Combining cross-crystal averaging and MRSAD to phase a 4354-amino-acid structure

Jason Nicholas Busby, J. Shaun Lott and Santosh Panjikar


IUCr Journals
CRYSTALLOGRAPHY JOURNALS ONLINE

Copyright © International Union of Crystallography

Author(s) of this paper may load this reprint on their own web site or institutional repository provided that this cover page is retained. Republication of this article or its storage in electronic databases other than as specified above is not permitted without prior permission in writing from the IUCr.

For further information see http://journals.iucr.org/services/authorrights.html
Combining cross-crystal averaging and MRSAD to phase a 4354-amino-acid structure

Jason Nicholas Busby,a,* J. Shaun Lottaa and Santosh Panjikarb,c,*

aSchool of Biological Sciences, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand, bMX, Australian Synchrotron, 800 Blackburn Road, Clayton, Melbourne, VIC 3168, Australia, and cDepartment of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia. *Correspondence e-mail: j.busby@auckland.ac.nz, santosh.panjikar@synchrotron.org.au

The B and C proteins from the ABC toxin complex of *Yersinia entomophaga* form a large heterodimer that cleaves and encapsulates the C-terminal toxin domain of the C protein. Determining the structure of the complex formed by B and the N-terminal region of C was challenging owing to its large size, the non-isomorphism of different crystals and their sensitivity to radiation damage. A native data set was collected to 2.5 Å resolution and a non-isomorphous Ta₆Br₁₂-derivative data set was collected that showed strong anomalous signal at low resolution. The tantalum-cluster sites could be found, but the anomalous signal did not extend to a high enough resolution to allow model building. Selenomethionine (SeMet)-derivatized protein crystals were produced, but the high number (60) of SeMet sites and the sensitivity of the crystals to radiation damage made phasing using the SAD or MAD methods difficult. Multiple SeMet data sets were combined to provide 30-fold multiplicity, and the low-resolution phase information from the Ta₆Br₁₂ data set was transferred to this combined data set by cross-crystal averaging. This allowed the Se atoms to be located in an anomalous difference Fourier map; they were then used in *Auto-Rickshaw* for multiple rounds of autobuilding and MRSAD.

1. Introduction

The phase problem remains one of the key difficulties in X-ray crystallography. The number of structures deposited in the Protein Data Bank (PDB) increases every year (Berman et al., 2000) and this provides an increasing pool of structures for use in molecular replacement, but experimental phasing is still required when investigating proteins for which no good molecular replacement models exist. The crystallographer’s next tool of choice tends to be SAD phasing using selenomethionine (SeMet)-substituted protein (Hendrickson et al., 1990). This technique is favoured for several reasons: SeMet can be incorporated into expressed proteins easily by supplementation during growth (Cowie & Cohen, 1957), SeMet incorporation can occur at high rates and the hydrophobic methionine residues are often buried in the hydrophobic core of proteins and are highly ordered. Another alternative is isomorphous replacement; however, this technique often suffers from non-isomorphism between different derivatives and the native crystal. Even relatively small changes in the unit cell between heavy-atom-soaked and native crystals can obliterate the isomorphous signal (Garman & Murray, 2003).

SAD or MAD phasing using SeMet-substituted protein can sometimes be very difficult. Complications can include crystals that fail to diffract to high resolution, limited anomalous signal, radiation damage and a very small or large number of methionine residues limiting the amount of phase information.
or making it difficult to determine the heavy-atom substructure. Some of these problems can be remedied by collecting highly redundant data from a single crystal, but with additional X-ray exposure comes additional radiation damage. Anomalous scatterers are often highly sensitive to radiation damage owing to their increased X-ray absorption cross-section, particularly at absorption peaks (Murray et al., 2005), so collecting additional data from a single crystal may not help. The advent of microfocus beamlines has enabled the collection of multiple data sets from different regions of a single crystal (Li et al., 2004; Moukhametzianov et al., 2008; Sanishvili et al., 2008), helping to combat radiation damage. Another method is to combine data from multiple different crystals (Liu et al., 2011, 2013, 2014). This can improve the signal-to-noise ratio, reduce crystal-specific systematic error and allow high completeness while maintaining a low dose to limit radiation damage.

There are a number of different phasing strategies available, and an even greater number of programs for pursuing them. Many of these programs work in different ways, and in difficult cases one program may succeed where others have failed, necessitating a large amount of trial and error. In some difficult cases a combination of phasing techniques can work where individual techniques have failed. Auto-Rickshaw is an automated structure-determination platform that is able to try a wide variety of phasing strategies separately and in combination (Panjikar et al., 2005, 2009). The techniques available include S-SAD, SAD, two-, three- and four-wavelength MAD, SIRAS, MR, MRSAD, MRSIRAS, RIP and MRRIP. The goal of Auto-Rickshaw (Panjikar et al., 2005) is to emulate an experienced crystallographer making decisions about which approaches to try and which programs to use on the fly during data evaluation in an automatic manner.

We have described the structure and function of the B–C component of the ABC toxin complex from Yersinia entomophaga previously (Busby et al., 2013). Here, we detail the methodologies undertaken to solve this difficult structure and their applicability to general X-ray crystallographic problems.

2. Methods and results

2.1. Cloning, expression and purification of YenB–YenC2<sub>NTR</sub>

The cloning, expression and purification of the YenB–YenC2<sub>NTR</sub> protein complex is described in Busby et al. (2013). The YenB and YenC2 genes from Y. entomophaga (GenBank accession No. DQ400808.1) were cloned into the pETDuet-1 co-expression vector (EMD Biosciences). YenB was cloned into multiple cloning site 1 (MCS1), including an N-terminal 6×His tag followed by a Tobacco etch virus (TEV) protease cleavage site. YenC2 was cloned into MCS2 with no tags.

Expression was performed in Escherichia coli Rosetta 2 (DE3) cells using ZYM-5052 auto-induction medium (Studier, 2005) at 291 K. The cells were lysed in 20 mM HEPES pH 7.5, 150 mM NaCl, 1 mM β-mercaptoethanol by a continuous-flow cell disruptor (Microfluidics M-110p) and insoluble material was removed by centrifugation. The protein was purified by immobilized metal-affinity chromatography (IMAC) followed by removal of the His tag by TEV cleavage, subtractive IMAC and size-exclusion chromatography (SEC).

When co-expressed with YenB, YenC2 was cleaved into N-terminal and C-terminal regions (YenC2<sub>NTR</sub> and YenC2<sub>CTR</sub>) and all three peptides remained tightly associated throughout purification. This complex was used for crystallization trials, but no crystals could be obtained. The YenB–YenC2<sub>NTR</sub> complex was dialysed against acetate buffer overnight (20 mM sodium acetate pH 4.5, 150 mM NaCl, 1 mM β-mercaptoethanol), causing YenC2<sub>CTR</sub> to dissociate from the complex and precipitate. This precipitate was removed by filtration and the remaining supernatant was subjected to SEC in acetate buffer before dialysis against buffer to restore the pH to 7.5. This resulted in a YenB–YenC2<sub>NTR</sub> complex that was subsequently used for crystallization trials.

2.2. Crystallization

The YenB–YenC2<sub>NTR</sub> protein complex was concentrated to 7.3 mg ml<sup>−1</sup> and crystallization trials were carried out with nanolitre-dispensing robotics using the conditions described in Moreland et al. (2005) and Gorrec (2009). Several conditions were found to produce crystals.

Crystals were fine-screened on a larger scale by hanging-drop vapour diffusion using 24-well plates. The best crystals were produced with well solution consisting of 18% (w/v) PEG 3350, 0.15 M K<sub>2</sub>HPO<sub>4</sub> pH 4.8. Microseeding (Luft & DeTitta, 1999) was used to further optimize the crystallization, significantly improving the size, quality and reproducibility of the crystals.

Figure 1
SeMet (a) and tantalum-cluster-derivatized (b) crystals of YenB–YenC2<sub>NTR</sub>. Scale bars are 100 μm in length.

2.3. Heavy-atom derivatization

No structures of proteins homologous to either YenB or YenC2 were available, so experimental phasing was pursued. SeMet-labelled protein was produced by expression in LB supplemented with selenomethionine. The protein was purified using the same methods as used for the native protein and was crystallized in the same condition using microseeding with native crystals. The crystals were cryoprotected by soaking them briefly in mother liquor with the addition of 5–20% glycerol as a cryoprotectant.

Hexatantalum dodecabromide (Ta\textsubscript{6}Br\textsubscript{12}) is a cluster compound that has been used to successfully phase the structure of the Australian Synchrotron. Typically, images were collected at a temperature of 100 K with 0.5 or 1.0 s exposure time, 0.5° or 1.0° oscillation and a total of 360° or 720° collected (Table 1). Data were integrated using XDS (Kabsch, 2010), with careful attention paid to the change in anomalous correlation as additional frames were added. If the anomalous correlation began to drop this was taken as a sign of radiation damage and data were truncated before this point (e.g. data set SeMet 1). Integrated data were scaled and merged using AIMLESS from the CCP4 suite (Evans & Murshudov, 2013; Winn et al., 2011).

Diffraction data from a native crystal were collected at 0.9537 Å to a resolution of ~2.26 Å, but ultimately a high-resolution cutoff of 2.5 Å was used. This crystal showed significant differences in unit-cell dimensions from other SeMet-labelled and tantalum-cluster-soaked crystals (Table 1) and so could not be used for isomorphous replacement directly.

Diffraction data from SeMet-labelled crystals were collected above the selenium peak (0.9791 Å) and at the inflection point (0.9795 Å). The anomalous signal from the best peak data set (SeMet1) only extended to 5.2 Å resolution, based on the resolution at which CC\textsubscript{anom} drops below 0.3. This data set had to be truncated owing to radiation damage, resulting in low completeness. When peak data sets from multiple isomorphous crystals were combined, however, the anomalous correlation was found to improve, so additional data sets were added in an iterative fashion until a point at which adding additional data did not improve CC\textsubscript{anom}. This resulted in a data set (Combined SeMet) with high multiplicity (~30) and anomalous signal that extended to ~5 Å resolution (Fig. 2).

<table>
<thead>
<tr>
<th>Space group</th>
<th>SeMet1</th>
<th>SeMet2</th>
<th>SeMet3</th>
<th>Combined SeMet</th>
<th>Ta1 HREM</th>
<th>Ta1 LREM</th>
<th>Ta2 HREM</th>
</tr>
</thead>
<tbody>
<tr>
<td>P\textsubscript{2}\textsubscript{1}2\textsubscript{1}2\textsubscript{1}</td>
<td>P\textsubscript{2}\textsubscript{1}2\textsubscript{1}2\textsubscript{1}</td>
<td>P\textsubscript{2}\textsubscript{1}2\textsubscript{1}2\textsubscript{1}</td>
<td>P\textsubscript{2}\textsubscript{1}2\textsubscript{1}2\textsubscript{1}</td>
<td>P\textsubscript{2}\textsubscript{1}2\textsubscript{1}2\textsubscript{1}</td>
<td>Ta1 HREM</td>
<td>Ta1 LREM</td>
<td>Ta2 HREM</td>
</tr>
</tbody>
</table>

| No. of unique reflections | 189788 (9300) | 131321 (4767) | 122012 (5408) | 123352 (5835) | 141318 (5939) | 89256 (4841) | 88515 (4823) | 129593 (6081) |

† Data were limited owing to radiation damage becoming apparent. Data were truncated at the point where adding additional frames no longer increased the anomalous signal. ‡ R\textsubscript{int} (R\textsubscript{int}1) and R\textsubscript{int}2 are as defined in Weiss (2001). § R.m.s. correlation ratio between two half data sets (Evans, 2011). ¶ Anomalous resolution is defined as the resolution at which CC\textsubscript{anom} drops below 0.3.
Diffraction data from tantalum-cluster-derivatized crystals was collected above and below the tantalum edge (1.2516 and 1.2580 Å, respectively). The high-energy remote data set showed a very high anomalous signal at low resolution (Fig. 2), but this signal dropped off very sharply at moderate resolution (~5.7 Å), a phenomenon that has been identified previously in tantalum-cluster derivatization (Banumathi et al., 2003).

2.4.1. Initial phasing attempts. Phasing was initially attempted using SeMet SAD/MAD phasing. The protein complex under study contains 30 non-initiation methionine residues and the unit-cell parameters suggested two heterodimers in the asymmetric unit (Matthews coefficient 2.78 Å³ Da⁻¹, solvent content 56%; Matthews, 1968), giving a total of 60 heavy-atom sites. Initial attempts at substructure determination were carried out with the individual SeMet data sets and the combined data-set sites using SHELXC v.2006/4.3 (Sheldrick, 2010), phenix.hyss v.1.8_1069 (Grosse-Kunstleve & Adams, 2003; McCoy et al., 2004) and SnB v.2.3 (Rappeleye et al., 2002), but these attempts were unsuccessful.

The most likely explanation for this is the limited resolution of the phase information (Fig. 2), the large number of sites (60) and possible radiation damage.

Tantalum bromide cluster compounds have been used to successfully phase large structures, where their extremely large anomalous and isomorphous signal and propensity to bind to relatively few sites are advantages (Knäblein et al., 1997; Banumathi et al., 2003; Gomis-Ruth & Coll, 2001). Several data sets were collected from crystals derivatized with tantalum clusters and they showed very strong anomalous signal at low resolution (Fig. 2 and Table 1), but this dropped off sharply at around 6 Å. We searched for heavy-atom sites for SAD/MAD phasing using SnB and for SIRAS using SHELX. The best tantalum-cluster data set (Ta1 in Table 1) had very different unit-cell parameters compared with the native, and could be used for SAD or two-wavelength MAD but not SIRAS. To identify a suitably isomorphous crystal for SIRAS, we compared all tantalum-cluster and SeMet data sets to find the two with the most similar unit-cell parameters. Ultimately, we used a second tantalum-cluster data set (Ta2 in Table 1) as a derivative and the SeMet4 data set to act as the native data set. Seven heavy-atom sites could be easily found using either method, and the same sites were found using both methods, giving us confidence that they were correct.

As there are most likely to be two heterodimers in the asymmetric unit, twofold noncrystallographic symmetry (NCS) may be present. To test for this, we inspected a self-rotation function. The self-rotation function contained several peaks in addition to the origin, one of which had a much higher Rf/σ than the others, suggesting that NCS is present. We also tried to detect NCS from the positions of the tantalum clusters identified earlier. The CCP4 program PROFESSION (Winn et al., 2011) found a set of operators that matched the heavy-atom positions, and the rotation angles were the same as those determined from the self-rotation function (self-rotation function, 84.69, 90.00, 180.00°; PROFESSION, 84.78, 90.40, 179.33°), giving confidence in both the NCS determination and the positions of the tantalum clusters.

SAD, MAD and SIRAS phasing were attempted using the tantalum-cluster data sets. Density modification produced solvent masks that, in retrospect, showed the correct molecular shape (Fig. 3), but the maps were not sufficiently well resolved to enable model building a priori. For example, the map correlation coefficient between a density-modified

Figure 2
Anomalous correlation versus resolution for data sets. Data were collected at the peak Se wavelength (0.9791 Å) or above the tantalum edge (1.2516 Å).

Figure 3
Horizontal cross-section of the final refined structure (above) compared with a cross-section of the electron density from SIRAS phasing of the Ta2 data set after density modification (below). The overall shape is clearly similar, indicating that phasing has been successful, but the electron density was not of sufficient quality to allow model building.
In order to provide high-resolution phase information, we attempted to locate SeMet sites by using the anomalous difference Fourier method. The best tantalum-cluster data set (Ta1) has significant differences in unit-cell dimensions compared with the best SeMet data sets (Tables 1 and 2), so in order to use the low-resolution phase information from the tantalum cluster with the SeMet data cross-crystal averaging was performed. The position of tantalum-cluster sites and a self-rotation function (MOLREP; Vagin & Teplyakov, 2010) could be used to find the NCS present in the tantalum-cluster two-wavelength MAD maps. These maps were density modified with NCS averaging (DM; Cowtan, 1994) and the electron density was cut out and placed into the SeMet data set by molecular replacement (Phaser; McCoy et al., 2007). The low-resolution phase information thus provided was combined

NCS-averaged SIRAS map and the final refined model map is just 0.19 and the phase error is 85.9° (calculated using the CCP4 program CPHASEMATCH; Winn et al., 2011) to 2.8 Å resolution. The most likely explanation for this is the poor starting phase information. The Ta6Br12 cluster compound has a bipyramidal arrangement of Ta atoms at its centre and at low resolution (less than 6 Å) these atoms essentially scatter in phase, allowing them to be treated as a single ‘super atom’. At higher resolution the Bijvoet differences drop sharply, reaching a local minimum at ~4.5 Å before increasing again at higher resolution (Banumathi et al., 2003). If the anomalous diffraction data are of high enough quality, the tantalum cluster can be fully modelled and correctly oriented within the cluster density, allowing phasing at high resolution. Sadly this was not the case for our data set, as the anomalous signal only extended to ~5.7 Å resolution. The tantalum clusters could not be accurately oriented within the density, and the phase information was therefore limited to low resolution.

In order to provide high-resolution phase information, we attempted to locate SeMet sites by using the anomalous difference Fourier method. The best tantalum-cluster data set (Ta1) has significant differences in unit-cell dimensions compared with the best SeMet data sets (Tables 1 and 2), so in order to use the low-resolution phase information from the tantalum cluster with the SeMet data cross-crystal averaging was performed. The position of tantalum-cluster sites and a self-rotation function (MOLREP; Vagin & Teplyakov, 2010) could be used to find the NCS present in the tantalum-cluster two-wavelength MAD maps. These maps were density modified with NCS averaging (DM; Cowtan, 1994) and the electron density was cut out and placed into the SeMet data set by molecular replacement (Phaser; McCoy et al., 2007). The low-resolution phase information thus provided was combined

NCS-averaged SIRAS map and the final refined model map is just 0.19 and the phase error is 85.9° (calculated using the CCP4 program CPHASEMATCH; Winn et al., 2011) to 2.8 Å resolution. The most likely explanation for this is the poor starting phase information. The Ta6Br12 cluster compound has a bipyramidal arrangement of Ta atoms at its centre and at low resolution (less than 6 Å) these atoms essentially scatter in phase, allowing them to be treated as a single ‘super atom’. At higher resolution the Bijvoet differences drop sharply, reaching a local minimum at ~4.5 Å before increasing again at higher resolution (Banumathi et al., 2003). If the anomalous diffraction data are of high enough quality, the tantalum cluster can be fully modelled and correctly oriented within the cluster density, allowing phasing at high resolution. Sadly this was not the case for our data set, as the anomalous signal only extended to ~5.7 Å resolution. The tantalum clusters could not be accurately oriented within the density, and the phase information was therefore limited to low resolution.
with the anomalous difference intensity data from the SeMet peak wavelength to calculate an anomalous difference Fourier map. 50 peaks in this map were of >5σ and were used to define starting heavy-atom positions. At this point, the data were submitted to Auto-Rickshaw for rounds of automated model building and MRSAD phasing (Panjikar et al., 2009). The phase information from the SeMet data proved to be sufficient for initial automated model building. The calculated phases from this model were combined with the experimental phase information, improving the anomalous difference Fourier and LLG map (McCoy & Read, 2010) and allowing additional heavy-atom sites to be found. This process was repeated and eventually all 60 non-initiation SeMet residues could be correctly placed and most of the structure could be built automatically using Buccaneer (Cowtan, 2006), RESOLVE (Terwilliger, 2000) and SHELXE (Sheldrick, 2010) and refined with REFMACS (Murshudov et al., 2011) and phenix.refine (Afonine et al., 2012) (Fig. 4).

2.4.2. Ex post facto analysis of phasing. While the strategy of combining data from several crystals into a high-multiplicity SeMet data set and cross-crystal averaging with tantalum-cluster phases was eventually successful in solving the YenB–YenC2NTR structure, the individual data sets seemed to be in the ‘twilight zone’ for experimental phasing. We therefore performed an ex post facto analysis on the requirements for successfully phasing this structure.

SHELX (SHELXC v.2014/2 and SHELXD v.2013/2) was used to attempt to locate heavy atoms in the individual SeMet data sets, searching for 60 sites and running 1000 trials, using a high-resolution cutoff of 4.7 Å and an E_{min} of 1.5. This high-resolution limit was determined through a trial-and-error approach using repeated SHELX runs with varying high-resolution limits. Phasing attempts were unsuccessful for any individual SeMet data set, but the heavy-atom substructure could successfully be determined for the combined SeMet data set (Fig. 5). This is somewhat surprising, as our original

![Figure 5](image-url)

Attempted substructure solution for SeMet data sets. The first four rows show individual SeMet data sets, for which the substructure could not be determined. The last row shows the combined data set containing all four individual data sets, from which the substructure was successfully determined. This can be seen by the difference between hands for contrast, connectivity and estimated CC(map), the sharp decrease in site occupancy and the fact that some trials had significantly higher CC_{all} and CC_{weak} than the majority.
Figure 6
Substructure determination for combinations of SeMet data sets. The best individual SeMet data set was SeMet4; the other SeMet data sets were added to this in various combinations and SAD phasing was attempted. Only the combination of data sets 4+1+2 and all four combined were successful when performing 1000 trials. The substructure of data sets 4+2+3 could be solved with 10 000 trials and data set 4 alone could be solved with 10 000 trials when using random-omit seeding rather than Patterson seeding.
Table 3
Data-collection and processing statistics for combinations of SeMet data sets.

The best individual SeMet data set was SeMet4. Other data sets were added to this to test whether the Se substructure could be solved.

<table>
<thead>
<tr>
<th>Space group</th>
<th>SeMet4+1</th>
<th>SeMet4+1+2</th>
<th>SeMet4+1+3</th>
<th>SeMet4+2</th>
<th>SeMet4+3</th>
<th>SeMet4+2+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit-cell parameters</td>
<td>$P_2_1_2_1_2_1$</td>
<td>$P_2_1_2_1_2_1$</td>
<td>$P_2_1_2_1_2_1$</td>
<td>$P_2_1_2_1_2_1$</td>
<td>$P_2_1_2_1_2_1$</td>
<td>$P_2_1_2_1_2_1$</td>
</tr>
<tr>
<td>$a$ (Å)</td>
<td>134.9</td>
<td>134.7</td>
<td>134.7</td>
<td>134.7</td>
<td>134.7</td>
<td>134.6</td>
</tr>
<tr>
<td>$b$ (Å)</td>
<td>150.2</td>
<td>150.3</td>
<td>150.3</td>
<td>150.4</td>
<td>150.3</td>
<td>150.4</td>
</tr>
<tr>
<td>$c$ (Å)</td>
<td>276.3</td>
<td>276.4</td>
<td>276.4</td>
<td>276.5</td>
<td>276.4</td>
<td>276.5</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.9791</td>
<td>0.9791</td>
<td>0.9791</td>
<td>0.9791</td>
<td>0.9791</td>
<td>0.9791</td>
</tr>
<tr>
<td>No. of frames</td>
<td>856</td>
<td>1216</td>
<td>1216</td>
<td>1080</td>
<td>1080</td>
<td>1440</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>96.50–2.77</td>
<td>96.46–2.77</td>
<td>96.46–2.77</td>
<td>96.47–2.77</td>
<td>96.46–2.77</td>
<td>96.46–2.77</td>
</tr>
<tr>
<td>$R_{	ext{meas}}$</td>
<td>0.335 (2.112)</td>
<td>0.371 (2.080)</td>
<td>0.356 (2.042)</td>
<td>0.381 (2.409)</td>
<td>0.365 (2.388)</td>
<td>0.388 (2.339)</td>
</tr>
<tr>
<td>$R_{	ext{i.m.}}$</td>
<td>0.112 (0.930)</td>
<td>0.104 (0.919)</td>
<td>0.100 (0.902)</td>
<td>0.113 (1.122)</td>
<td>0.108 (1.113)</td>
<td>0.100 (1.099)</td>
</tr>
<tr>
<td>$\langle f^2 / \sigma^2 \rangle$</td>
<td>14.7 (1.2)</td>
<td>15.1 (1.2)</td>
<td>14.6 (1.2)</td>
<td>14.6 (1.0)</td>
<td>14.6 (1.0)</td>
<td>14.6 (1.0)</td>
</tr>
<tr>
<td>CC1/2</td>
<td>99.4 (47.7)</td>
<td>99.3 (48.8)</td>
<td>99.4 (49.3)</td>
<td>99.2 (45.1)</td>
<td>99.4 (46.3)</td>
<td>99.5 (45.2)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.4 (89.0)</td>
<td>99.4 (87.4)</td>
<td>99.4 (87.8)</td>
<td>98.9 (82.5)</td>
<td>98.9 (82.9)</td>
<td>98.8 (80.4)</td>
</tr>
<tr>
<td>Multiplicity</td>
<td>17 (7.3)</td>
<td>23.4 (7.3)</td>
<td>23.4 (7.3)</td>
<td>20.8 (6.7)</td>
<td>20.9 (6.7)</td>
<td>27.3 (6.6)</td>
</tr>
<tr>
<td>CC\text{anom} (%)</td>
<td>9.4 (–0.7)</td>
<td>11.0 (–0.3)</td>
<td>11.0 (–0.4)</td>
<td>11.5 (–0.6)</td>
<td>11.7 (–1.7)</td>
<td>13.6 (–2.5)</td>
</tr>
<tr>
<td>RCR\text{anom} (%)</td>
<td>1.099 (0.993)</td>
<td>1.117 (0.969)</td>
<td>1.116 (0.997)</td>
<td>1.123 (0.994)</td>
<td>1.124 (0.984)</td>
<td>1.147 (0.975)</td>
</tr>
<tr>
<td>Anomalous resolution (Å)</td>
<td>5.4</td>
<td>5.1</td>
<td>5.2</td>
<td>5.1</td>
<td>5.3</td>
<td>5.0</td>
</tr>
<tr>
<td>No. of unique reflections</td>
<td>141965 (6422)</td>
<td>141967 (6315)</td>
<td>141967 (6330)</td>
<td>141324 (5936)</td>
<td>141327 (5967)</td>
<td>141327 (5813)</td>
</tr>
</tbody>
</table>

† $R_{	ext{meas}}$ ($R_{	ext{i.m.}}$) and $R_{	ext{i.m.}}$ are as defined in Weiss (2001). ‡ R.m.s. correlation ratio between two half data sets (Evans, 2011). § Anomalous resolution is defined as the resolution at which $CC_{\text{anom}}$ drops below 0.3.

Many of these phasing attempts found only a few successful runs out of 1000 trials. The fact that data set 4+1+2 succeeded while 4+2+3 failed was somewhat unexpected as the latter data set has higher multiplicity, so we repeated this phasing run using an even larger number of trials (10 000). In this case, the substructure was successfully determined (Fig. 6), emphasizing the need for a very large number of trials when dealing with weak anomalous data. Klinke et al. (2015) recently determined the structure of a large protein with a large number of anomalous sites by S-SAD. One of the conclusions that they drew is that when a large number of anomalous scatterers are present, Patterson seeding can adversely affect the phasing process. Therefore, we attempted substructure determination using each individual data set with 10 000 trials and either Patterson or random-omit seeding. Substructure determination was unsuccessful for SeMet1, SeMet2 and SeMet3 (not shown), but was successful in two of the 10 000 trials with SeMet4 when using random-omit seeding (Fig. 6). This supports the use of random seeding over Patterson when searching for large numbers of heavy atoms.

For the any of the combinations of data sets where heavy atoms could be located the maps were sufficient to allow autobuilding, and phase information from this initial model could be combined with the experimental phases using MR-SAD, allowing additional heavy-atom sites to be located and phases to be improved. This process is automated in Auto-Rickshaw, allowing the structure to be fully determined from any of these starting points.

2.4.3. Isomorphism of data sets. Despite growing in identical conditions, crystals of YenB–YenC2 NTR showed varying levels of non-isomorphism from batch to batch. This may be attributed to intrinsic differences in the crystals, changes caused by soaking in heavy-atom solutions (in the case of the tantalum-cluster crystal) or varying levels of dehydration occurring during harvesting, cryoprotection and cryocooling. The largest changes are seen between the native and tantalum-cluster data sets, with changes in individual unit-cell dimensions of 1.4, 5.3 and 1.9 Å (a change of 1.0, 3.6 and 0.7%, respectively). This results in an increase in the unit-cell volume of 5.4%.

In order to compare the imperfect isomorphism of the various data sets, we scaled them against each other in a pairwise fashion and calculated the cross-crystal $R$ factors (Table 2) using the CCP4 program SCALEIT (Winn et al., 2011). The SeMet data sets were all relatively isomorphic to...
one another, with data sets 2 and 3 being the most similar. This justifies our decision to combine all four data sets to increase the multiplicity for SAD phasing. The tantalum-cluster and native data sets, in comparison, showed relatively large $R$ factors. This is not surprising considering the large differences in unit-cell parameters (Table 1) and confirms our suspicion that non-isomorphism would make isomorphous replacement phasing difficult with these data sets.

### 2.4.4. Refinement

The autobuilt model was manually corrected and several rounds of refinement model building took place. The phase information was then transferred to the higher resolution native data set by molecular replacement. Rounds of manual model building using Coot (Emsley et al., 2010) and refinement using phenix.refine (Afonine et al., 2012) followed. The final refinement statistics and the details and the interpretation of the structure have been published in Busby et al. (2013).

### 3. Discussion

The crystal structure of the YenB–YenC2NTR complex presented a difficult phasing problem. The complex is large, with a total of 4354 amino acids in the asymmetric unit, 64 of which are methionines. SeMet-labelled crystals diffracted relatively poorly, the anomalous signal was low and the large number of heavy-atom sites initially made SAD/MAD methods unsuccessful. A tantalum-cluster data set provided a much higher anomalous signal and heavy atoms could be located, but the limited resolution and poor phase quality at high resolution meant that the electron density produced could not be interpreted. It was only with the combined use of both derivative data sets that the phase problem could be solved by using the low-resolution phase information from the tantalum-cluster data set to find heavy-atom sites in the SeMet data set and then using multiple rounds of autobuilding and MRSAD to bootstrap to a structure that could be refined.

Despite a strong anomalous signal at low resolution, the tantalum-cluster data set did not result in an interpretable electron-density map. A sharp decrease in anomalous signal at $\sim$6 Å is often seen with this cluster compound (Banumathi et al., 2003), and while heavy-atom sites could be easily located and the maps produced are (in retrospect) correct, the resolution was not sufficient to allow model building. This low-resolution phase information could, however, be used to determine the heavy-atom substructure of other derivatives. If these derivatives were isomorphous, this process would have been relatively straightforward, analogous to the process used in MIRAS (Panjikar & Tucker, 2002). A similar process can be used even with non-isomorphous derivatives by performing molecular replacement using electron density rather than an atomic model. This approach should be generally applicable in cases where an anomalous data set is available but is not quite good enough to determine the heavy-atom substructure on its own. An external source of phase information, even low-resolution or poor-quality phases, can aid in determining the substructure and allow bootstrapping into the build-and-refine cycle. Such phase information could come from a low-resolution heavy-atom data set, as is the case here, or from a poor molecular-replacement solution.

Our SeMet data sets appear to be on the border of solvability. The heavy-atom substructure could be determined in the high-multiplicity combined data set and in certain combinations of three data sets with our initial settings, but the single-crystal data set could only be solved by using very large numbers of trials with random-omit seeding rather than Patterson seeding. The benefit of highly redundant SAD data has been known for some time for single crystals (Dauter et al., 2002), and recently averaging data sets collected from a number of different crystals has been shown to be advantageous at low resolution (Liu et al., 2011, 2013, 2014; Mancusso et al., 2012; Akey et al., 2014; Wöhlert et al., 2014; Bleichert et al., 2015; Kim et al., 2015; Klinke et al., 2015). This is often necessary for successful phasing using very weak anomalous signal such as from S-SAD, but can also help in other difficult cases such as with a large number of anomalous scatterers, large asymmetric units, poor resolution or crystals that are very sensitive to radiation damage. High-multiplicity data, particularly combining data from multiple crystals, has several benefits for phase determination. Aside from the obvious benefit that more measurements means greater accuracy of measurement (which is particularly important when measuring very small anomalous differences), combining isomorphous data from multiple crystals allows the creation of a high-completeness data set with a much lower total dose, limiting the effects of radiation damage. It also helps to remove systematic errors owing to individual crystal defects, as these will tend to average out over multiple crystals.

In summary, the work presented here emphasizes several key points about the experimental phasing of difficult structures. (i) Combining data sets from multiple isomorphous crystals can allow phasing where individual crystal data sets could not. Several groups have observed this in recent years and multiple-crystal data collection is becoming more common (Cherezov et al., 2007; Liu et al., 2011, 2014; Rasmussen et al., 2011). (ii) Very high levels of multiplicity can aid in substrate determination and phasing. (iii) When solving a substructure with large numbers of heavy atoms, higher numbers of trials (>10 000) may be necessary, and random-omit seeding may work better than Patterson seeding. (iv) Low-resolution phase information that is insufficient for model building on its own (such as from a poor molecular-replacement model or limited experimental phases) can be useful in locating heavy atoms in other data sets. When data sets are isomorphous, the phases can be used directly, but even when non-isomorphous they can be used via cross-crystal averaging and, when available, NCS averaging can assist this process.

### Acknowledgements

This work was supported by a subcontract to JSL and JNB from the New Zealand Foundation for Research, Science and Technology contract C10X0804 awarded to Mark R. H. Hurst. All X-ray diffraction data were collected on the MX1 and MX2 beamlines of the Australian Synchrotron with the
support of the New Zealand Synchrotron Group for synchrotron-access arrangements. We wish to thank all of the beamline staff of the MX beamlines for their support in X-ray data collection.

References


