Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Exploiting subtle structural differences in heavy-atom derivatives for experimental phasing

Jimin Wang,* Yue Li and Yorgo Modis

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

Correspondence e-mail: jimin.wang@yale.edu

Received 7 March 2014
Accepted 19 April 2014

Structure determination using the single isomorphous replacement (SIR) or single-wavelength anomalous diffraction (SAD) methods with weak derivatives remains very challenging. In a recent structure determination of glycoprotein E2 from bovine viral diarrhea virus, three isomorphous uranium-derivative data sets were merged to obtain partially interpretable initial experimental maps. Small differences between them were then exploited by treating them as three independent SAD data sets plus three circular pairwise SIR data sets to improve the experimental maps. Here, how such subtle structural differences were exploited for experimental phasing is described in detail. The basis for why this approach works is also provided: the effective resolution of isomorphous signals between highly isomorphous derivatives is often much higher than the effective resolution of the anomalous signals of individual derivative data sets. Hence, the new phasing approaches outlined here will be generally applicable to structure determinations involving weak derivatives.

1. Symbols and abbreviations

F_{H}, $F_{\mathrm{H}}(+)$, $F_{\mathrm{H}}(-)$, $F_{\mathrm{H}}(-)^*$: structure factors, Friedel mates and the complex conjugates of heavy-atom (H) substructures.

$F_{\mathrm{H}o}$, $F_{\mathrm{H}o}(+)$, $F_{\mathrm{H}o}(-)$, $F_{\mathrm{H}o}(-)^*$, $\mathrm{F}_{\mathrm{H}i}$, $\mathrm{F}_{\mathrm{H}i}(+)$, $\mathrm{F}_{\mathrm{H}i}(-)$, $\mathrm{F}_{\mathrm{H}i}(-)^*$: constant ($o$) and anomalous dispersion (i) portions of structure factors, Friedel mates and the complex conjugates of heavy-atom substructures, respectively.

$F_{\mathrm{H}(I)}$: structure factors of the Ith heavy-atom substructure.

F_{P}, $F_{\mathrm{P}}(+)$, $F_{\mathrm{P}}(-)$, $F_{\mathrm{P}}(-)^*$: structure factors, Friedel mates and the complex conjugates of the parent (P) macromolecular structure of interest.

F_{PH}, $F_{\mathrm{PH}}(+)$, $F_{\mathrm{PH}}(-)$, $F_{\mathrm{PH}}(-)^*$: structure factors, Friedel mates and the complex conjugates of derivatives, *i.e.* parent plus heavy atom, PH.

$F_{\mathrm{PH}o}(+)$, $F_{\mathrm{PH}o}(-)$, $F_{\mathrm{PH}o}(-)^*$: constant portion of structure factors, Friedel mates and the complex conjugates of derivatives.

$F_{\mathrm{PH}(I)}$: structure factors of the Ith closely related derivative.

3SAD: single anomalous dispersive (SAD) phasing with three independent data sets.

3SIR: single isomorphous replacement (SIR) phasing with three independent data sets.

3SAD+3SIR: phasing from a combination of 3SAD and 3SIR.

BVDV sE2: the soluble ectodomain fragment of envelope protein E2 from *Bovine viral diarrhea virus* lacking domain I (PDB entry 4ild; Li *et al.*, 2013).

BVDV E2-ECD: the full ectodomain of bovine viral diarrhea virus E2 containing domains I–III (PDB entry 4jnt; Li *et al.*, 2013).

doi:10.1107/S1399004714008943

FOM: figure of merit.
HA: heavy atom.
NCS: noncrystallographic symmetry.
U1, U2, U3, U12, U23 and U123: three individual BVDV sE2 uranium-derivative data sets and the corresponding pairwise and triply merged U data sets.

2. Introduction

A major challenge in initial structure determination with crystals diffracting to medium and low resolution using the single isomorphous replacement (SIR) or single-wavelength anomalous diffraction (SAD) methods is that heavy-atom (HA) signals of amplitude differences between derivative and native pairs and between different wavelengths, or between Friedel mates, are often too weak to be accurately measured because of radiation damage and statistical errors in the measurement. During structure determination at resolutions of 5 Å or lower, the most difficult step is *de novo* HA substructure determination (Figs. 1 and 2). Even with the correct HA trial solutions, weak signals may prevent HA refinement from generating interpretable maps. We show how this can be overcome with a model-independent phasing procedure involving redetermination and re-refinement of the HA substructure during the course of structure determination of a soluble ectodomain fragment of glycoprotein E2 from *Bovine viral diarrhea virus* (BVDV sE2; Li *et al.*, 2013). In this procedure, the HA models to be refined do not contribute to phase calculations and have no influence on HA parameter refinement themselves. Instead, somewhat independent phase information serves as external phases for such refinement. This procedure was absolutely necessary for initial map interpretation and phase-restrained structure refinement. We found during this procedure that subtle differences existed among merged uranium-derivative data sets. Here, we document how we extracted additional phase information from such differences.

BVDV and hepatitis C virus (HCV) belong to the *Flaviviridae* family. BVDV is often used as a model system for studying HCV, a serious human pathogen for which there is currently no effective vaccine (De Francesco & Migliaccio, 2005). The outer lipid envelope of these viruses contains two glycoproteins, E1 and E2, which drive cell entry by fusing the viral and cellular membranes together. Thus, structural information on E1 and E2 is critical for understanding the infection mechanisms of these viruses. BVDV E2 forms covalently cross-linked dimers with a maximal dimension nearly twice as long as the longest unit-cell parameter (Fig. 3). Each monomer has three domains: domains I, II and III. The latter can be divided further into three subdomains: IIIa, IIIb and IIIc (Li *et al.*, 2013). We initially crystallized the intact BVDV E2 protein in the low-symmetry space group *C*2 but were unable to determine the structure owing to limited diffraction resolution and severe radiation damage (Fig. 4; Supplementary Fig. S1[1]). Through proteolysis, we identified a shorter E2 construct, namely sE2, and also crystallized it in space group *C*2 but with much improved diffraction resolution. Nevertheless, it was still a low-resolution structure-determination problem owing to severe diffraction anisotropy, radiation sensitivity and a low yield of high-quality crystals (Li *et al.*, 2013).

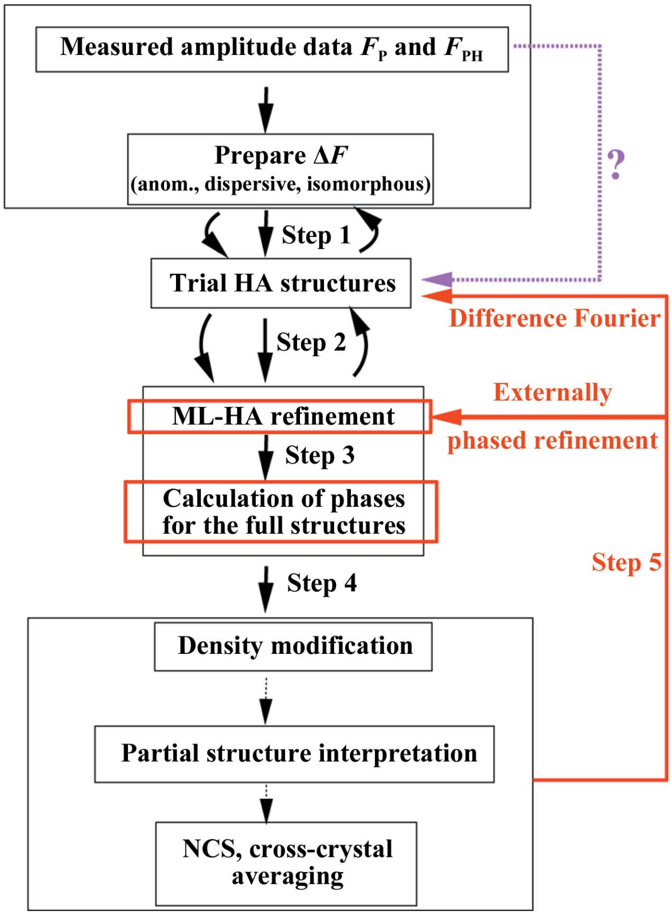


Figure 1
Computational flowchart for medium- and low-resolution structure determination with weak derivatives. In this study we identify Step 5 (red), an iterative HA structure re-determination by difference Fourier methods and re-refinement by externally phased HA refinement, as a potentially decisive addition to the conventional structure-determination steps Steps 1–4 (bold black arrows). With relatively weak HA signals, the most difficult step in structure determination is to obtain initial HA solutions, and we have to systematically examine all of a dozen or sometimes hundreds of nondiscriminatory HA trial solutions to determine whether they are correct or partially correct using a maximum-likelihood HA refinement procedure. In order to carry out Step 5, it is necessary to separate, or 'open the box' between, ML-HL HA refinement and phase calculation for the full structure (see Fig. 2) into two distinct steps.

3. When and how to merge derivative data sets and when not to merge them

We identified the U derivative from the first data set collected to about 6 Å resolution from a U-treated crystal. Soaking with uranium changed the unit-cell parameters (see Supporting Information) and resulted in detectable anomalous signals

[1] Supporting information has been deposited in the IUCr electronic archive (Reference: DZ5327).

(Dauter, 2006) as well as outstanding peaks in anomalous difference Patterson maps (Fig. 4). Similar features were observed in 12 further U-treated data sets collected over a period of several years at resolutions of up to 3.4 Å. Although the resolution of these data sets was up to 3.25–3.4 Å in the strongest direction, severe diffraction anisotropy reduced the overall resolution. According to *phenix.xtriage* analysis as implemented in *PHENIX* (Adams *et al.*, 2010; Zwart *et al.*, 2005), the effective resolution of the usable anomalous signals varied from 5.74 to 4.45 Å among the three best BVDV sE2 U data sets (Table 1), namely U1, U2 and U3. With such low-resolution anomalous signals, *de novo* determination of the HA sites remained the most challenging aspect of structure determination (Fig. 1). In fact, we failed to determine HA substructures *de novo* from each of over 30 potential derivative data sets using conventional methods. In addition to the U derivatives, we identified several other weak derivatives, including Pt, Os and Se (from partially selenomethionine-incorporated sE2 protein), which typically diffracted to lower than 6 Å resolution.

Amplitude isomorphous differences between pairs from U1, U2 and U3 varied from 8.9 to 12.6% (Table 1), as calculated using the *CCP*4 program *SCALEIT* (Winn *et al.*, 2011). The corresponding intensity differences were smaller than the merging *R*-factor statistics (such as $R_{\rm p.i.m.}$ or $R_{\rm merge}$) within each data set. This prompted us to merge these data sets together to generate a derivative data set, U123, with stronger overall anomalous signals. Because U-treated crystals diffracting to high resolution were rare, we often collected multiple data sets from each crystal until it ceased diffracting, and decided afterwards which images to include for data processing. Some images had streaky features and were excluded from data reprocessing. Another criterion used to determine whether specific images should be included in data processing of the U derivative prior to merging of data sets was whether the images contributed to recognizable peaks and features in anomalous difference Patterson maps.

Table 1
Cross-isomorphous amplitude differences among U-derivative data sets calculated using *SCALEIT* in *CCP*4.

	U2	U3	U23	U123
U1	12.6	11.4	9.6	6.6
U2		8.9	7.8	8.9
U3			4.7	6.6
U23				3.9

The three U derivatives had slightly different extents of diffraction anisotropy, making it difficult to merge them, even though they were reasonably isomorphous to each another (Table 1). We initially merged them together using two different approaches starting with (i) unmerged integrated intensities from *XDS* (Kabsch, 2010) or (ii) pre-merged intensities from *SCALEPACK* in the *HKL*-2000 suite (Otwinowski & Minor, 1997). Merging of pre-merged intensities should reveal the relative strength of each data set, providing guidance as to whether the inclusion of a specific data set in merging could be beneficial. Merging a very weak data set with a strong data set may increase the completeness

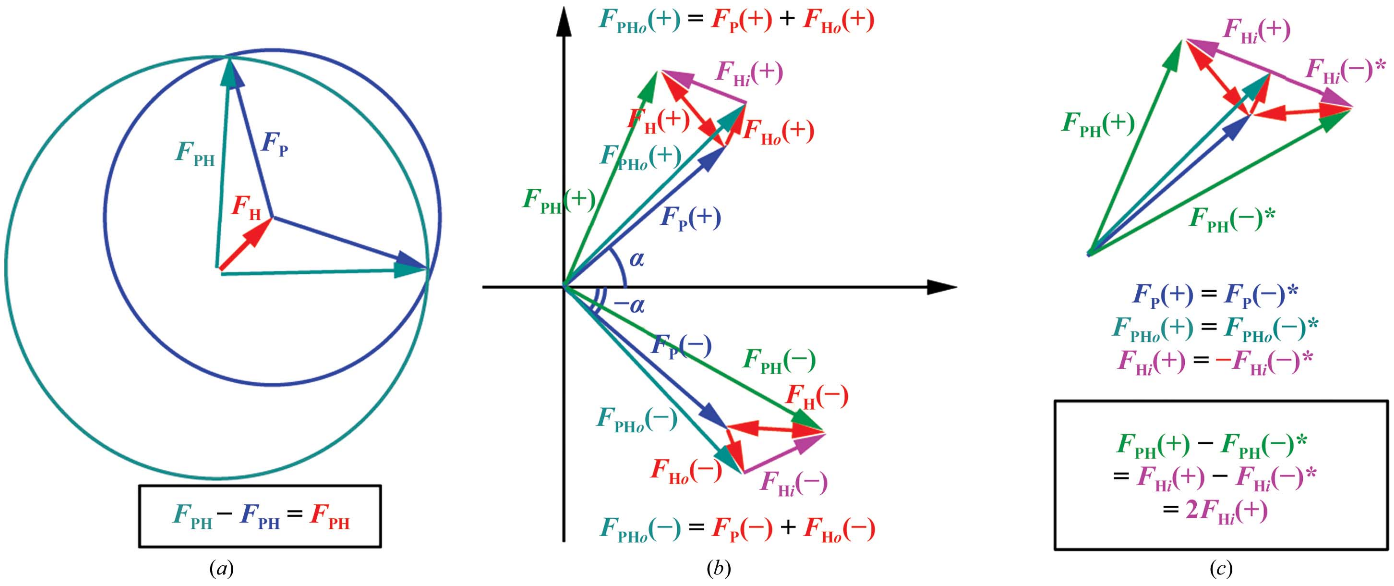


Figure 2
Basis for algebraic calculation of phases for the full structure through Harker construction for single isomorphous replacement and single-wavelength anomalous dispersion scattering. (*a*) There are two equally probable solutions to the SIR phase equation ($F_{\rm PH} - {\rm F_P} = {\rm F_H}$), giving rise to a phase-solution ambiguity. The two solutions with their corresponding phase-equation triangles are mirror-symmetry related with respect to the $F_{\rm H}$ vector. (*b*) In the presence of anomalous dispersive scatterers, the heavy-atom structure factors have two components, a constant portion ($F_{{\rm H}o}$) and an anomalous portion ($F_{{\rm H}i}$), where the anomalous portion has a constant phase shift of 90°. The anomalous dispersion breaks down the complex-conjugation relationship: $F_{{\rm H}i}(+) \neq F_{{\rm H}i}(-)^*$, $F_{\rm H}(+) \neq F_{\rm H}(-)^*$ and $F_{\rm PH}(+) \neq F_{\rm PH}(-)^*$, whereas $F_{\rm P}(+) = F_{\rm P}(-)^*$, $F_{{\rm H}o}(+) = F_{{\rm H}o}(-)^*$, $F_{\rm P}(+) + F_{{\rm H}o}(+) = F1_{\rm P}(-)^* + F_{{\rm H}o}(-)^*$. (*c*) There are also two solutions to the SAD phase equation $[F_{\rm PH}(+) - F_{\rm PH}(-)^* = F_{{\rm H}i}(+) - {\rm F}_{{\rm H}i}(-)^*]$. They are related by mirror symmetry with respect to the $F_{{\rm H}i}(+)$ vector (not shown).

of the data set but does not necessarily enhance the anomalous signals. Moreover, all existing intensity-merging procedures (such as in *XDS* or *SCALEPACK*) apply only isotropic scaling. This could reduce or wipe out anomalous signals in the highest resolution shells, particularly if the two crystals have different diffraction anisotropy. Indeed, this merging caused a reduction in the effective resolution of the anomalous signals (U12 and U123 in Table 2), whereas it generally increased the anomalous signals in the lowest resolution shells. The effective anomalous resolution in U123 was only 6.07 Å, much lower than that of each individual U data set. Besides the different anisotropy, another reason why the anomalous signals in the highest resolution shells among the three U derivatives were not correlated was because the derivatives had slightly different U substructures. After anisotropic scaling using the *CCP*4 program *SCALEIT*, merging of pre-merged intensities helped to improve the effective resolution of anomalous signals. However, such merging of pre-merged data could also misleadingly increase the apparent resolution of anomalous signals if large erroneous differences existed and could not properly be excluded, particularly in lower symmetry space groups (U23 in Table 1). Thus, there were no simple criteria or statistics to determine the best way to merge data.

A surprising finding was that only the U123 data set from merging of pre-merged intensities resulted in partially interpretable initial experimental maps, not the individual U data sets or pairwise merged data sets (Li *et al.*, 2013). Also unexpectedly, an attempt to produce a merged U123 data set using unmerged intensities failed to generate the correct HA solution, and did not produce better anomalous difference Patterson maps than merging the pre-merged intensities. This required further investigation and an explanation, which was the main motivation for this study. With the known structures of BVDV sE2 and the U derivatives, we set out to determine whether we may have unknowingly dismissed partial correct HA models along with some interpretable features in experimental maps during previous unsuccessful structure-determination attempts.

Merging multiple derivative data sets can enhance anomalous signals from dominant sites but suppress signals from minor sites, which may help to identify initial HA solutions. Not merging data sets allows their differences to be exploited for phasing in the later steps after an initial HA solution is found. This can be explained in the context of the following SIR phase equations:

$$F_{\rm PH} = \sum_{I=1}^{N} w_I F_{{\rm PH}(I)}/N; \quad F_{\rm P} + F_{\rm H} = F_{\rm PH}, \tag{1}$$

$$F_{\rm P} + F_{{\rm H}(I)} = F_{{\rm PH}(I)} \quad (I = 1, \ldots, N). \tag{2}$$

In (1), N closely related data sets of the same derivative are merged through amplitude (or intensity) averaging, weighted (w_I) by overall F/σ_Φ ratios in each data set. Let us consider an ideal case such that all derivative data sets represent exactly the same HA substructure and should be merged together. Amplitude averaging increases the F/σ_F ratio as well as the anomalous ratio by $N^{1/2}$ because this average suppresses the noise level in the original data. Another ideal situation is when measurement errors in each data set are strictly responsible for errors in resulting SIR experimental phases. Without amplitude averaging, N sets of closely related experimental phases can then be generated. By proper phase combination through addition of Hendrickson–Lattmann phase probability coefficients from each data set, assuming that they are independent of each other (see below), phase errors originating from measurement errors may also be reduced. Therefore, once the initial HA solution is known, not merging similar derivatives with real differences between them should generate better maps, as illustrated in this study. The same arguments can be made for SAD data sets but with a slightly modified phase equation:

$$F_{\rm PH}(-)^* + 2F_{{\rm H}i}(+) = F_{\rm PH}(+). \tag{3}$$

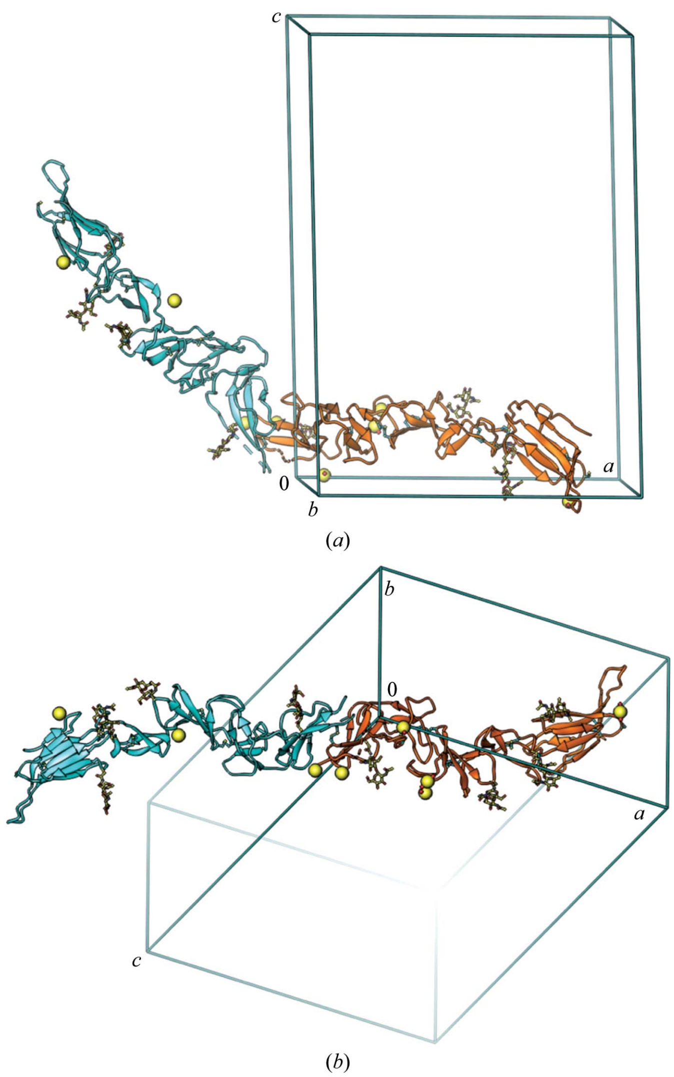
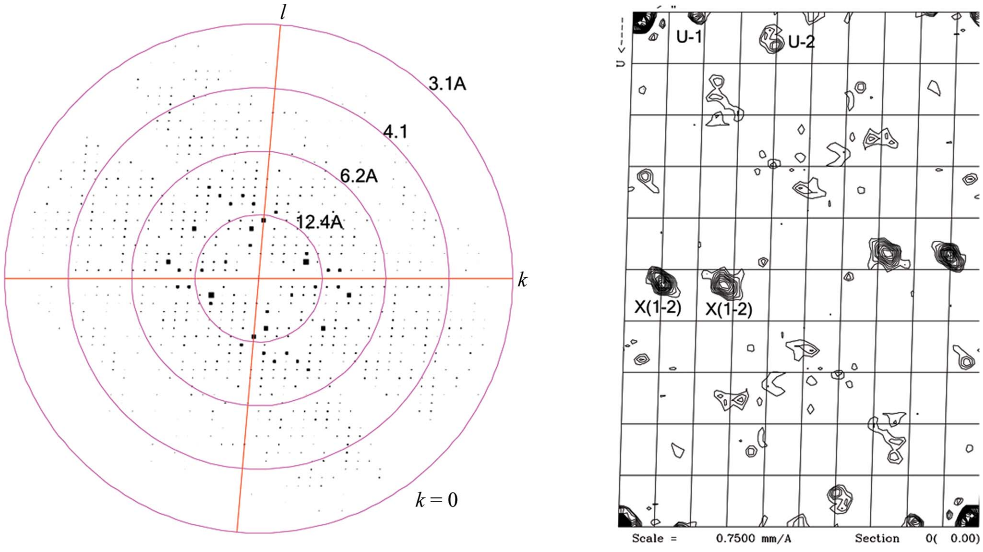


Figure 3
Disulfide cross-linked elongated BVDV sE dimer in the crystallographic asymmetric unit. (*a*) Viewed approximately along the *b* axis. (*b*) Viewed approximately along the NCS dyad. U atoms are shown as large yellow spheres; carbohydrates attached to Asn and disulfide bonds are shown in ball-and-stick representation.

Table 2
Statistics from *AutoSol* runs of various U-derivative data sets using the single-wavelength anomalous diffraction (SAD) method.

AutoSol was rerun under identical conditions for test cases in this study assuming that there were eight HA sites. The results of the U123 SAD *AutoSol* run with the partially interpretable initial experimental maps are shown in bold. Interestingly, its inverted HA structure also produced a clear solvent boundary as apparent from the relatively low *R*-factor statistics between the observed and back-transformed amplitudes during density modification. FOM stands for figure of merit for experimental phases. When correct solutions are found, one expects a large difference in overall score between the correct (original) and inverted HA substructure solutions. In addition to a low *R* factor in density modification, a large map skew value and high correlation coefficients (CC) are good indicators of a correct solution at medium and high resolution. They may not be as reliable at low and medium resolutions. Thus, visual inspection of resulting experimental maps remains the best method to evaluate the correctness of the HA substructure solutions.

	Heavy-atom search statistics				Density-modification statistics		
U HA hand	Anomalous resolution (Å)	Sites	FOM	Overall score	*R* factor (%)	Map skew	Density CC
U1 original	5.10	10	0.334	31.79 ± 16.85	39.32	0.11	0.73
U1 inverted				30.89 ± 17.02	39.28	0.11	0.73
U2 original	4.45	6	0.336	21.94 ± 17.04	45.15	0.06	0.81
U2 inverted				21.82 ± 17.02	45.52	0.06	0.82
U3 original	5.74	6	0.321	23.10 ± 17.16	40.20	0.07	0.71
U3 inverted				21.10 ± 16.90	39.50	0.05	0.68
U12 original	5.67	7	0.287	21.59 ± 16.97	39.06	0.06	0.65
U12 inverted				19.43 ± 16.49	Failed		
U23 original	3.75	10	0.304	22.70 ± 17.13	39.77	0.06	0.72
U23 inverted				21.99 ± 17.04	41.23	0.06	0.71
U123 original	6.07	7	0.376	**34.25 ± 16.14**	**25.70**	**0.12**	**0.76**
U123 inverted				24.39 ± 17.20	28.11	0.07	0.65



(*a*) (*b*)

Figure 4
Diffraction data of BVDV sE2. (*a*) Severe anisotropy in a representative data set of BVDV sE2 viewed in the (*h*0*l*) zone. (*b*) Harker section of an anomalous difference Patterson map using amplitude-averaged triple isomorphous uranium (U123) data. Two strong peaks are actually cross-peaks labeled X(1–2), coincidentally located near the Harker section, generated from for the two major U sites labeled U-1 and U-2 at the Harker section.

4. Pseudo-symmetry and determination of U-derivative substructures

Prior to initial HA solution, recognizable features in anomalous difference Patterson maps were our most reliable metric for whether to merge data. Any combinations of merging the U1, U2 and U3 data sets in pairs or as a triplet resulted in increased heights in the two most outstanding peaks compared with Patterson maps calculated using any individual U-derivative data set (Fig. 4). This was the main reason that we pursued the ultimately successful SAD experimental phasing using the merged U123 data set (Li *et al.*, 2013).

We reran the *AutoSol* routine in this study using the merged U123 data set under the SAD option (Zwart *et al.*, 2008; Adams *et al.*, 2010; Terwilliger *et al.*, 2009). In this map, we could identify the main features of domain II with a recognizable β-strand sandwich core for both molecules in the asymmetric unit. However, we could not derive the NCS matrix for the two non-dyad-related domains II for NCS averaging. The automatic NCS-detection algorithm in *AutoSol* failed to identify the NCS using the known HA sites because all U sites were located on the surfaces of dimers involved in the packing of dimers in the crystal and did not obey the dyad symmetry. In addition, there was very little electron density at the dimer interface near the pseudo-dyad, which was outside the initially defined protein boundary. Moreover, the elongated dimer spans 190 Å, extending into two adjacent unit cells (Fig. 3), and increasingly deviates from a dyad relationship as the distance from the dyad increases. We therefore had to generate experimental maps of sufficient quality to properly define the boundary of each domain before any NCS domain averaging could be performed.

We reran the *AutoSol* routines in this study using the SAD option with all individual U data sets as well as various pairwise merged U data sets under the optimal conditions to reaffirm our original conclusion that any of these combinations indeed failed to result in interpretable experimental maps (Table 2). Pairing each U derivative with a pseudo-native data set generated with back-soaking methods also failed to identify HA solutions using the SIRAS option (Li *et al.*, 2013). This analysis rules out the possibility that our inability to obtain the correct HA solution in our original study with these combinations was owing to inadvertent dismissal of correct solutions or suboptimal conditions.

We examined each HA solution in *AutoSol* reruns using all individual U data or all pairwise or triplet merged U data, comparing the observed anomalous difference Patterson maps with the Patterson maps calculated using the corresponding HA coordinates from each run. Unexpectedly, the two highest outstanding peaks in the observed anomalous difference Patterson maps (Fig. 4) were actually cross-peaks for two major sites rather than true Harker peaks, although they coincidentally appeared near the Harker section. The correct Harker peaks for the first two major U sites with the highest occupancy sites were much smaller than the cross-peaks, and were unambiguously identifiable only in the maps calculated using the merged U123 data set. Nevertheless, most of the unsuccessful *AutoSol* runs correctly identified these cross-peaks as such.

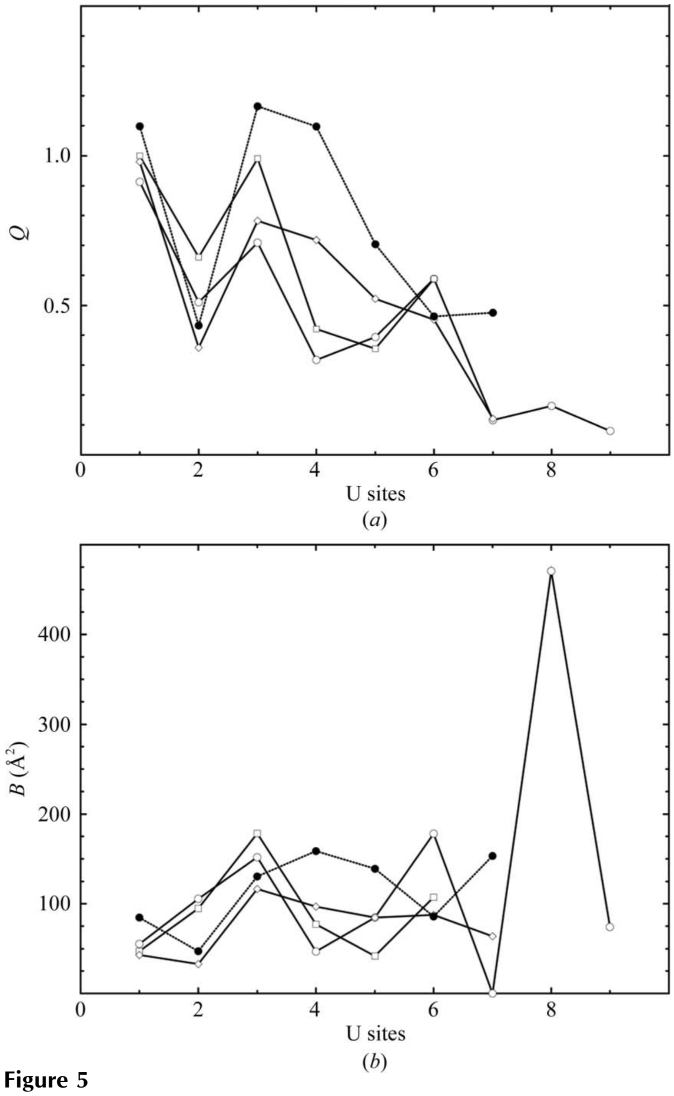


Figure 5
Heavy-atom substructures after iterative refinement. Four isomorphous U derivatives, U1, U2, U3 and U4, are shown as open circles, open squares, open diamonds and filled circles, respectively. (*a*) Anomalous occupancy (Q). (*b*) Temperature factors (B). Subtly different occupancies (and temperature factors) were exploited for isomorphous difference phasing between U data sets.

We then examined the symmetry of the HA models by calculating anomalous difference Fourier maps using experimental phases before and after density modification in the two SAD phasing attempts using either U123 or U23. The peak heights in the failed U23 attempt were more similar for the two dominant sites related by the strong cross-peaks than in the successful U123 attempt, suggesting possible centrosymmetry within the HA sites. When major HA sites have the same y value in any polar space groups such as C2, the y plane in which the sites are located is a mirror plane, independent of whether the reference y value of this plane is arbitrarily set to be zero or nonzero. Phases of such HA models have only two possible values, α_{ky} or $\alpha_{ky} + \pi$, where α_{ky} is a value that is only dependent on the product of Miller index k and the reference y value. The mirror-symmetry property of the HA model is passed onto protein phases in SIR phasing, resulting in electron density with mixed handedness. The *Phase-O-Phrenia* plot (Grosse-Kunstleve & Adams, 2003) and the heights of the other HA peaks confirmed the presence of pseudo-centrosymmetry in all HA solutions, including the correct HA solution of the U123 SAD *AutoSol* run. In all of our unsuccessful *AutoSol* runs, the extent of pseudo-centrosymmetry in the HA solutions was greater than in the successful run. For example, the amplitude-merged U23 data set exhibited a higher degree of pseudo-centrosymmetry in the two dominant sites than the amplitude-merged U123 data set. Thus, despite having only a 3.9% amplitude difference between the U23 and U123 data sets (Table 1), the ability to break down the HA pseudo-centrosymmetry was the tipping point for the vastly divergent results of the *AutoSol* runs using these two data sets (Table 2). The presence of pseudo-centrosymmetry in the HA substructure made it difficult to distinguish between the correct and incorrect HA solutions. Incorrect centrosymmetric HA solutions may have had overinflated scores, and the correct centrosymmetric HA solutions may thus have been rejected by automated routines.

5. Exploiting subtle differences among individual U-derivative substructures

Given the relatively low resolution of anomalous signals in the U123 data set, the quality of the initial experimental maps was not sufficient for complete interpretation of the sE2 backbone or determination of the NCS between the two molecules in the asymmetric unit. To improve the experimental phases, we employed two parallel approaches: (i) finding HA solutions of other weak derivative structures and (ii) exploiting subtle structural differences between U derivatives. Additional weak derivative structures could not be solved at this stage because the U derivatives had very different unit-cell parameters from all other derivatives or true native data sets. In the first approach, an optimal transfer of the best experimental phases from U derivatives to other potential derivatives through multiple-crystal averaging required accurate knowledge of the positions of each domain in the asymmetric unit, which was missing at this stage. The second approach was successful after we established that subtle structural differences existed between U derivatives using cross-difference Fourier methods.

We found that the U1, U2 and U3 derivatives were not identical based on the anomalous difference Fourier maps as well as the final refined HA parameters (Fig. 5). The present analysis was carried out using the final best experimental phases, which were obtained through an iterative multi-crystal multi-domain averaging procedure (see Supporting Information) and were no longer biased towards any HA sites. The peak heights at the two major U sites in the anomalous difference Fourier maps of each individual U data set were lower than those in the merged data set, *i.e.* merging the data enhanced the anomalous signals for the two major sites. The peak heights at the remaining minor sites were increased in U1 and U2 but decreased in U3; that is, merging the data also suppressed the anomalous signals for minor sites, making the HA structure appear more pseudo-symmetric with reduced effective resolution and a smaller number of dominant sites. The overall gaps between major and minor sites in the three U derivatives were reduced relative to the merged structure. Fewer dominant sites and increased gaps between major and minor sites were the main reasons why an initial HA solution was found for the merged U123 derivative.

Differences in anomalous occupancy between different U data sets were further confirmed in the cross-isomorphous difference Fourier maps. For example, in the U2–U3 difference Fourier maps we expected a positive peak in site 2 and a negative peak in site 3 according to their relative occupancies as established in anomalous difference Fourier maps as well as refined HA parameters (Fig. 5). Indeed, we observed exactly what we expected. We found a general correlation between peak heights in anomalous difference Fourier maps and the signs of peaks in isomorphous difference Fourier maps. Lastly, when we calculated the HA structure factors using refined real occupancies, we found that the correlation of their amplitudes was less than 30% among individual data sets beyond 6.5 Å resolution. Together, these observations confirmed that the U derivatives were similar but not identical.

If isomorphous differences between the U2 and U3 data sets can be explained by differences in their HA substructures, we would expect to be able to derive new phase information from their isomorphous differences by treating one data set as a reference pseudo-native data set using the SIR phase relationship similar to equation (2). There are three possible pairs: U1–U2, U1–U3 and U2–U3. In fact, when we calculated experimental phases from three SAD data sets and three SIR pairs and combined them together in the procedures described below, we observed a much clearer solvent boundary in the resulting experimental map than in the initial map generated by the U123 SAD *AutoSol* run prior to density modification, *i.e.* the *Phaser*-derived experimental map (McCoy *et al.*, 2007). The SIR method described here differs from conventional SIR in that the 'native' set is actually a derivative set but is treated as a native set. Whereas closely related derivatives can often be included to calculate pseudo-MIR experimental phases, the extraction of extra experimental phase information by directly pairing derivatives with each other has not been reported previously in detail. When paired derivative/(pseudo-native) derivatives are more isomorphous to each other than any pairing of each derivative to a true native data set, the derivative/(pseudo-native) derivative pairing provide more accurate phase information than conventional derivative–native pairing.

Density modification such as solvent flattening consistently results in phase improvement (Wang, 1985), especially with high solvent contents such as the 60% in our structure. However, the success of density modification requires a reasonably accurate model of the solvent boundary. In the case of SAD phasing with strong HA signals, a combination of direct-methods approaches has been reported to resolve phase ambiguities (Wang *et al.*, 2004). Our study through phase combination of 3SAD+3SIR among three highly isomorphous U derivatives is one of only a few studies dealing with low-resolution weak derivatives. This combination provided the minimal information required for determination of the solvent boundary and successful density modification.

We analyzed the back-soaked U data sets using isomorphous difference Fourier methods and found that the majority of U atoms had indeed diffused out of the crystal within the first 30 s of back-soaking (see Supporting Information). An analysis of the major U-binding sites in our structure shows that they are made of multiple carboxylates between molecules that bind weakly to UO_2^{2+} (see Supporting Information). After back-soaking, repulsive interactions among carboxylates without UO_2^{2+} could disrupt the crystal lattice. We also reaffirmed that soaking with 0.2 m*M* UO_2^{2+} did not result in any binding according to results of our isomorphous difference Fourier analysis (Li *et al.*, 2013). Using the back-soaked data set as a pseudo-native for the U derivatives, we produced further improved experimental phases as judged from the clearer solvent boundary than with any pairs of U derivatives prior to density modification. The U and back-soaked native pair provides SIR phasing information mainly at very low resolution, whereas the U–U pair provides phasing at higher resolution since there are no significant differences between U derivatives at very low resolution.

6. Model-independent phasing method with externally phased HA refinement

In the determination of the BVDV sE2 structure, the anomalous signals in U derivatives were limited to only very low resolution (up to 6.0 Å), whereas the isomorphous signals between U-derivative pairs were limited to medium resolution (3.5–5 Å; Li *et al.*, 2013). Because the resolution ranges of the SAD phasing and SIR phasing did not overlap, the ambiguity in the phase triangular relationship could not be resolved (Fig. 2). The resulting protein phases were hence the same as the HA model phases, and possessed pseudo-centosymmetry, which impaired determination of the solvent boundary.

HA model refinement is prone to HA model bias when (i) HA models are incomplete or inaccurate during the early stages of HA structure determination, (ii) HA signals are very weak and only at very low resolution and (iii) the resolution ranges of SAD and SIR phasing do not overlap, all of which problems we encountered during the determination of the BVDV sE2 structure. Incomplete HA models are not uncommon during the early stages of structure determination (Silvian *et al.*, 1999). They can generate an overinflated average cosine of estimated phase errors, or figure of merit (FOM), because irremovable errors are present in HA phases for the construction of the phase triangulation relationship during HA model refinement (see the Supporting Information for a specific example and see below for further discussion on the FOM inflation problem). Maximum-likelihood HA parameter refinement procedures provide the best protein phase information in such situations; however, they cannot mathematically resolve the phase-ambiguity problem when SAD and SIR have non-overlapping resolution ranges (Fig. 2).

HA model bias can be reduced by refining the HA models against externally generated phases without a direct contribution from the HA models to be refined. This is the basis of the model-independent phasing method presented in this study. We refined HA structures using a modified version of the maximum-likelihood refinement program *MLPHARE* and combined phases using *SIGMAA* as implemented in *CCP*4 (Winn *et al.*, 2011; Collaborative Computational Project, 1994; Read, 1997; Otwinowski, 1991). The novelty of our procedures is to bypass an internal decision-making process at the user interface. More specifically, we identified a specific optimal subset of the derivatives to be used for final phasing. When the experimental phases initially produced by the U123 SAD *AutoSol* run were used as external phases, HA refinement became very robust and stable. In this refinement, the HA models were effectively real-space refined by fitting into the anomalous or isomorphous difference Fourier densities with fixed phases derived from the previous iterative run and fixed HA signals of amplitude differences (Rould *et al.*, 1992; Wang, 2010). Nevertheless, the refinement was still carried out in reciprocal space for all reflections. Once the HA models were reliably refined, phases could be calculated for the full structure using simple algebraic or probabilistic procedures (Fig. 2). This iterative procedure is particularly suitable for the exploitation of subtle structural differences among U derivatives. This procedure eliminates any problem of possible correlation among closely related HA

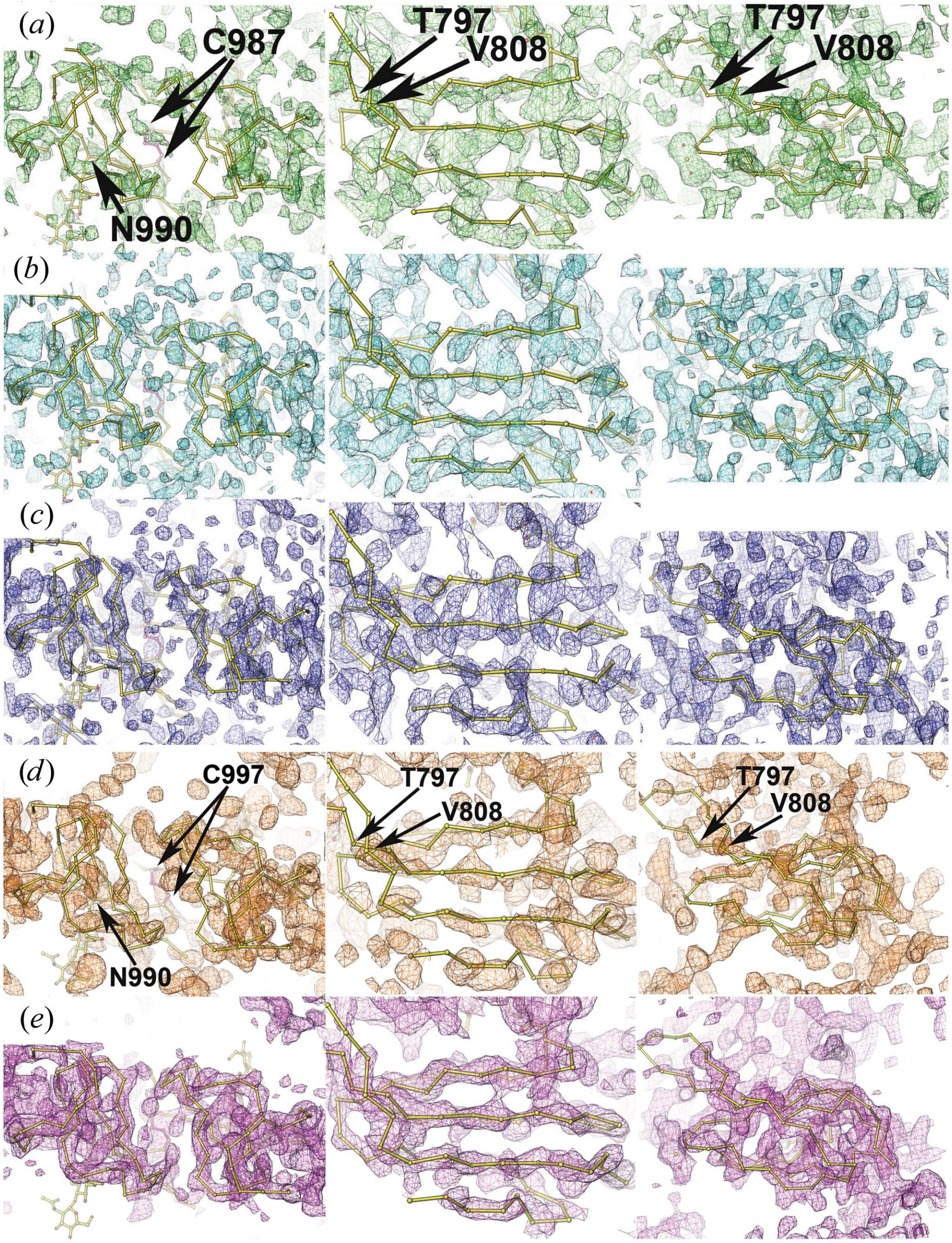


Figure 6
Experimental maps before and after density modification. (*a*) Amplitude-averaged SAD maps generated using the *PHENIX AutoSol* routine. (*b*, *c*) Phase-averaged 3SAD+1SIR maps from the second (*b*) and third (*c*) passes of iterative heavy-atom substructure refinement. Left, domain IIIc; middle and right, two orthogonal views of domain II. Maps are contoured at 1.5σ. The correct boundary for domain IIIc could not be determined in (*a*) but was identified in (*b*) and (*c*) by automated algorithms. (*d*) Amplitude-averaged SAD maps with the same sections and same contours as in Fig. 4. (*e*) Phase-averaged 3SAD maps.

Table 3
Effective anomalous resolutions of recently solved structures in our laboratories.

Structure	HA	Final No. of sites	Anomalous resolution (Å)	Methods for initial solutions	Finding HA solution *SHELXD*	Finding HA solution *AutoSol*	References
Gin–DNA	Os	2	11.8	SIRAS	Yes	No	Ritacco *et al.* (2013)
DnaB–DNA†	Ta_6Br_{12}	24	6.68	SIRAS, SAD	Yes	No	Itsathitphaisarn *et al.* (2012)
YfbU†	W_{12}	8	6.20	SIRAS, SAD	Yes	No	Wang & Wing (2014)
BVDV sE2	U	8	6.07	SAD	Partial	Yes	This study
PAN ENE†	Ir	8	5.73	SIRAS, MAD	Yes	No	Mitton-Fry *et al.* (2010)
Gin†	Hg	2	5.73	SIRAS, SAD	Yes	No	Ritacco *et al.* (2014)
Gly switch	Ir	11	5.39	SAD, MAD	Partial	No	Butler *et al.* (2011)
c-di-GMP	Ir	10	3.30	SAD, MAD	Yes	Yes	Smith *et al.* (2009)

† In these four cases, the percentage of finding correct or partially correct solutions in *SHELXD* runs using the SIRAS method was much higher than in the corresponding runs using the SAD method, by as much as 60-fold. However, the corresponding *PHENIX AutoSol* runs failed to find any correct solutions using either the SIRAS or SAD methods.

derivatives because the derivative structures to be refined do not contribute to phase calculation during refinement.

In the first pass, we calculated an experimental map using combined phases calculated from the anomalous signals of the three individual U1, U2 and U3 derivatives (3SAD) and compared it with the corresponding map generated from the U123 SAD *AutoSol* run prior to density modification. The solvent boundary was clearer in the new map. In the second pass, we calculated another map (3SAD+3SIR) by adding phases calculated from the isomorphous difference signals in the U1–U2, U1–U3 and U2–U3 pairs and compared it with the 3SAD map. In addition to improvement of the solvent boundary in this new map, some β-strands could now be recognized prior to density modification, particularly in domain IIIc (Fig. 6), for which the density was very poor in earlier maps. With such improvement, we were able to identify the dyad-related NCS relationship between the two molecules in the asymmetric unit. Domain II of the second molecule could then be located and the corresponding NCS operator could be calculated.

In our experience of phase combination using *SIGMAA* for two sets of phase probability curves described by Hendrickson–Lattmann coefficients, whenever we observed visual improvement of the solvent boundary the FOM in the same resolution range also increased. Conversely, decreases in the FOM were always accompanied by a deterioration of the combined experimental maps. However, increases in the FOM were not always accompanied by improvement of the solvent boundary or phases. For example, the 3SIR (U1–U2, U1–U3 and U2–U3) experimental map had a reasonably clear solvent boundary, and so did the U1–native paired SIR map, where the ‘native’ was actually a 30 s back-soaked pseudo-native (see Supporting Information). When their phases were combined, the FOM increased but the resulting experimental maps deteriorated, suggesting that the FOM was inflated and no longer described the accuracy of the resulting phases objectively, as discussed above. In general, phase combination using anomalous signals (SAD) causes less FOM inflation than using isomorphous signals (SIR). If the FOM is severely inflated, experimental phases may not improve after density modification. For example, the 3SIR+3SAD map had a clearer solvent boundary than the 3SAD map before density modification, but was less informative about the structure after density modification. Major inflation occurred when the least isomorphous SIR data-set pairs were included (U1–U2 and U2–U3). By excluding these two SIR pairs, the remaining 1SIR+3SAD produced better maps after density modification (Fig. 6). Unfortunately, our studies do not provide guidelines or predictions on how to merge all possible phase information from various possible pairs, which remains to be determined on a trial-and-error basis. Our calculations suggest that whereas any isomorphous pairs may provide new phase information, only those pairs that have the largest difference peaks in isomorphous difference Fourier maps and the smallest non-isomorphous differences can provide experimental phases with the most accurately estimated FOM. The inflation of the FOM is a major problem in earlier stages of the SIR/SAD phasing when the external phases still contain some ghost phase solutions that could inflate the occupancies of HA models. With gradual improvement of experimental phases after elimination of ghost solutions in iterative procedures, the FOM becomes more reliable. In turn, this facilitates HA model refinement and new phase calculation.

Using the same stepwise procedures, we were able to add a fourth U derivative in SAD/SIR phasing and two pseudo-native data sets (30 and 60 s back-soaked) in SIR phasing (Li *et al.*, 2013). By adding these weak derivative pairs to experimental phasing, the boundaries of each domain in both molecules in the asymmetric unit could be identified, and all NCS matrices could be established for all domains. This permitted multi-crystal multi-domain averaging during density modification (see SOM), which further improved the experimental maps (Supplementary Fig. S2).

It should be noted that the SIR+SAD iterative HA refinement in this study is different from conventional SIR/AS in two aspects, as is 3SIR from MIR. Firstly, in conventional SIR/AS the real and anomalous occupancies are in proportion to each other as a function of the type of HA and the wavelength used for data collection. In the SIR+SAD case in this study the real component of the HA structure is actually a difference structure between two derivatives and bears no relationship to the corresponding anomalous component of the HA model. These two components are refined independently. Additionally, negative real occupancy is permitted when the native and derivative data (or the high- and low-occupancy derivative data) are intentionally switched if the high-occupancy data are more complete and extend to higher resolution than the low-occupancy data. In the case of MIR of closely related derivatives, HA substructures are highly correlated and proper correlation matrices should be included in refinement (Kohlstaedt *et al.*, 1992; Sygusch, 1985; Bricogne *et al.*, 2003; Pannu *et al.*, 2003; Pannu & Read, 2004). Secondly, the HA refinement is carried out with externally supplied phases with no direct biases to any given derivative being refined. In this case, all phase information derived from independently measured anomalous signals can be treated as truly independent sources. It is important to note that although phase information derived from our 3SIR data may appear to be independent, by the nature of triangulation phase information from only two of the three possible circularly permutable pairs (U1–U2, U2–U3 and U3–U1) can be treated as independent sources.

The model-independent phasing method described in this study has been routinely carried out in our laboratories in the last few years when dealing with medium- and low-resolution structure determination (Ritacco *et al.*, 2013, 2014; Mitton-Fry *et al.*, 2010; Wang & Wing, 2014; Itsathitphaisarn *et al.*, 2012; Butler *et al.*, 2011; Smith *et al.*, 2009). In all of these cases, with the exception of BVDV sE2 described in this study (Table 3), initial HA solutions were derived using *SHELXD* (Sheldrick, 2010). With the *SHELXD*-derived HA sites, initial experimental maps were often calculated using the *SOLVE*/*RESOLVE* program suite (Terwilliger & Berendzen, 1999; Terwilliger, 2000). In all of these cases, whenever the effective resolutions of anomalous signals were relatively low (Table 3) iterative HA parameter re-refinement greatly improved the experimental maps. Although the initial Ta_6Br_{12} sites were correctly determined using *SHELXD* under both the SAD and SIRAS options at 5.6 Å resolution during structure determination of the DnaB helicase–DNA complex (Itsathitphaisarn *et al.*, 2012; Sheldrick, 2010), these sites were also found using direct methods with the amplitudes of the derivative data alone (Miller *et al.*, 2007). In fact, the much higher success rates of finding correct sites in *SHELXD* under the SIRAS option (about 90 out of 100 trials) than under the SAD option (none within the first 100 trials but three out of 200 trials) are attributable mainly to the larger amplitudes of the derivative data.

7. Concluding remarks

This study has demonstrated that there are advantages in collecting multiple data sets of the same derivative from single or multiple crystals for experimental phasing. By merging them together, statistical errors are reduced, noise is filtered out and HA signals are enhanced. Once initial phases have been obtained, comparison of the unmerged data sets allows subtle structural differences in HA substructures to be exploited for additional power for experimental phasing. These differences can be radiation-induced changes in HA substructures or inherent variations in the crystals. Thus, any experimental design for medium- and low-resolution structure determination should include the collection of at least three data sets from a known derivative for application of the methods described in this study as well as by Hendrickson and coworkers (Liu *et al.*, 2012). Although maximum-likelihood HA refinement can somewhat account for the effects of non-isomorphism problems, the best approach remains to empirically find the optimal isomorphous native–derivative pair. This study has also shown that the *de novo* generation of the correct HA solution from weak anomalous or isomorphous signals at low resolution remains the most challenging aspect of structure determination. Given that the effective resolution of amplitudes is much higher than the effective resolution of either anomalous or isomorphous signals calculated from amplitude differences between paired reflections, it remains to be seen whether the amplitudes themselves, not just their differences, can assist in the identification of initial HA solutions.

This work was supported by a Burroughs Wellcome Investigator in the Pathogenesis of Infectious Disease Award and National Institutes of Health Grants P01 GM022778 and R01 GM102869 (to YM) and by the Steitz Center for Structural Biology, Gwangju Institute of Science and Technology, Republic of Korea.

References

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

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

Butler, E. B., Xiong, Y., Wang, J. & Strobel, S. A. (2011). *Chem. Biol.* **18**, 293–298.

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

Dauter, Z. (2006). *Acta Cryst.* D**62**, 867–876.

De Francesco, R. & Migliaccio, G. (2005). *Nature (London)*, **436**, 953–960.

Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Acta Cryst.* D**59**, 1974–1977.

Itsathitphaisarn, O., Wing, R. A., Eliason, W. K., Wang, J. & Steitz, T. A. (2012). *Cell*, **151**, 267–277.

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

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A. & Steitz, T. A. (1992). *Science*, **256**, 1783–1790.

Li, Y., Wang, J., Kanai, R. & Modis, Y. (2013). *Proc. Natl Acad. Sci. USA*, **110**, 6805–6810.

Liu, Q., Dahmane, T., Zhang, Z., Assur, Z., Brasch, J., Shapiro, L., Mancia, F. & Hendrickson, W. A. (2012). *Science*, **336**, 1033–1037.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.

Miller, R., Shah, N., Green, M. L., Furey, W. & Weeks, C. M. (2007). *J. Appl. Cryst.* **40**, 938–944.

Mitton-Fry, R. M., DeGregorio, S. J., Wang, J., Steitz, T. A. & Steitz, J. A. (2010). *Science*, **330**, 1244–1247.

Otwinowski, Z. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous scattering*, edited by A. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 303–326.

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

Pannu, N. S. & Read, R. J. (2004). *Acta Cryst.* D**60**, 22–27.

Read, R. J. (1997). *Methods Enzymol.* **277**, 110–128.

Ritacco, C. J., Kamtekar, S., Wang, J. & Steitz, T. A. (2013). *Nucleic Acids Res.* **41**, 2673–2682.

Ritacco, C. J., Steitz, T. A. & Wang, J. (2014). *Acta Cryst.* D**70**, 685–693.

Rould, M. A., Perona, J. J. & Steitz, T. A. (1992). *Acta Cryst.* A**48**, 751–756.

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

Silvian, L. F., Wang, J. & Steitz, T. A. (1999). *Science*, **285**, 1074–1077.

Smith, K. D., Lipchock, S. V., Ames, T. D., Wang, J., Breaker, R. R. & Strobel, S. A. (2009). *Nature Struct. Mol. Biol.* **16**, 1218–1223.

Sygusch, J. (1985). *Methods Enzymol.* **115**, 15–22.

Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.

Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H. & Hung, L.-W. (2009). *Acta Cryst.* D**65**, 582–601.

Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.

Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.

Wang, J. (2010). *Acta Cryst.* D**66**, 988–1000.

Wang, J. W., Chen, J. R., Gu, Y. X., Zheng, C. D., Jiang, F., Fan, H. F., Terwilliger, T. C. & Hao, Q. (2004). *Acta Cryst.* D**60**, 1244–1253.

Wang, J. & Wing, R. A. (2014). *Acta Cryst.* D**70**, 1491–1497.

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

Zwart, P. H., Afonine, P. V., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., McKee, E., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K., Storoni, L. C., Terwilliger, T. C. & Adams, P. D. (2008). *Methods Mol. Biol.* **426**, 419–435.

Zwart, P. H., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *CCP4 Newsl. Protein Crystallogr.* **43**, contribution 7.