin crystals.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.

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

Beck, T., Gruene, T. & Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 374–380.

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

Flot, D., Mairs, T., Giraud, T., Guijarro, M., Lesourd, M., Rey, V., van Brussel, D., Morawe, C., Borel, C., Hignette, O., Chavanne, J., Nurizzo, D., McSweeney, S. & Mitchell, E. (2010). *J. Synchrotron Rad.* **17**, 107–118.

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

Kabsch, W. (2010*a*). *Acta Cryst.* D**66**, 133–144.

Kabsch, W. (2010*b*). *Acta Cryst.* D**66**, 125–132.

Leal, R. M. F., Bourenkov, G. P., Svensson, O., Spruce, D., Guijarro, M. & Popov, A. N. (2011). *J. Synchrotron Rad.* **18**, 381–386.

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. G. (2006). *Structure*, **14**, 791–800.

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

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.

Panjikar, S., Mayerhofer, H., Tucker, P. A., Mueller-Dieckmann, J. & de Sanctis, D. (2011). *Acta Cryst.* D**67**, 32–44.

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

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

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

Rudiño-Piñera, E., Ravelli, R. B. G., Sheldrick, G. M., Nanao, M. H., Korostelev, V. V., Werner, J. M., Schwarz-Linek, U., Potts, J. R. & Garman, E. F. (2007). *J. Mol. Biol.* **368**, 833–844.

Sanctis, D. de *et al.* (2012). *J. Synchrotron Rad.* **19**, 455–461.

Sanctis, D. de, Tucker, P. A. & Panjikar, S. (2011). *J. Synchrotron Rad.* **18**, 374–380.

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.

Schönfeld, D. L., Ravelli, R. B., Mueller, U. & Skerra, A. (2008). *J. Mol. Biol.* **384**, 393–405.

Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.

Thorn, A. & Sheldrick, G. M. (2011). *J. Appl. Cryst.* **44**, 1285–1287.

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

Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.