Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Application of general formulas for the correction of a lattice-translocation defect in crystals of a lentiviral integrase in complex with LEDGF

Stephen Hare,[a] Peter Cherepanov[a] and Jimin Wang[b]*

[a]Division of Medicine, St Mary's Campus, Imperial College London, Norfolk Place, London W2 1PG, England, and [b]Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520-8114, USA

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

Received 1 May 2009
Accepted 22 June 2009

The symmetry inherent to many biological macromolecular assemblies has been implicated in a range of crystal pathologies, including lattice-translocation defects (LTDs). Crystals suffering from classic LTDs contain two lattices that are shifted with respect to each other but nonetheless remain within the length of coherent interference. LTD introduces an undesirable intensity modulation into diffraction data, resulting in scrambled or partially scrambled electron densities. In this report, LTD theory is extended and a new general method for determining defect fractions is developed based on the heights of the non-origin peaks observed in native Patterson maps. The application of this method to crystals of lentiviral integrase in complex with its cofactor, where the observed translocation vector does not equal a small integral fraction of a unit-cell edge, is reported and its general application to all classic LTD cases is predicted.

1. Introduction

Homo-multimeric macromolecules related by noncrystallographic symmetry (NCS) can sometimes allow alternative alignments of successive layers in a growing crystal, resulting in a polysynthetic twin containing two (or more) identical coherently diffracting lattices. Such crystal pathology, which is commonly referred to as a lattice-translocation defect (LTD), order–disorder twinning or one-dimensional disorder, can occur in crystals of homo-multimeric as well as monomeric macromolecules (Howells & Perutz, 1954; Bragg & Howells, 1954; Trame & McKay, 2001; Wang, Kamtekar *et al.*, 2005; Wang, Rho *et al.*, 2005; Hwang *et al.*, 2006; Tanaka *et al.*, 2008; Zhu *et al.*, 2008). One common case of classic LTD results in a pair of lattices whose crystallographic origins maintain the same environment, for example $\Delta z = 1/2$, resulting in an origin shift to an equivalent but not identical crystallographic position in monoclinic space groups (Wang, Kamtekar *et al.*, 2005). Systematic phase shifts between the two lattices result in strong modulations in the observed diffraction intensities. Owing to the presence of multiple crystallographic origins in such crystals, the averaged contents of the polysynthetic crystal can be represented by overlapping structures with different occupancies (Wang, Kamtekar *et al.*, 2005; Zhu *et al.*, 2008). Polysynthetic macromolecular crystals belong to general polytypic structures; in crystals of small molecules, layer structures of multiple distinct unit-cell contents can be modeled owing to sufficiently large observation-to-parameter ratios (Durovic, 1992). Our previous and current treatments are designed to reduce the multiple unit-cell contents to the single unit-cell content in the absence of sufficient observations in the diffraction data from macromolecular crystals (Wang, Kamtekar *et al.*, 2005; Wang, Rho *et al.*, 2005).

Howells, Perutz and Bragg, who studied crystals of tetrameric imidazole-methemoglobin, reported the first classic LTD case in 1954; to our knowledge, the structure of this crystal form remains undetermined (Howells & Perutz, 1954; Bragg & Howells, 1954). In the past half century, new cases of LTDs have sporadically been reported, with some theoretical consideration but no practical proposals for structure determination (Glauser & Rossmann, 1966; Pickersgill, 1987). More recent reports and efforts notwithstanding, the full extent of the LTD problem in macromolecular crystallography remains unknown because most undetermined structures remain unreported. The hallmark of crystals with LTD is an unusual diffraction pattern with subsets of sharp and streaked Bragg spots. The extent of streaked reflections is a function of the randomness of the distribution of the translocated layers in the crystals (Cochran & Howells, 1954; Wang, Kamtekar *et al.*, 2005). No streaks will be present if translocation occurs regularly, for example every other or every third layer, which would double or triple the unit-cell parameters. Streaks only occur for those Bragg spots that suffer from negative interference in structure factors from the alternate lattices. The extent of diffraction-spot smearing depends on the frequency of the lattice translocation. Streaks may not be observable when the affected Bragg spots suffer from minimal interference because of a very low defect fraction or from maximal interference, which could result in zero intensity in some cases. In addition, the LTD problem may escape detection when streaks only occur in a small subset of reflections. It is important to note that the streaks observed in diffraction images from crystals affected by LTD differ from those caused by correlated dynamic motions of macromolecular complexes within a given single-crystal lattice. In the latter case, most or all of the Bragg spots are affected (as well as non-Bragg reflections), but the contribution of thermal diffuse scattering to their total intensities is small (Doucet & Benoit, 1987). In the LTD case, the smearing is more pronounced and is distributed unevenly among the Bragg spots. Another hallmark of LTD is the presence of strong non-origin peaks in native Patterson maps which correspond to physically impossible packing defined by the translocation vector(s) relating crystallographic origins of the intermingled lattices. Structure determination and analysis of crystal packing may be required to distinguish between LTD and translational NCS.

2. General correction formulation for the LTD problem

Five decades after the initial description of the classic LTD (Howells & Perutz, 1954; Bragg & Howells, 1954), it was shown to be possible to unscramble the unit-cell contents in two special cases of LTD (Wang, Kamtekar *et al.*, 2005; Wang, Rho *et al.*, 2005). In the first case, the translocation vector $\mathbf{t}_d$ was (0, 0, 1/2), which affects the intensity of reflections with odd l indices only, and in the second case $\mathbf{t}_d$ was (0, 0, 1/3), which affects the intensity of reflections with l indices that are not divisible by three. In both cases, it was possible to estimate the defect fractions by examining the extent of the modulation of the affected reflections. However, such a specialized method is not applicable to a generalized translocation vector, which may not represent a small integral fraction of a unit-cell edge, for example $\mathbf{t}_d = (0.096, 0, -0.096)$, as in a case discussed here. Therefore, a more general approach to the LTD problem is needed.

Following the theory described in a previous study (Wang, Kamtekar *et al.*, 2005), the phase shift between the two crystallographic origins introduced by the LTD is $\exp(2\pi i\mathbf{h}\mathbf{t}_d)$, where $\mathbf{h}$ and $\mathbf{t}_d$ are the reciprocal and translocation vectors, respectively. Let $F_o(\mathbf{h})$ be the unit structure factor and κ the frequency of translocation (the volume contributions of the two lattices will thus be κ and $1-\kappa$); the total structure factor $\mathbf{F}_{total}$ for a crystal with an LTD can then be formulated as

$$\mathbf{F}_{total}(\mathbf{h}) = \mathbf{F}_o(\mathbf{h})[(1-\kappa) + \kappa\exp(2\pi i\mathbf{h}\mathbf{t}_d)]. \quad (1)$$

The interference resulting from the phase shift leads to total observed intensities that are related to the intensities of a single layer (or single unit lattice) by the following equation, where f is the factor for undesirable intensity modulation and $1/f$ is the correction factor to be applied to observed data to remove the undesirable modulation,

$$I_{total}(\mathbf{h}) = [(2\kappa^2 - 2\kappa + 1) + 2\kappa(1-\kappa)\cos(2\pi\mathbf{h}\mathbf{t}_d)]I_o = fI_o. \quad (2)$$

Using this formula with the derived global parameters κ and $\mathbf{t}_d$ from native Patterson maps, it was possible to solve the structures of φ29 DNA polymerase and the HslV–HslU complex from crystals suffering from LTDs (Kamtekar *et al.*, 2004; Wang, Kamtekar *et al.*, 2005; Wang, Rho *et al.*, 2005). We applied (2) for intensity correction in most cases of φ29 DNA polymerase (Wang, Kamtekar *et al.*, 2005) because there was a single non-origin peak in native Patterson maps with $\mathbf{t}_d = (0, 0, 1/2)$. This method has been successfully applied to solve other LTD-affected structures such as SARS S1 receptor-binding domain in complex with a neutralizing antibody, the bacterial carboxysome shell protein CcmL and the 1918 H1N1 influenza neuraminidase (Hwang *et al.*, 2006; Tanaka *et al.*, 2008; Zhu *et al.*, 2008).

In one case of φ29 DNA polymerase (Wang, Kamtekar *et al.*, 2005), there were two non-origin peaks related by inversion symmetry, with $\mathbf{t}_d = (0, 0, 0.4735)$ or $\mathbf{t}_d = (0, 0, 0.5265)$. (2) is still applicable for intensity correction because this formula is a cosine function and the modulation factor f is independent of the choice of either of the two vectors $\mathbf{t}_d$ or is the same as the averaged value from using both vectors. We note that the summation of the modulation contributions from the two inversion symmetry-related vectors in this case must be made by intensity addition. If the summation is made by complex structure-factor addition, two imaginary components directly cancel out and the length of the resulting vector doubles. Such predicted modulation failed to explain the periodicity of the observed intensity modulation or to remove the observed modulation.

Here, we report another example of LTD in crystals containing the N-terminal and catalytic core domains of maedi-visna virus (MVV) integrase ($IN_{NTD+CCD}$) in complex

Table 1
Data-collection and refinement statistics for crystal form 2.

(*a*) Data processing. Values in parentheses are for the highest resolution shell.

	Data set 1 (κ = 0.22)	Data set 2 (κ = 0.17)
Space group	$P2_1$	$P2_1$
Unit-cell parameters (Å, °)	a = 103.1, b = 83.4, c = 115.5, β = 101.8	a = 102.9, b = 83.2, c = 115.3, β = 102.0
Resolution (Å)	40–2.6 (2.74–2.6)	40–2.64 (2.71–2.64)
R_{merge} (%)	14.1 (52.8)	10.2 (58.6)
Multiplicity	2.8 (2.9)	3.4 (3.3)
$I/\sigma(I)$	5.0 (2.0)	8.1 (2.1)
Completeness (%)	98.1 (97.8)	99.5 (99.4)

(*b*) Refined structures. Incomplete or final models refined in *REFMAC* (v.5.5.0088) using matched NCS and TLS groups, identical geometric restraints and scaling settings against original or corrected data sets, respectively. Because further model building was not possible prior to data correction, the statistics indicate feasibly achievable outcomes in terms of model-to-data agreement and model quality prior to and following data correction.

	Data set 1 (κ = 0.22)		Data set 2 (κ = 0.17)	
	Original	Corrected	Original	Corrected†
Reflections (work)	52333	52333	50264	50264
Reflections (test)	2940	2940	2827	2827
R_{work} (%)	27.2	22.4	27.0	22.6
R_{free} (%)	30.0	24.5	29.8	25.3
Weighted R_{work} (%)	28.2	23.1	26.8	22.4
Weighted R_{free} (%)	30.8	25.2	30.0	25.2
No. of protein atoms	8631	8625	8454	8625
No. of ligand/ion atoms	4	43	4	43
No. of water molecules	0	110	0	110
R.m.s.d. bonds (Å)	0.018	0.012	0.018	0.013
R.m.s.d. angles (°)	1.82	1.31	1.80	1.41
Average B factor (Å^2)	55.9	53.0	55.9	59.0
Ramachandran plot (%)				
Favored	90.7	97.1	95.8	96.6
Allowed	6.9	2.9	3.5	3.2
Outliers	2.4	0.0	0.7	0.2

† The final model refined against corrected data set 2 was deposited in the Protein Data Bank (PDB code 3hph).

with the integrase-binding domain of lens epithelium-derived growth factor ($LEDGF_{IBD}$) (reviewed in Engelman & Cherepanov, 2008). In this case, the translocation-defect fraction (κ) could not be explicitly derived from the modulation of the observed intensities as was performed in the previous special LTD cases. To treat this problem, we developed a novel method to determine the defect fraction, in which we systematically apply a range of trial values of κ to demodulate the data, using the native Patterson function to gauge the effectiveness of demodulation for each given κ. Critically, this procedure allows us to establish the relationship between the defect fraction and the height of the non-origin peaks observed in Patterson maps.

In the case discussed here and all previous cases, the formula only deals with one single translocation vector for given space groups and their inversion symmetry-related vectors are ignored because they have identical f in (2). When the LTD problem occurs in a high-symmetry space group, the translocation must occur pairwise and the translocation fractions for all symmetry-related $\mathbf{t}_d$ pairs must be identical so that the original crystallographic symmetry is statistically retained. Otherwise, crystallographic symmetry will be reduced to noncrystallographic symmetry. In the case of higher symmetry space groups, the modulation factor f is a weighted average from all N symmetry-related $\mathbf{t}_d$. In this formula,

$$I_{total}(\mathbf{h}) = \frac{1}{N}\sum[(2\kappa^2 - 2\kappa + 1) + 2\kappa(1-\kappa)\cos(2\pi\mathbf{h}\mathbf{t}_d)]I_o = fI_o, \quad (3)$$

where summation is carried out for all symmetry-related $\mathbf{t}_d$, all inversion symmetry-related $\mathbf{t}_d$ may be excluded. In the current implementation, the minimal modulation factor is set to be 0.05 so that the correction factor will not exceed 20. It should be noted that Trame & McKay (2001) have also proposed a similar formula for correcting the LTD problem, which is a specific case of our formulation with κ fixed at 0.5.

3. Experimental procedures for detection of LTD

3.1. Protein expression, purification and crystallization

Details of the expression, purification and crystallization of the MVV $IN_{NTD+CCD}$–$LEDGF_{IBD}$ complex have been described elsewhere (Hare *et al.*, 2009). Briefly, crystal form (CF) 1 grew in the presence of 25–30% Jeffamine M600 (Hampton Research) as the main precipitant. CF2, which was initially identified using microseed matrix screening (D'Arcy *et al.*, 2007), grew in the presence of 0.7–0.9 *M* dibasic ammonium phosphate and 2–5% Jeffamine M600.

3.2. Data collection and processing

Diffraction data were collected at Diamond Light Source (Oxfordshire, England) on undulator beamlines I02 and I04. CF1 diffracted to ∼3.3–3.5 Å resolution and belonged to space group $P2_1$ (unit-cell parameters a = 91.1, b = 148.9, c = 91.1 Å, β = 113.4°); the structure was solved by molecular replacement using the program *MOLREP* (Vagin & Teplyakov, 1997, 2000) with individual search domains of HIV-1 integrase (from PDB entries 2b4j and 1k6y) and LEDGF (PDB entry 2b4j) (Cherepanov *et al.*, 2005; Wang *et al.*, 2001). The final model together with the experimental data was deposited in the RCSB Protein Data Bank with accession code 3hpg. In a bid to visualize details of the protein–protein complex that were not defined in CF1, we identified an additional crystal form CF2 that typically diffracted to ∼2.5–2.9 Å resolution. However, CF2 exhibited a pronounced LTD, which could be characterized following the initial structure determination.

The original data set collected for CF2 (data set 1) revealed the space group to be $P2_1$, with unit-cell parameters a = 102.7, b = 83.0, c = 115.3 Å, β = 101.8°. These data were integrated and merged to 2.6 Å with an overall R_{merge} of 14.1% using *MOSFLM* and *SCALA* (Evans, 1993; Leslie, 1992). Subsequently, a higher quality data set was collected from another crystal that could be processed in *MOSLFM*/*SCALA* or *XDS* (Kabsch, 1993) to 2.64 Å with an R_{merge} of 11.8% and 10.2%, respectively (Table 1). The majority of the diffraction images obtained from CF2 samples exhibited exclusively sharp Bragg reflections, which provided proper profiles for peak integration. However, a segment of images covering ∼50° of the

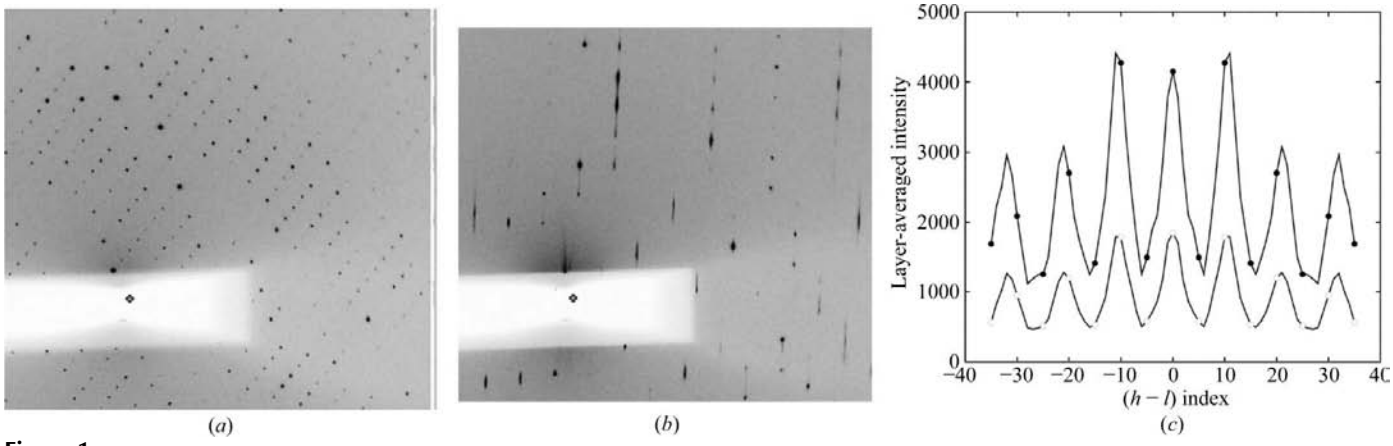

Figure 1
Diffraction data from LTD-affected crystal form 2. (*a*) Diffraction image showing exclusively sharp reflections. (*b*) Diffraction image showing both sharp and diffuse reflections. (*c*) Layer-averaged intensities along the $h - l$ index for data set 1 (white circles) and data set 2 (black circles). The two data sets cannot be scaled prior to correction and are not on the same scale.

spindle-rotation spectrum contained both sharp and streaked reflections (Figs. 1*a* and 1*b*). It was notable that the severity of the diffraction anomalies varied between individual crystals, precluding the indexing and/or processing of data sets collected from the majority (over 90%) of samples. Layer-averaged intensity along the diagonal ($h - l$) axis showed strong intensity modulation, which was as high as fourfold between adjacent layers (Fig. 1*c*). As in previous LTD cases (Bragg & Howells, 1954; Wang, Kamtekar *et al.*, 2005), the data could also be indexed in a larger 'statistically orthorhombic' unit cell of dimensions $a = 137.1$, $b = 169.8$, $c = 83.2$ Å. This type of 'twinning' is in contrast to merohedral (or pseudo-merohedral) twinning, where the cells of apparent higher order symmetry are related to the cell of correct symmetry by adding extra rotational symmetry from twinning operations.

3.3. Initial structure determination and characterization of the LTD in CF2

The CF2 structure was originally solved by molecular replacement in data set 1, using *MOLREP* with the MVV IN CCD dimer (from CF1) as a search model, followed by IBD of LEDGF (from PDB entry 2b4j) and finally MVV IN NTD. A pair of IN dimers were found to form a tetramer with four associated LEDGF chains. Initial refinement using *REFMAC* (Murshudov *et al.*, 1997) and *PHENIX* (Adams *et al.*, 2002) with manual building in *Coot* (Emsley & Cowtan, 2004) resulted in a model with an $R_{\rm work}$ and an $R_{\rm free}$ of 28% and 31%, respectively. Including TLS refinement and using data set 2, the R factors were further reduced to 27% and 30%, respectively (Table 1). Although the statistics were borderline acceptable, the resulting $F_{\rm o} - F_{\rm c}$ maps displayed significant swathes of uninterpretable positive density.

Inspection of the native Patterson maps revealed two non-origin peaks with heights of 22.9 and 5.7% of the origin peak for data set 2 with fractional coordinates of (0.096, 0.000, −0.096) and (0.192, 0.000, −0.192), respectively (Fig. 2). The same peaks were visible for data set 1 but with respective heights of 29.3 and 6.8%. The non-origin peak coordinates suggested a translational symmetry within the asymmetric unit (ASU) with identical structures separated by ∼16 and 32 Å along the $-a/c$ (or $a/-c$) diagonal. Because the tetramer present in the ASU is ∼90 Å across in this direction such a translation is physically impossible, even though a ghost density corresponding to the model shifted by 16 Å could indeed be observed in the $F_{\rm o} - F_{\rm c}$ map (Fig. 3). The features of the Patterson map as well as the density for a shifted structure strongly suggested a case of LTD. Another indicator was the presence of periodic sharp and streaked reflections in the diffraction patterns along the ($h - l$) index direction (Fig. 1).

4. Demodulation of data and final refinement

Using (3), it is possible to demodulate diffraction data from crystals with an LTD, provided the global parameters, the translocation vector ($\mathbf{t}_{\rm d}$) and the translocation frequency (κ), have been determined. The former can be derived from the native Patterson map and in this case $\mathbf{t}_{\rm d}$ is (0.096, 0, −0.096). Previous examples of demodulating data from crystals with LTDs relied on $\mathbf{t}_{\rm d}$ being equal to an integral fraction of a unit-cell parameter (for example, 1/2 or 1/3), which helped in determination of the defect fraction κ (Wang, Kamtekar *et al.*, 2005; Wang, Rho *et al.*, 2005; Hwang *et al.*, 2006). Because such rules do not apply in this case, we estimated the global parameter κ using a trial-and-error procedure (Fig. 4). The data were systematically demodulated with various values of κ and native Patterson functions were calculated for each demodulated data set (Fig. 2); the optimal value of κ was determined based on the flattening of the native Patterson function.

The new demodulation procedure is based on the assumption that the LTD is solely responsible for modulation of the data and for non-origin peaks in native Patterson functions. Proper treatment should minimize the intensity modulation as well as the non-origin peaks in the corrected data. The demodulation function is an inverse of the original modulation function that directly subtracts the contribution from the additional lattice to the observed diffraction intensities. In undercorrected data we expect to see weakening of the non-

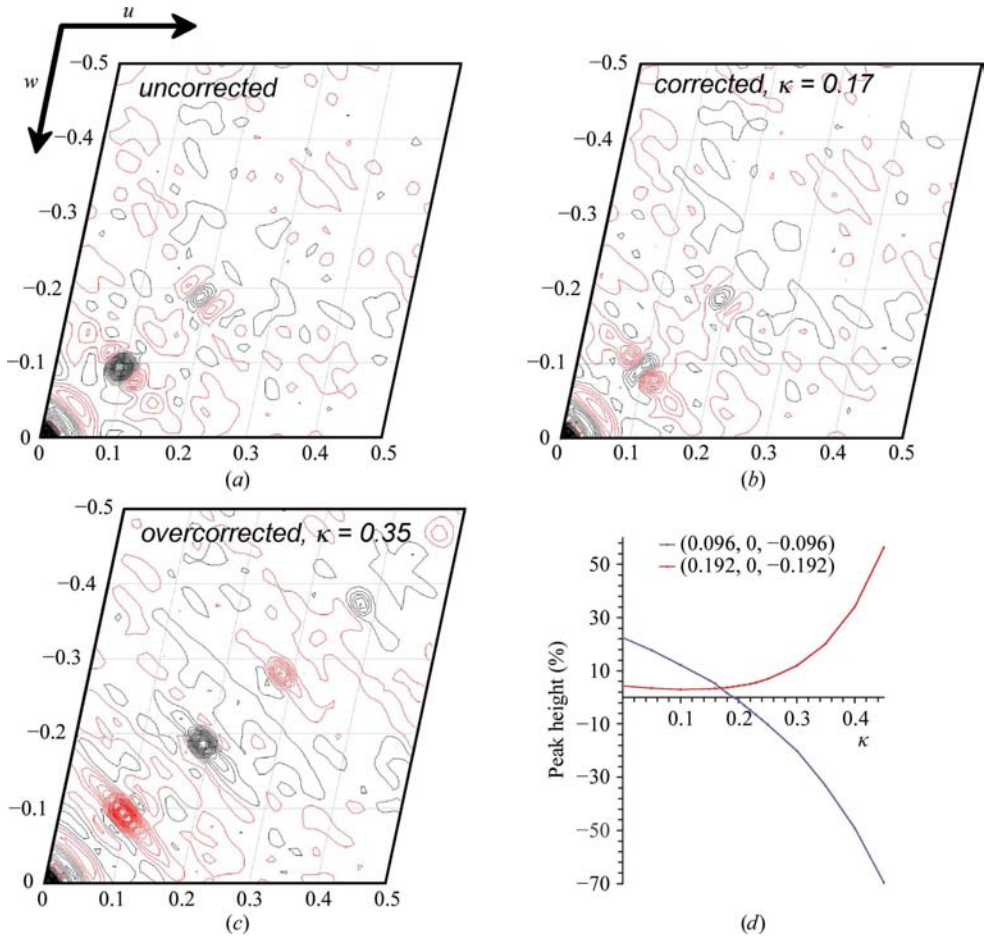

Figure 2
Patterson maps of native and demodulated data from data set 2. (*a*) Native uncorrected Patterson showing major non-origin peaks at (0.096, 0, −0.096) and (0.192, 0, −0192). Positive peaks are shown as black contours and negative peaks are shown as red contours. (*b*) Patterson from corrected data with $\kappa = 0.17$; off-origin peaks are minimized. (*c*) Patterson from overcorrected data ($\kappa = 0.35$); the (0.096, 0, −0.096) peak becomes negative and the (0.192, 0, −0.192) peak increases in height. (*d*) Graph showing the major non-origin Patterson peak heights as a function of κ values.

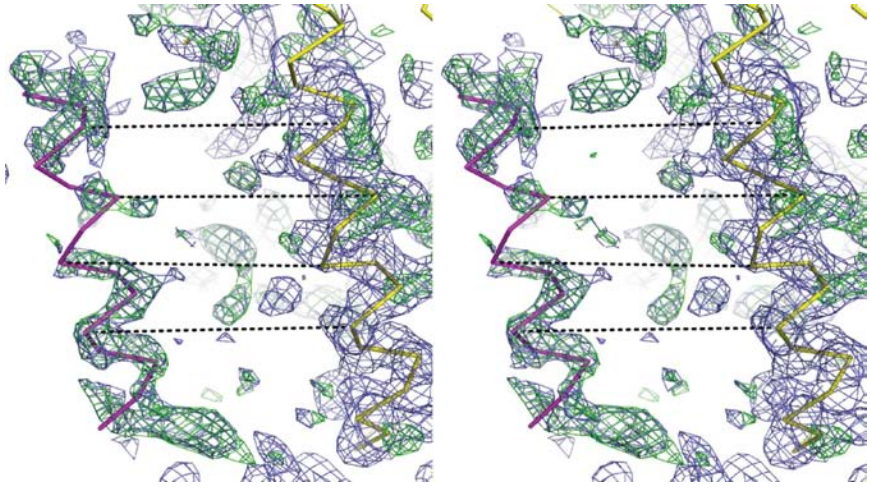

Figure 3
Stereoview of part of the model built into the native data. A helix of the actual model is shown as a yellow ribbon representation and the position of the same helix translocated by 16 Å along the *a*/−*c* diagonal is shown in pink. The black dashed lines represent the translocation vector. $2F_o - F_c$ (purple–blue) and $F_o - F_c$ (green) electron-density maps are shown contoured at 1.0σ and 2.5σ, respectively.

origin peaks in the native Patterson functions, together with some residual modulations, whereas overcorrection would result in an inverted modulation and the appearance of a negative non-origin peak at the same location (Fig. 4).

For data set 2, a value of $\kappa = 0.17$ (*i.e.* assuming that the translocated lattice accounts for 17% of the crystal volume) resulted in the best flattening of the non-origin Patterson peaks. Corrections using higher defect fractions led to inverted modulation along the $(h - l)$ indexes (Fig. 4). In the overcorrected data, the first-order intensity maximum appeared at the $(h - l)$ index where the first-order intensity minimum was located in the uncorrected data. The length of the apparent new translocation vector for the overcorrected data was doubled from that of the original vector in the uncorrected data and the length of the reciprocal vector was halved. Thus, with the overcorrected data we saw two new features in the native Patterson functions: a strong positive peak at $2\Delta\mathbf{t}$ = (0.192, 0.000, −0.192) and a negative peak at $\Delta\mathbf{t}$ = (0.096, 0.000, −0.096).

The percentage amplitude change between the corrected and uncorrected data (*i.e.* the linear cross *R* factor) for data set 2 was 13.5%. As expected, the demodulation led to a significant reduction of the crystallographic *R* factors for the same partially refined model (R_{work}/R_{free} decreased from 27.0/29.8% to 24.2/27.5%); our estimation suggested that the corrections to structure-factor amplitudes directly contributed to the reduction in the *R* factors. Furthermore, the resulting $F_o - F_c$ map was vastly improved, unambiguously allowing the placement of several new amino-acid residues and solvent molecules. Following additional cycles

of building and refinement in *REFMAC* (including TLS refinement), the final model had an R_{work}/R_{free} of 22.6/25.3% and good geometry (Table 1). The coordinates and the corrected data set 2 have been deposited in the RCSB Protein Data Bank (PDB code 3hph). Because the defect fractions in this crystal were relatively small, the interference from doubly translocated layers (that should in part account for the second non-origin peak in the original native Patterson function) could be ignored. For data set 1, the optimum value of κ was found to be ∼0.22, explaining the more pronounced intensity modulation and higher native non-origin Patterson peaks compared with those in data set 2. Interestingly, despite the relatively high R_{merge} value and lower signal-to-noise ratio in data set 1, the final model refined remarkably well against these data following the simple demodulation (Table 1).

5. Structural basis of the LTD in the integrase complex

The structural basis for the lattice translocation immediately became obvious on examination of the partially refined model (Fig. 5*a*). The twofold NCS axis relating the two halves of the ASU is perpendicular to but does not intersect the crystallographic 2_1 screw axis (Fig. 5*b*). As a consequence, symmetry mates related by the 2_1 axis are shifted with respect to each other by ∼8 Å along the $-a/c$ diagonal. The internal symmetry of the ASU allows an alternative packing, *i.e.* an 8 Å

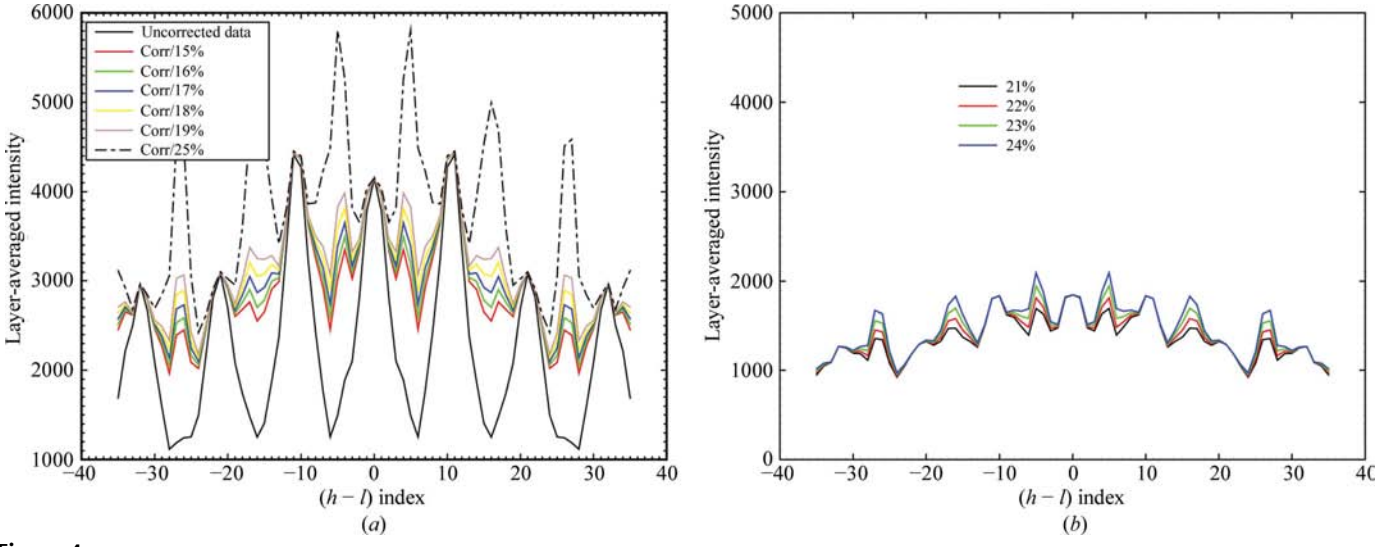

Figure 4
Details of the demodulation process. (*a*) Data set 2 demodulated with defect fractions κ near the correct value of 0.17 and with overcorrected data ($\kappa = 0.25$; dashed line). (*b*) Data set 1 demodulated with defect fractions κ near the correct value of 0.22. This figure is on the same scale as Fig. 1(*c*), with the baseline in (*a*) offset by 1000.

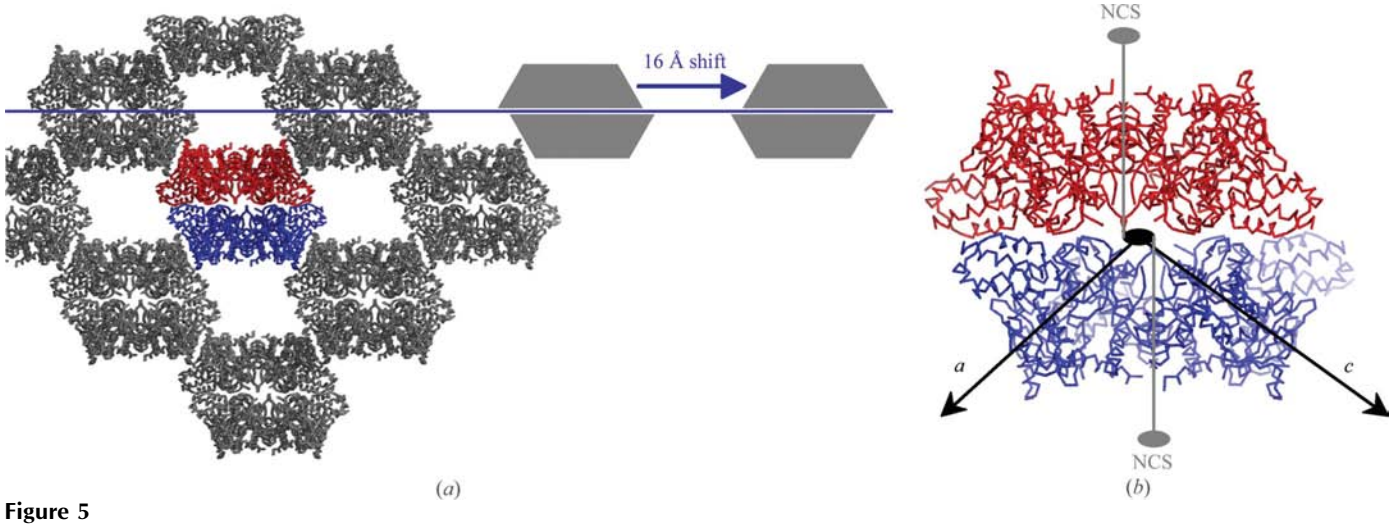

Figure 5
Structural basis of the lattice-translocation defect. (*a*) Molecular packing within crystal form 2, as viewed along the crystallographic 2_1 axis. The asymmetric unit contains a tetramer of IN with four associated LEDGF chains, which pack against each other in layers (dark blue line). Two asymmetric units are highlighted in blue and red. It is possible for another packing arrangement to be produced resulting from a 16 Å shift of one layer with respect to another. (*b*) Closer view of the crystal packing. The blue and red ASUs are related by the crystallographic twofold axis (black oval). The ASU has its own internal twofold symmetry (gray lines) that runs perpendicular to but does not intersect the crystallographic twofold axis.

shift in the opposite direction, resulting in the occasional layer translocation by ~16 Å (Fig. 5*a*). Such a translocation would explain the extra density present in the $F_o - F_c$ map and the (0.096, 0, −0.096) vector observed in the native Patterson map.

6. Corrections for high-order lattice-translocation defects

The crystal packing in CF2 (Fig. 5*a*) does not allow a single layer translocation by $\mathbf{t}_d$ = (0.192, 0, −0.192), which would correspond to a 32 Å shift between two consecutive layers. Minimally, lattices related by a 32 Å shift must be separated by at least one layer or one block. We hypothesize that the secondary non-origin Patterson peak at (0.192, 0, −0.192) can be explained by a rare occurrence of three interfering blocks. Here, if we choose the intervening block to be the reference, the Patterson peaks at (0.192, 0, −0.192) would correspond to combined lattice shifts of $+\mathbf{t}_d$ and $-\mathbf{t}_d$, where the primary $\mathbf{t}_d$ is (0.096, 0, −0.096), with respect to that reference. Each layer can only have one translocation vector of either $+\mathbf{t}_d$ or $-\mathbf{t}_d$, but not both. Otherwise, the layer has no net translocation.

Obviously, the occurrence of a third translocated block with volume fraction ζ is a function of κ. If the two translocated blocks on either side of the reference block have the same sign in $\mathbf{t}_d$, the formula for the three translocated blocks is reduced to the formula of the two translocated blocks as defined by (1) (*i.e.* the addition of three structure-factor vectors from the three blocks is independent of the order of the vectors). Importantly, the existence of multiple interfering blocks may not be visible in the X-ray data if the number of intervening layers exceeds the length of X-ray coherence (a situation that rarely occurs in polytypic structures of small molecules). Thus, in general $\zeta \leq \kappa$. Only when the two translocated blocks have opposite signs ($\mathbf{t}_d$ and $-\mathbf{t}_d$) with respect to the reference block, a high-order interference occurs after modifying (1) and (2) as follows, where $\alpha = 2\pi\mathbf{h}\mathbf{t}_d$,

$$\begin{aligned}\mathbf{F}_{\rm total}(\mathbf{h}) &= \mathbf{F}_o(\mathbf{h})\{(1-\kappa)+\kappa\exp(\alpha)[(1-\zeta)+\zeta\exp(\alpha)]\}\\ &= \mathbf{F}_o(\mathbf{h})[(1-\kappa)+\kappa(1-\zeta)\exp(\alpha)+\kappa\zeta\exp(2\alpha)], \quad (4)\end{aligned}$$

$$I_{\rm total}(\mathbf{h}) = fI_o(\mathbf{h}), \quad (5a)$$

$$\begin{aligned}f = {} & [1-2\kappa(1-\kappa)-2\kappa^2\zeta(1-\zeta)]\\ & + [2\kappa(1-\kappa)(1-\zeta)+2\kappa^2(1-\zeta)\zeta]\cos(\alpha)\\ & + 2\kappa\zeta(1-\kappa)\cos(2\alpha). \quad (5b)\end{aligned}$$

These equations (4 and 5) have two variables, κ and ζ, to be determined. When $\zeta = 0$ these equations return to (1) and (2). Because $\zeta \leq \kappa$, we can estimate the maximal contribution of the higher order interference by assuming $\zeta = \kappa$. The coefficient ratio between the $\exp(2i\alpha)$ and $\exp(i\alpha)$ terms in (4) is $\kappa/(1-\kappa)$, which is relatively small when $\kappa < 0.2$. The coefficient ratio between the $\cos(2\alpha)$ and $\cos(\alpha)$ terms is $\kappa/(1-\kappa+\kappa^2)$, which is even smaller. Using a trial-and-error procedure similar to that described above, we estimated ζ values of ~0.05 and 0.04 for data sets 1 and 2, respectively, which were much smaller than the corresponding κ values. When we applied the correction factor derived from (5) for the two data sets, we observed further flattening of the secondary non-origin Patterson peak at (0.192, 0, −0.192). However, this higher order correction did not lead to an additional significant improvement of the refinement statistics compared with that using the single translocation model based on (1). Thus, at least in this case, the higher order interference appears to be negligible.

7. Prospective remarks

LTD is a relatively common problem that may often have escaped detection. In fact, the LTD problem in CF2 was only discovered following considerable efforts to complete model building and refinement, when ghost densities for a translocated helix were initially noticed (Fig. 3). Prior to the recognition of the LTD, the structure could be determined by molecular replacement and refined using uncorrected data to obtain crystallographic $R_{\rm work}$ and $R_{\rm free}$ values of ~27% and ~30%, respectively, which are borderline acceptable for a correct structure. However, some of the ghost densities had been interpreted as ordered solvent or Jeffamine polymer, which seemed to improve the refinement statistics but did not make physical sense. Hence, identification of the LTD for the protein–protein complex, followed by correction using the new methods described here, significantly improved the quality of the resulting model (Table 1). We believe that this demodulation study highlights a new methodology that could be used for the detection and correction of hidden LTD problems among reported structures whose statistics are borderline acceptable. In previously published structures where LTD was not recognized (Bochtler *et al.*, 2000; Ishikawa *et al.*, 2000; Sousa *et al.*, 2000; Wang, 2001; Wang, Rho *et al.*, 2005), an incorrect quaternary arrangement of the HslU–HslV complex with a disordered interface between HslU and HslV was observed, which was part of the shifted layer–layer structure but with all layer interactions maintained elsewhere.

In principle, layer–layer interactions within a crystal containing an LTD are identical within and between the alternate lattices, perhaps only limited by imprecision of the NCS as in the case of the integrase complex. One should expect a 50:50 distribution of the two lattices in CF2 with no obvious energetic difference between regular layers and the translocated layers (Fig. 5*a*). In practice, however, the observed defect fractions significantly varied between individual crystals; the relative abundances of the alternate lattices are likely to be determined by the direction of crystal growth and asymmetric interactions of layers with the surrounding environment; for example, with cover slips or at the solvent–air interface of crystallization droplets. Thus, correct estimation of the defect fraction is critical to demodulation of twinned data sets because the defect fraction is not always at 50%:50%. We note that demodulation with a fixed κ value of 0.5 as implied from a physical model (Trame & McKay, 2001) leads to over-correction of the data set, which can be seen from the occurrence of large negative peaks in the native Patterson maps calculated from the treated data. In their uncorrected data, the

heights of the non-origin peaks were 10.6% of the origin peak, which was smaller than their theoretical value of 16.7% (or 1/6) for $\kappa = 0.5$. Furthermore, in the remaining data sets from the same work the heights of the non-origin Patterson peaks varied from 9.0% to 15.0%, suggesting that κ was indeed variable between different data sets.

Most previous examples of demodulation of LTD data have relied on the translocation vector being an integral fraction of a unit-cell dimension, so that an explicit method can be derived for the calculation of the defect fraction κ. Alternatively, Tanaka *et al.* (2008) estimated this value empirically by rigid-body refinement of multiple overlapping copies of a partially refined solution with varying occupancy values; the occupancy giving the lowest R factors was used as κ. This multiple packing conformer approach with variable occupancy can explain diffraction data well when there are sufficient observation-to-parameter ratios in high-resolution small RNA structures (Shah & Brunger, 1999). Here, we have described a more straightforward method which is not reliant on integral fraction translocation or even on the possession of a partially refined model. This procedure may allow the detection of many other unrecognized LTD problems without the necessity of examining original diffraction images. Furthermore, it could be possible to use this method in standard crystallographic software and apply it automatically from within a structure-refinement routine, as has already been performed for dealing with cases of merohedral and pseudomerohedral twinning in the current *PHENIX* and *REFMAC* engines (Adams *et al.*, 2002; Murshudov *et al.*, 1997). An automatic comparison of the presence of strong non-origin peaks in the observed native Patterson maps with their absence in calculated native Patterson map from models can help to detect a potentially hidden LTD problem in the data.

References

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1948–1954.

Bochtler, M., Hartmann, C., Song, H. K., Bourenkov, G. P., Bartunik, H. D. & Huber, R. (2000). *Nature (London)*, **403**, 800–805.

Bragg, W. L. & Howells, E. R. (1954). *Acta Cryst.* **7**, 409–411.

Cherepanov, P., Sun, Z. Y., Rahman, S., Maertens, G., Wagner, G. & Engelman, A. (2005). *Nature Struct. Mol. Biol.* **12**, 526–532.

Cochran, W. & Howells, E. R. (1954). *Acta Cryst.* **7**, 412–415.

D'Arcy, A., Villard, F. & Marsh, M. (2007). *Acta Cryst.* D**63**, 550–554.

Doucet, J. & Benoit, J. P. (1987). *Nature (London)*, **325**, 643–647.

Durovic, S. (1992). *International Tables for Crystallography*, Vol. *C*, edited by A. J. C. Wilson, pp. 667–680. Dordrecht: Kluwer Academic Publishers.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Engelman, A. & Cherepanov, P. (2008). *PLoS Pathog.* **4**, e1000046.

Evans, P. R. (1993). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by L. Sawyer, N. Isaacs & S. Bailey, pp. 114–122. Warrington: Daresbury Laboratory.

Glauser, S. & Rossmann, M. G. (1966). *Acta Cryst.* **21**, 175–177.

Hare, S., Di Nunzio, F., Labeja, A., Wang, J., Engelman, A. & Cherepanov, P. (2009). In the press.

Howells, E. R. & Perutz, M. F. (1954). *Proc. R. Soc. London Ser. A*, **225**, 315–329.

Hwang, W. C., Lin, Y., Santelli, E., Sui, J., Jaroszewski, L., Stec, B., Farzan, M., Marasco, W. A. & Liddington, R. C. (2006). *J. Biol. Chem.* **281**, 34610–34616.

Ishikawa, T., Maurizi, M. R., Belnap, D. & Steven, A. C. (2000). *Nature (London)*, **408**, 667–668.

Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.

Kamtekar, S., Berman, A. J., Wang, J., Lazaro, J. M., de Vega, M., Blanco, L., Salas, M. & Steitz, T. A. (2004). *Mol. Cell*, **16**, 609–618.

Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Pickersgill, R. W. (1987). *Acta Cryst.* A**43**, 502–506.

Shah, S. A. & Brunger, A. T. (1999). *J. Mol. Biol.* **285**, 1577–1588.

Sousa, M. C., Trame, C. B., Tsuruta, H., Wilbanks, S. M., Reddy, V. S. & McKay, D. B. (2000). *Cell*, **103**, 633–643.

Tanaka, S., Kerfeld, C. A., Sawaya, M. R., Cai, F., Heinhorst, S., Cannon, G. C. & Yeates, T. O. (2008). *Science*, **319**, 1083–1086.

Trame, C. B. & McKay, D. B. (2001). *Acta Cryst.* D**57**, 1079–1090.

Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.

Vagin, A. & Teplyakov, A. (2000). *Acta Cryst.* D**56**, 1622–1624.

Wang, J. (2001). *J. Struct. Biol.* **134**, 15–24.

Wang, J., Kamtekar, S., Berman, A. J. & Steitz, T. A. (2005). *Acta Cryst.* D**61**, 67–74.

Wang, J. Y., Ling, H., Yang, W. & Craigie, R. (2001). *EMBO J.* **20**, 7333–7343.

Wang, J., Rho, S.-H., Park, H. H. & Eom, S. H. (2005). *Acta Cryst.* D**61**, 932–941.

Zhu, X., Xu, X. & Wilson, I. A. (2008). *Acta Cryst.* D**64**, 843–850.