Acta Crystallographica Section D
**Biological Crystallography**
ISSN 1399-0047

# Space-group and origin ambiguity in macromolecular structures with pseudo-symmetry and its treatment with the program *Zanuda*

**Andrey A. Lebedev[a]* and Michail N. Isupov[b]**

[a]CCP4, Research Complex at Harwell, STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot OX11 0FA, England, and [b]Henry Wellcome Building for Biocatalysis, Biosciences, College of Life and Environmental Sciences, University of Exeter, Stocker Road, Exeter EX4 4QD, England

Correspondence e-mail: andrey.lebedev@stfc.ac.uk

Received 4 April 2014
Accepted 23 June 2014

The presence of pseudo-symmetry in a macromolecular crystal and its interplay with twinning may lead to an incorrect space-group (SG) assignment. Moreover, if the pseudo-symmetry is very close to an exact crystallographic symmetry, the structure can be solved and partially refined in the wrong SG. Typically, in such incorrectly determined structures all or some of the pseudo-symmetry operations are, in effect, taken for crystallographic symmetry operations and *vice versa*. A mistake only becomes apparent when the $R_{\rm free}$ ceases to decrease below 0.39 and further model rebuilding and refinement cannot improve the refinement statistics. If pseudo-symmetry includes pseudo-translation, the uncertainty in SG assignment may be associated with an incorrect choice of origin, as demonstrated by the series of examples provided here. The program *Zanuda* presented in this article was developed for the automation of SG validation. *Zanuda* runs a series of refinements in SGs compatible with the observed unit-cell parameters and chooses the model with the highest symmetry SG from a subset of models that have the best refinement statistics.

## 1. Introduction

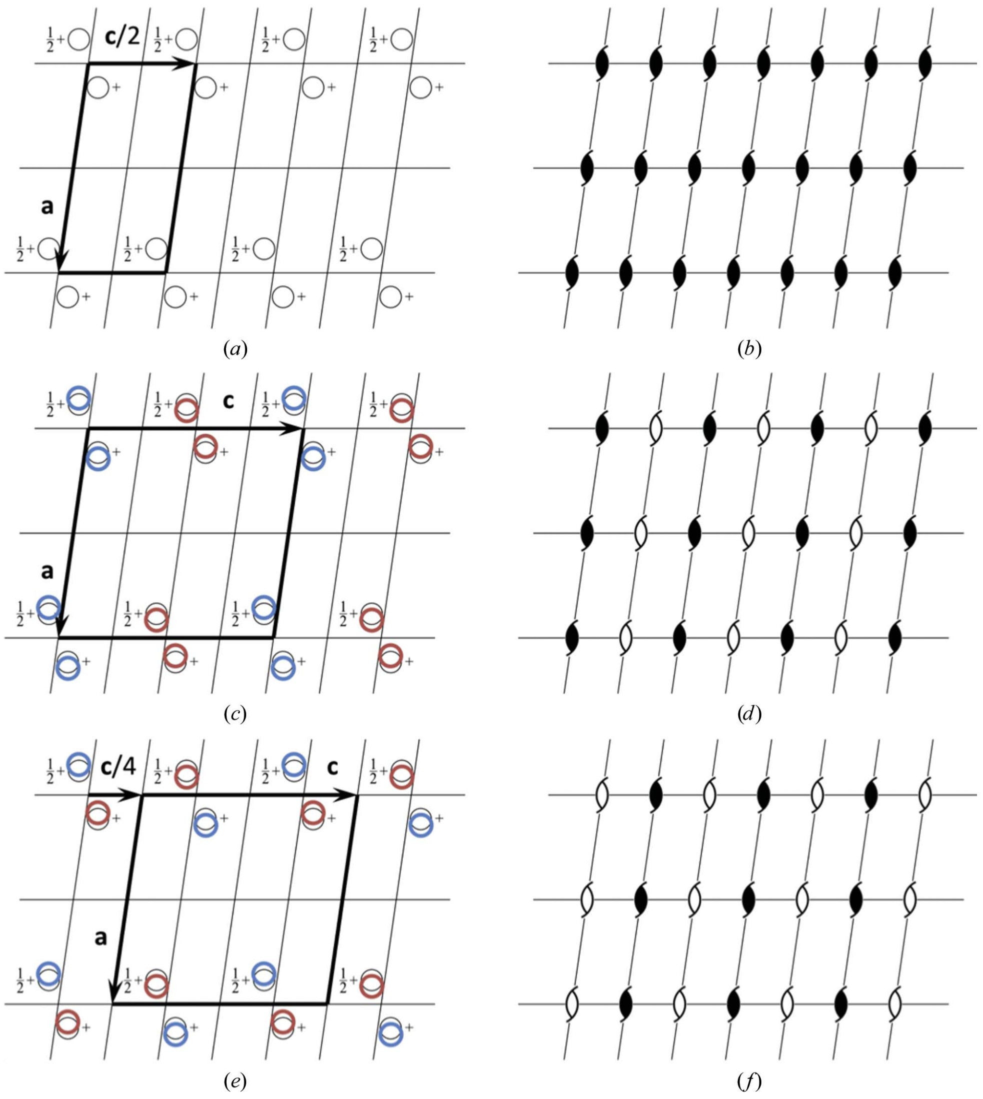
A routine macromolecular structure determination starts from diffraction images and involves several steps, including data integration, data reduction, phasing and refinement. Preliminary unit-cell and point-group symmetry assignment is performed at the stage of data integration (Otwinowski & Minor, 1997; Leslie & Powell, 2007; Kabsch, 2010). However, only the lattice symmetry can be taken into account in this step. Subsequent point-group analysis, scaling and merging steps (Evans, 2006, 2011) provide more accurate definition and may require repeated integration of the data. However, in the presence of twinning by (pseudo)merohedry, a low $R_{\rm merge}$ in the composite symmetry group of the twinned crystal may disguise the point-group symmetry of the individual twin domain. Timely warning can come from twinning tests (*e.g.* the *L*-test; Padilla & Yeates, 2003), but some of these may, in turn, be misleading if the crystal has pseudo-symmetry (Lee *et al.*, 2003). Nonmerohedral twins, as any other case of multiple lattices, should not cause problems for indexing and point-group assignment (Powell *et al.*, 2013), although they may require more attention during data reduction or even at the refinement stage (Rye *et al.*, 2007).

The space group (SG) of a crystal is often assigned at the data-reduction stage on the basis of known point group and axial systematic absences, although enantiomorphic SGs (*e.g.*

$P4_3$ and $P4_1$) cannot be resolved. Other unfortunate possibilities include cases where axial conditions for systematic absences are obscured by the general conditions (*e.g.* $I23$ and $I2_13$), cases where the crystal axis was parallel to the spindle axis of the goniometer and axial reflections were not measured, and crystals where reflections that should have been extinct have significant intensities owing to partial crystal disorder. On the other hand, the apparent systematic absences may be misleading for crystals with pseudo-symmetry. Substantially more sophisticated crystal arrangements are possible in which the SG varies between crystal domains, as in allotwins (Dauter *et al.*, 2005). In limiting cases of partial crystal disorder, the SG may be considered to be undefined and its assignment a matter of convenience (Trame & McKay, 2001; Pletnev *et al.*, 2009).

Experimental phasing (Green *et al.*, 1954; Carter & Sweet, 1997; Vonrhein *et al.*, 2007; Sheldrick, 2010; Skubák & Pannu, 2013) or molecular replacement (MR; Crowther & Blow, 1967; Rossman, 1972; Vagin & Teplyakov, 2000; McCoy *et al.*, 2007) provide the next opportunity for revision of the SG assignment, as the structure solution can be attempted in several candidate SGs that have not been eliminated at an earlier stage. The term SG ambiguity is used sometimes in this context to describe the state of the current knowledge.

In this paper, we provide five examples (Table 1) in which the correct SG was only established at the stage of refinement and model building. The situation to be discussed here is quite common in the presence of pseudo-translation and manifests itself most clearly as a shift of the crystallographic origin from its position in the true structure; therefore, it may be referred to as a pseudo-origin problem. In the course of presenting these examples, we introduce the program *Zanuda* which was written to assist in resolving SG ambiguity at the stage of refinement, particularly in cases when pseudo-origin problems may be encountered.



(*a*) (*b*) (*c*) (*d*) (*e*) (*f*)

**Figure 1**
The pseudo-translation **c**/2 in SG $P2_1$. (*a*, *b*) An approximate structure in which the pseudo-translation **c**/2 acts as a crystallographic translation. This structure belongs to SG $P2_1$ with the basis of lattice vectors (**a**, **b**, **c**/2). Such a structure may result from MR using a reduced data set in which weak reflections $l = 2n + 1$ were ignored. The true structure with the basis of lattice vectors (**a**, **b**, **c**) is not uniquely defined in this case. There are two possible solutions (*c*, *d*) and (*e*, *f*), both belonging to SG $P2_1$. Note that the positions of crystallographic and pseudo-symmetry axes (filled and open shapes, respectively) are swapped in (*d*) and (*f*). Accordingly, the relative positions of symmetry-related atoms, displayed as circles of the same colour in (*c*) and (*e*), are different. In addition, the crystallographic origins differ in (*c*) and (*e*), as illustrated by the positions of the unit cells (thick black lines). This is because, by convention, the origin is located on one of the crystallographic axes. Accordingly, in the first approximation, the crystallographic *x* coordinates of corresponding atoms in (*c*) and (*e*) differ by **c**/4.

## 2. Pseudo-origin solutions

### 2.1. Pseudo-symmetry space group

The crystal SG contains all of the symmetry operations that map the crystal structure onto itself. Each operation defines a rotation and a translation of the crystal such that each atom in the repositioned copy matches a certain atom in the original. Pseudo-symmetry operations are defined similarly, except that the coordinates of matching atoms are not required to coincide exactly. Therefore, it is convenient to define a pseudo-symmetry space group (PSSG) which contains both all of the operations

**Table 1**
Overview of examples.

Key characteristics of symmetry and pseudo-symmetry for the structures discussed in §3. These include the Hermann–Mauguin symbol for the true SG (SG), the number of monomers per asymmetric unit (AU), the Hermann–Mauguin symbol for the PSSG (PSSG), the pseudo-translation vector (PT) and the r.m.s.d. over $C^{\alpha}$ atoms calculated between globally superposed pseudo-origin and true structures (R.m.s.d.). The relative shifts of two structures required for the best superposition are detailed in Tables 2–5 for individual examples.

| Example | PDB code | SG | AU | PSSG | PT | R.m.s.d. (Å) |
|---|---|---|---|---|---|---|
| 1. Monoclinic aminotransferase | 4b9b | $P2_1$; $a = 80.4$, $b = 133.2$, $c = 162.0$ Å, $\beta = 92°$ | 8 | $P2_1$; $a = 80.4$, $b = 133.2$, $c = 81.0$ Å, $\beta = 92°$ | **c**/2 | 1.18 |
| 2a. Orthorhombic aminotransferase–gabaculine complex | 4b98 | $P2_12_12_1$; $a = 119.2$, $b = 192.5$, $c = 77.3$ Å | 4 | $A2_122$; $a = 119.2$, $b = 192.5$, $c = 77.3$ Å | **b**/2 + **c**/2 | 0.45 |
| 2b. Native orthorhombic aminotransferase | 4bq0 | $P2_12_12$; $a = 112.0$, $b = 192.2$, $c = 76.7$ Å | 4 | $A2_122$; $a = 112.0$, $b = 192.2$, $c = 76.7$ Å | **b**/2 + **c**/2 | 0.97 |
| 3. GAF domain of CodY | 2gx5 | $P4_322$; $a = b = 90.2$, $c = 205.6$ Å | 4 | $P4_222$; $a = b = 90.2$, $c = 102.8$ Å | **c**/2 | 1.80 |
| 4. CLEC5A | 2yhf | $P3_1$; $a = b = 109.1$, $c = 84.9$ | 9 | $P3_121$†; $a = b = 63.0$, $c = 84.9$ Å | **a**/3 + 2**b**/3 | 1.24 |

† PSSG shown for the substructure containing chains *A*–*F*.

from the crystal SG and all of the pseudo-symmetry operations.

It is noteworthy that noncrystallographic symmetry (NCS) and pseudo-symmetry are different concepts. An NCS operation is local and is defined by the best overlap of two NCS-related molecules after applying the NCS operation to one of them. In contrast, the pseudo-symmetry operation is global and is defined by the best match between the entire crystal and its transformed copy. Thus, the NCS operation and the pseudo-symmetry operation relating the same two molecules are in general different operations and may coincide only in special cases.

In structures with one molecule per asymmetric unit there is no pseudo-symmetry and the PSSG coincides with the SG of the crystal. In many cases of NCS, such as, for example, in crystals with five identical molecules per asymmetric unit, the global mapping of the crystal onto itself cannot be defined even formally and the PSSG remains equal to the SG of the crystal. Even in the cases when a nontrivial PSSG can be formally defined, the match between the structure and its transformed copy can be too poor to agree with the intuitive perception of pseudo-symmetry. Therefore, dependent on the purpose, a certain threshold may be set on the precision of the operations from the PSSG.

### 2.2. Pseudo-translations and space-group ambiguity

In this article, we discuss structures with pseudo-translations. Notably, the latter term is used by some authors to describe any translational NCS; however, for consistency with the definition of pseudo-symmetry in the previous subsection we will discriminate between the two concepts and assume that operations of pseudo-translation act on the whole crystal and therefore are elements of the PSSG.

Let us consider a structure with SG symmetry $P2_1$ and pseudo-translation vector **c**/2 (Fig. 1, Table 2). The PSSG of this structure is also $P2_1$, but with the basis of lattice vectors (**a**, **b**, **c**/2) (Figs. 1*a* and 1*b*). There are two interesting $P2_1$ subgroups of the PSSG, both having the basis (**a**, **b**, **c**) compatible with the experimentally observed unit-cell parameters. Let the first of these two subgroups be the true SG of the crystal structure (Figs. 1*c* and 1*d*). The second one is then associated with the pseudo-origin structure in which pseudo-symmetry axes are treated as crystallographic axes and *vice versa* (Figs. 1*e* and 1*f*). The two structures are different because different subsets of atoms are related by crystallographic symmetry (note the colour legend in Fig. 1).

**Table 2**
Subgroups of the PSSG for a $P2_1$ structure with the pseudo-translation **c**/2.

The SG Hermann–Mauguin symbol (SG), basis of lattice vectors (Basis), position of the standard origin relative to the standard origin in the true structure (Origin) and references to the panels of Fig. 1 are shown for five subgroups of the PSSG including the PSSG itself (Ref 1). The subgroup (Ref 4) is assumed to be the SG of the true structure. Among an infinite number of possible subgroups of the PSSG, the subgroups shown have either smallest unit cells (Refs 1 and 2) or the same basis of lattice vectors as in the true structure (Refs 3, 4 and 5). The origin positions indicated are the closest ones, among all of the equivalent positions, to the origin in the true structure. The symbol **0** indicates the zero vector.

| Ref | SG | Basis | Origin | |
|---|---|---|---|---|
| 1 | $P1$ | (**a**, **b**, **c**/2) | **0** | — |
| 2 | $P12_11$ | (**a**, **b**, **c**/2) | **0** | Figs. 1(*a*) and 1(*b*) |
| 3 | $P1$ | (**a**, **b**, **c**) | **0** | — |
| 4 | $P12_11$ | (**a**, **b**, **c**) | **0** | Figs. 1(*c*) and 1(*d*) |
| 5 | $P12_11$ | (**a**, **b**, **c**) | **c**/4 | Figs. 1(*e*) and 1(*f*) |

To clarify the concept of pseudo-origin structure, we discuss the following questions. How likely is it for a pseudo-origin structure to emerge as a result of the structure-determination procedure? At what stage does it become clear that the solution is incorrect, and how will the pseudo-origin solution manifest itself? The true and the pseudo-origin structures may be superimposed with an r.m.s.d. of 1 Å, for instance. If refinement starts from a pseudo-origin solution, why does it not converge to the correct structure?

It appears that for a PSSG with an r.m.s.d. in the range 0.4–2 Å the probabilities of obtaining a pseudo-origin MR solution and the true solution are nearly equal. Five examples in §3 fall into this r.m.s.d. range, and for all of them the pseudo-origin structure was the first to be found. Two more cases can be added to this series: Anti-TRAP from *Bacillus licheniformis* (Isupov & Lebedev, 2008; PDB entry 3lcz) and UDP-glucose 4-epimerase from *B. anthracis* (Au *et al.*, 2006; PDB entry 2c20); overall, this amounts to a significant percentage of cases in the authors' experience.

An incorrect origin assignment only becomes apparent when the $R_{\rm free}$ (Brünger, 1992) ceases to decrease below 0.39 or even a higher value, as in the examples below, and no further model rebuilding and refinement can improve it. At this point the electron-density map remains imperfect (breaks in the main-chain electron density, poor solvent peaks) and does not suggest any particular ways of model improvement.

Technically, macromolecular refinement deals with the content of a single asymmetric unit. An equivalent viewpoint is that an infinite crystal is refined, but symmetry-related molecules are kept identical, and their relative positions and orientations are dictated by crystallographic symmetry. As shown in Fig. 1, the subsets of molecules constrained to be identical in the true and the pseudo-origin structures have different configurations. Suppose now that a reference molecule can be moved arbitrarily, and its motion defines, *via* crystallographic symmetry, the motion of all other molecules. In this manner the pseudo-origin structure can be transformed into the true structure, with **c**/4 being the shortest displacement to achieve this. Regrettably, such a shift is far too large for MX refinement, which is a local minimization method.

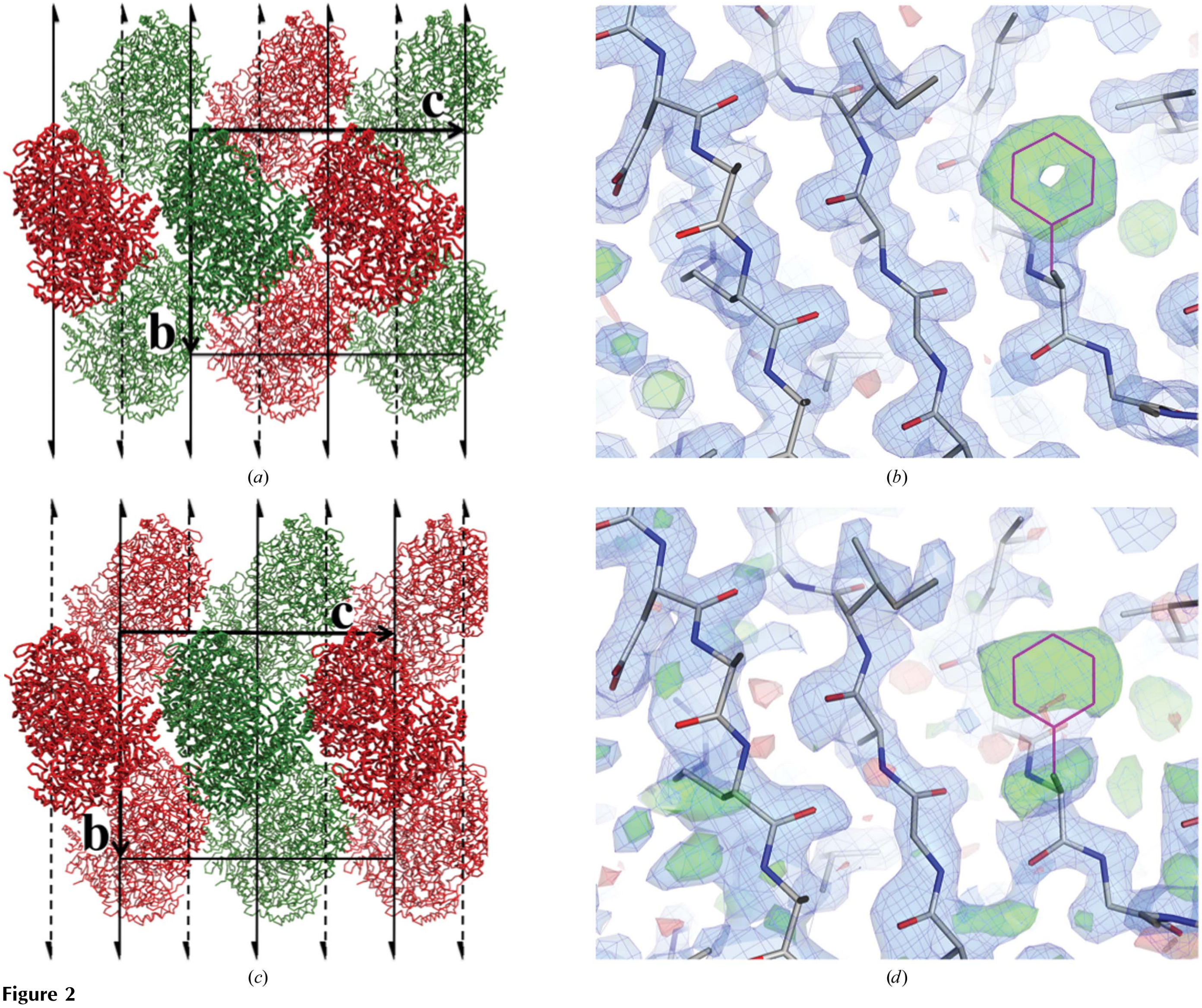


**Figure 2**
Crystal structure of *Pseudomonas* holo AT, an example of a $P2_1$ structure with **c**/2 pseudo-translation. (*a*) The true (PDB entry 4b9b) and (*c*) the pseudo-origin (MR solution) structures of AT correspond to Figs. 1(*c*, *d*) and 1(*e*, *f*), respectively. Crystallographic and pseudo-symmetry axes are shown by solid and dashed black lines, respectively, and the unit cells by rectangles. Tetramers related by crystallographic symmetry are shown in the same colour (red or green). Electron density for (*b*) the true and (*d*) the pseudo-origin structure is shown around residue Phe422 with $2F_{\rm o} - F_{\rm c}$ maps contoured at 1.1$\sigma$ (blue), $F_{\rm o} - F_{\rm c}$ maps contoured at 4.0$\sigma$ for the true structure and 2.5$\sigma$ for the pseudo-origin structure (green) and $F_{\rm o} - F_{\rm c}$ maps contoured at $-2.7\sigma$ for both structures (red). Phe422 side-chain atoms beyond $C^{\beta}$ (magenta lines) were omitted for density calculation. Some parts of the electron density for the pseudo-origin structure closely resemble the corresponding fragment of the true electron density, with the missing Phe422 side chain visible. However, in other locations main-chain density breaks can be observed, with the electron-density maps giving no hints for model improvement. Figs. 2 and 3 were prepared using *PyMOL* (DeLano, 2002).

**Table 3**
Refinements performed by *Zanuda* for the monoclinic AT structure.

The input pseudo-origin $P2_1$ structure was generated from PDB entry 4b9b in two steps: (i) after removal of ligands and solvent the protein molecules were moved into pseudo-origin positions using the 'transform only' *Zanuda* option and (ii) this structure was extensively refined to emulate the original structure-solution process. The transformations of the input model and refinements in subgroups of the PSSG were performed in a single *Zanuda* run. As in Table 2, the subgroups are indicated by their Hermann–Mauguin symbols and relative shift of the crystallographic origin. For each subgroup shown, *Zanuda* performed 24 cycles of *REFMAC*5 rigid-body refinement and eight cycles of restrained refinement. Each refinement series is represented by the r.m.s.d. between the initial and the refined structure and $R_{\rm cryst}$ and $R_{\rm free}$ for the refined structure. A shift of **c**/4 of the origin *versus* the true origin indicates the pseudo-origin structure. Models and maps from the refined true and pseudo-origin $P2_1$ structures were used to generate Fig. 2.

| Hermann– Mauguin symbol | Origin *versus* true origin | R.m.s. shift (Å) | $R_{\rm cryst}$ | $R_{\rm free}$ |
|---|---|---|---|---|
| $P1$ | **0** | 1.16 | 0.263 | 0.324 |
| $P2_1$ | **0** | 1.18 | 0.260 | 0.323 |
| $P2_1$ | **c**/4 | 0.33 | 0.400 | 0.466 |

## 3. Examples

The five examples in this section present cases from the authors' experience in which pseudo-origin solutions were dealt with in the course of structure determination (Table 1). Examples 1, 2a and 2b originate from an aminotransferase project (Sayer *et al.*, 2013). Example 1 is the simplest possible example of a pseudo-origin structure; it illustrates the scheme represented in Fig. 1. Examples 2a and 2b describe two nearly isomorphic structures, such that some crystallographic axes in one become pseudo-symmetry axes in the other and *vice versa*. Examples 3 and 4 are more sophisticated: there is more than one pesudo-origin solution. Example 3 instigated the development of the *Zanuda* program (§4), which was instrumental in the solution of example 4.

### 3.1. Analysis of pseudo-symmetry in the monoclinic aminotransferase

The monoclinic aminotransferase ($P2_1$; PDB entry 4b9b) presents the simplest example of the pseudo-origin problem; the nature of the problem and its solution can be clearly illustrated in a two-dimensional drawing (Figs. 2*a* and 2*c*). The structure was solved by MR using a low-homology model; electron density was visible for the missing side chains, suggesting the correct MR solution. However, the structure did not refine beyond an $R$ factor of 0.49. As the model contained nearly 3400 residues, a significant effort had to be put into model rebuilding before the pseudo-origin problem became apparent and was solved by repositioning of the whole model.

**3.1.1. Structure solution**. The *Pseudomonas aeruginosa* $\beta$-alanine:pyruvate aminotransferase (AT) and its complexes were extensively studied at Exeter University (Sayer *et al.*, 2013). The native protein crystallized in SG $P2_1$ with unit-cell parameters $a$ = 80.4, $b$ = 133.2, $c$ = 162.0 Å, $\beta$ = 92°; the asymmetric unit contained two tetrameric molecules. The native Patterson synthesis of AT calculated at 3 Å resolution contained a pseudo-translation peak with a height of 35% of the origin peak at (0, 0, 0.5), which indicated the presence of a pseudo-translation **c**/2 relating the two tetramers.

The initial MR solution was obtained using *MOLREP* (Vagin & Teplyakov, 2010) and a dimeric model of a related AT from *Chromobacterium violaceum* (Sayer *et al.*, 2013; PDB entry 4ah3) which shared 30% sequence identity with the target. Four dimers were positioned to form two tetrameric AT molecules with a correlation coefficient (Vagin & Teplyakov, 2000) of 0.419 at 4 Å resolution. As with the choice of the crystallographic origin, the choice between the true origin and the pseudo-origin is made when the first copy of the search model is positioned. In our case, the two top translation-function peaks for the first dimer had nearly equal correlation coefficients and therefore this choice became essentially random. As a result, the MR solution proved to be a pseudo-origin solution (Fig. 2*a*; compare with the true structure in Fig. 2*c*). The pseudo-origin problem was noticed and dealt with later, when the refinement statistics did not improve after a few rounds of model rebuilding.

**3.1.2. Structure correction**. *REFMAC*5 (Murshudov *et al.*, 2011) was used for both rigid-body refinement of the MR solution at 15–4 Å resolution and subsequent restrained refinement. The phases obtained by eightfold NCS averaging using *DM* (Cowtan, 2010) were further used for *REFMAC*5 refinement with external phases input (Pannu *et al.*, 1998) and the improved maps were used for model rebuilding with *Coot* (Emsley *et al.*, 2010). This resulted in a significant decrease in $R_{\rm cryst}/R_{\rm free}$ from 0.72/0.72 to 0.44/0.49 at 1.8 Å resolution. A very high starting $R$ factor is a common feature of MR solutions in the presence of pseudo-translation. The substantial drop in $R_{\rm free}$ is rather indicative of a correct MR solution. However, the $R_{\rm free}$ of 0.49 was the best value that could be achieved, and the quality of the maps ceased to improve even after this extensive rebuilding and refinement.

In fact, the electron-density maps were good enough to adjust the conformation of some loops and to assign side-chain rotamers for most of the amino acids that differed between the model and the target structure (Fig. 2*d*). However, there were breaks in the main-chain density and poor density for some side chains and for the solvent. Even in the regions where the electron density fitted the model well, many uninterpretable additional features were present. Therefore, the pseudo-origin solution was suspected to be the problem and two actions were carried out: (i) by applying crystallographic symmetry operations to individual dimers the model was rearranged in such a way that it consisted of two tetramers related by pseudo-translation and (ii) the rearranged model was translated by **c**/4. The corrected structure refined to $R_{\rm cryst}/R_{\rm free}$ of 0.39/0.44 before any manual rebuilding. The model was subsequently improved and refined to $R_{\rm cryst}/R_{\rm free}$ of 0.18/0.22 at 1.7 Å resolution (Sayer *et al.*, 2013; Fig. 2*b*).

Table 3 presents a test run of *Zanuda* with this example. It shows statistics of refinements in the relevant subgroups of the PSSG. For refinement in $P1$, the input $P2_1$ model was expanded by the addition of a symmetry-related copy. One of the two $P2_1$ refinements did not require any rearrangements of the input model, while the other was preceded by rearrangements equivalent to those described above. All transformations of the models were performed automatically. Low $R$ factors for the $P1$ refinements indicate that the molecules have restored their correct relative positions despite the correct symmetry constraints not being reinforced. However, further actions would be required for the transformation of this refined $P1$ model into the correct $P2_1$ model; these would include reduction to the new asymmetric unit and a shift of $\mathbf{c}/4$. There are examples (see, for example, Fig. 2 in Lebedev & Isupov, 2012) in which refinement in $P1$ does not work as expected. The use of correct symmetry constraints (here refinement in the true SG $P2_1$) increases the chances of a refinement program converging to the correct global minimum and of a consequent drop in the $R$ factors. Therefore, the algorithm implemented in *Zanuda* includes, as an intermediate step, independent refinements in all of the subgroups of the PSSG with the basis of the lattice vectors matching the experimentally determined unit-cell parameters. Further details of the *Zanuda* algorithm are given in §4.2.

### 3.2. Structures of two orthorhombic AT crystal forms

Several more AT structures were subsequently analysed, including the gabaculine–AT complex (PDB entry 4b98; Sayer *et al.*, 2013), which crystallized in SG $P2_12_12_1$ with unit-cell parameters $a$ = 119.2, $b$ = 192.5, $c$ = 77.3 Å. Another, more recently characterized, crystal form of native AT (PDB entry 4bq0)[1] has a similar orthorhombic cell with unit-cell parameters $a$ = 112.0, $b$ = 192.2, $c$ = 76.7 Å; however, its SG is $P2_12_12$.

**3.2.1. Cause of space-group ambiguity**. For both crystal forms, the data were merged in point group 222 and systematic absences were observed along all coordinate axes. The native Patterson synthesis calculated at 3 Å resolution contained a strong pseudo-translation peak at (0, 0.5, 0.5) with a height of 71% of the origin peak for the gabaculine complex and 44% for the native enzyme.

The pseudo-translation vector $(\mathbf{b} + \mathbf{c})/2$, which is evident from the Patterson map, and crystallographic twofold axes along $\mathbf{b}$ and $\mathbf{c}$ generate parallel pseudo-symmetry twofold axes. However, because the pseudo-translation is a diagonal translation, the generated axes are screw axes if the crystallographic axes are proper axes and *vice versa*. Therefore, the PSSG does not depend on whether the crystallographic axes along $\mathbf{b}$ and $\mathbf{c}$ are screw or proper axes and, in the crystal settings under consideration, it is either $A222$ or $A2_122$. (The alternative settings for SGs $C222$ and $C222_1$ are used for consistency with the standard setting of the $P2_12_12$ holoenzyme crystal with $a$ = 112.0, $b$ = 192.2, $c$ = 76.7 Å.) Given systematic absences along $\mathbf{a}^*$, there was no reason to doubt the twofold crystallographic screw axis along $\mathbf{a}$, and the PSSG of both crystals is therefore $A2_122$. However, the observed systematic absences along $\mathbf{b}$ and $\mathbf{c}$ do not necessarily mean that screw twofold axes in these directions are crystallographic axes; these absences can be pseudo-absences caused by pseudosymmetic screw twofold axes.

With this analysis it is clear what kind of problem could be expected (and was indeed encountered) in the course of the MR structure determination. Here, the true structure and a pseudo-origin solution, in which the crystallographic axes are misinterpreted as pseudo-symmetry axes, differ by the type of axes along $\mathbf{c}$ and $\mathbf{b}$. From a practical point of view, the situation is a little simpler in comparison to the previous example (and the two further examples), as the alternative solutions are unambiguously specified by the Hermann–Mauguin symbol of the SG and the choice has to be made from $P2_122$, $P2_122_1$, $P2_12_12$ and $P2_12_12_1$, which is quite a common situation in MR structure determination. The difference from a routine case is that prominent MR solutions could be expected for all SGs in this set.

**3.2.2. Structure solution**. Both orthorhombic AT structures were solved by MR with *MOLREP* using data in the resolution range 20–3 Å and the tetrameric AT structure from §3.1 as a search model. The rotation search for both cases gave clear solutions.

For the gabaculine complex, the translational search resulted in high-contrast solutions in all four candidate SGs, with the two top correlation coefficients being nearly identical at 0.622 and 0.629. These were obtained in SGs $P2_12_12$ and $P2_12_12_1$, respectively. Subsequent refinement favoured the second SG; after 60 cycles of restrained refinement with *REFMAC*5 at 1.65 Å resolution the $R_{\rm free}$ converged to 0.394 for the $P2_12_12$ structure and to 0.311 for the $P2_12_12_1$ structure. For the native AT the translational search in SGs $P2_12_12$ and $P2_12_12_1$ also gave close correlation coefficients of 0.612 and 0.608, respectively. The $R_{\rm free}$ difference was larger in this case: the MR solutions refined to an $R_{\rm free}$ of 0.327 in the true SG $P2_12_12$ and 0.457 in $P2_12_12_1$ at a resolution of 1.8 Å. This native structure was eventually refined to an $R_{\rm free}$ of 0.276. The equivalent cross-sections of the two crystal structures are shown in Fig. 3.

Because the Patterson peak corresponding to pseudo-translation $(\mathbf{b} + \mathbf{c})/2$ was so strong (71% of the origin peak) for the gabaculine complex, we could not completely exclude the possibility that this peak corresponded to the true crystallographic translation and that the SG was actually $A2_122$, with half of the measured reflections being merely noise. The program *REINDEX* from the *CCP*4 program suite (Winn *et al.*, 2011) was used to change the crystal setting from $A2_122$ to the conventional $C222_1$ ($a$ = 77.3, $b$ = 192.5, $c$ = 119.2 Å) and to exclude reflections with $h + k = 2n + 1$ (in the new setting). The MR solution found in this SG contained two monomers and refined to an $R_{\rm free}$ of 0.343 at 1.65 Å resolution. This figure looks comparable to the $R_{\rm free}$ of 0.311 for the SG $P2_12_12_1$ observed earlier. However, if SG $C222_1$ were the true space

[1] This structure has not been previously described elsewhere. Crystals were grown by the microbatch method from 10 mg ml$^{-1}$ protein solution containing 20% PEG 3000, 100 m$M$ NaCl, 50 µ$M$ PLP, 100 m$M$ citrate at pH 5.5 and 20 m$M$ of the amino-group acceptor substrate pyruvate. Diffraction data for this crystal form were collected at 100 K using a PILATUS detector on Diamond Light Source beamline I24. The data were processed using *XDS* (Kabsch, 2010) through the *xia*2 pipeline (Winter, 2010). The presence of citrate in the crystallization solution resulted in sequestering of Ca$^{2+}$ ions from the interface of the catalytic dimers, which were thought to be important for tetramer stability (Sayer *et al.*, 2013). However, the AT retained its tetrameric structure in this crystal form.

group and the excluded reflections were merely noise, the $R_{\rm free}$ obtained in SG $P2_12_12_1$ would have been significantly higher than that in $C222_1$. Therefore, for the gabaculine complex the subsequent model refinement and rebuilding was carried out in SG $P2_12_12_1$ with an $R_{\rm free}$ of 0.260 for the refined structure.

### 3.3. Structure solution of the GAF domain of CodY

The structure of the dimeric GAF domain of CodY was originally solved in complex with isoleucine (PDB entry 2b18; Levdikov *et al.*, 2006). This model was used to solve the nonligated structure (Levdikov *et al.*, 2009; PDB entry 2gx5). The GAF domain is a dimer in both solution and in the crystal, with the interface formed by the basal three $\alpha$-helical bundle contributed by each subunit. MR was complicated by substantial conformational changes of both the monomer and the dimer upon ligand binding and by space-group and origin ambiguity. Here, we focus on the pseudo-symmetry of the nonligated structure and describe several approaches to the SG assignment.

**3.3.1. Structure and pseudo-symmetry**. The nonligated GAF domain of CodY crystallizes in SG $P4_322$ with unit-cell parameters $a = b = 90.2$, $c = 205.6$ Å; data were collected to 1.74 Å resolution (Levdikov *et al.*, 2009). The crystal structure had translational pseudo-symmetry with translation vector **c**/2 and an r.m.s.d. of 1.8 Å over matching C$^\alpha$ atoms. The asymmetric unit contained four subunits.

The GAF-domain structure is presented in Fig. 4(*a*). The crystal is formed by cylindrical assemblies of molecules spanning the whole crystal in the **c** direction. The approximate symmetry of a single cylinder includes an eightfold screw axis along **c** and twofold axes perpendicular to it. One quarter of all symmetry operations of the cylinder are crystallographic operations in the three-dimensional crystal.

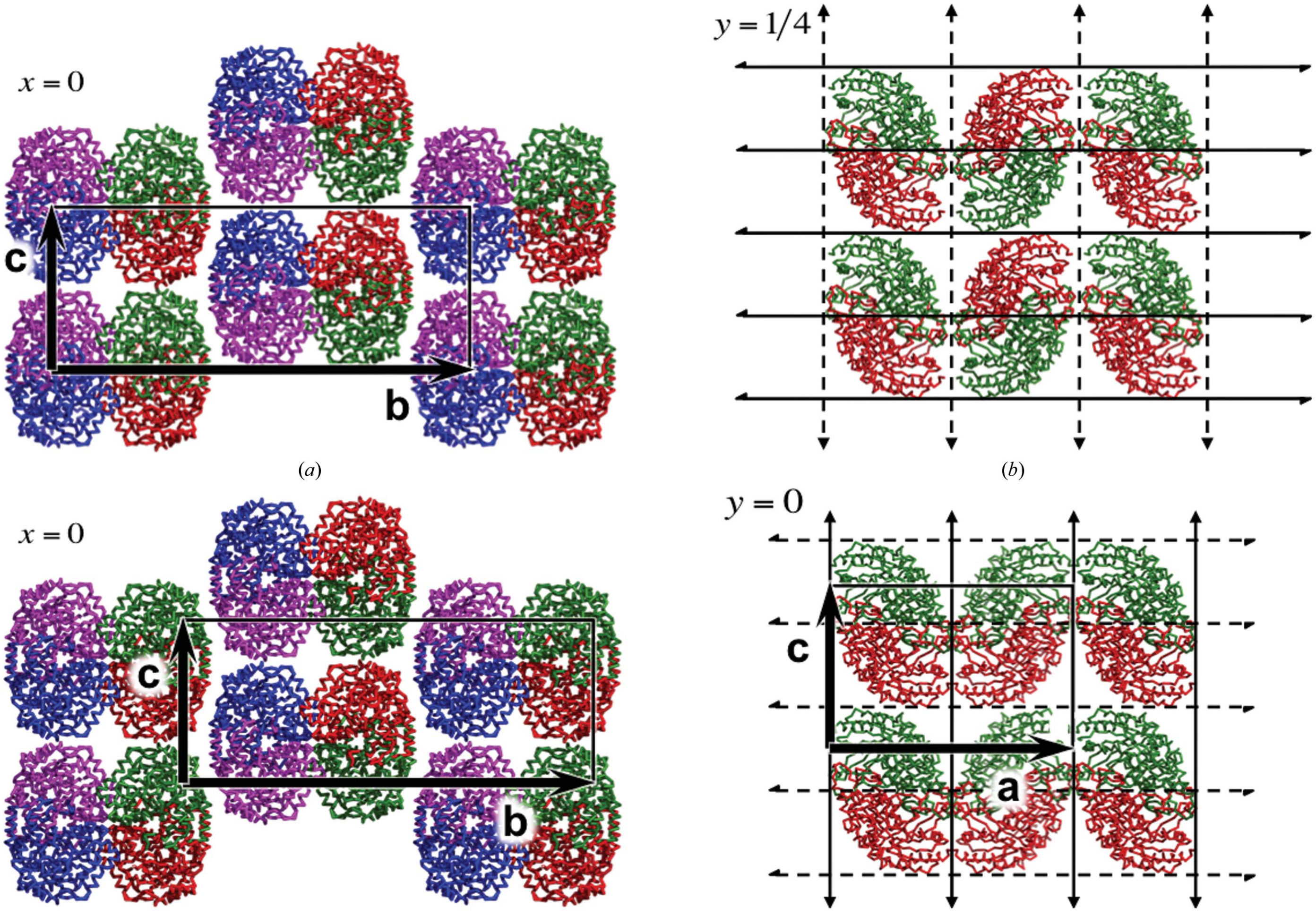


**Figure 3**
Organization of two orthorhombic AT crystals. C$^\alpha$ traces show the packing in (*a*, *b*) the gabaculine complex ($P2_12_12_1$ crystal form) and (*c*, *d*) native AT ($P2_12_12$ crystal form). The unit cells in (*a*), (*c*) and (*d*) are shown as boxes with the basis lattice vectors represented by thick lines and arrows. Symmetry-related monomers are in the same colour. Two orthogonal views are given for each crystal form, which demonstrate their close similarity. Both SGs are subgroups of $A2_122$ (alternative setting of $C222_1$) with the crystallographic axes (solid lines) and pseudo-symmetry axes (dashed lines) swapped between them in the corresponding planes orthogonal to **b***, as shown in (*b*) and (*d*). The difference in the crystallographic and pseudo-symmetry axes results in a different position of the standard crystallographic origin relative to corresponding fragments of the two structures, as shown by the position of the unit cells in (*a*) and (*c*). The unit cell is omitted in (*b*) to highlight that the crystallographic origin is not in the plane shown. Besides, the two crystals have somewhat dissimilar unit-cell parameters, with a maximum difference of 7 Å in the *a* parameter.

Fig. 4($b$) shows two neighbouring slices of a single cylinder, such that each slice contains a pair of biological dimers residing on the same pseudo-symmetry twofold axis. The two dimers are related by the crystallographic twofold axis in the plane of the drawing (and by another pseudo-symmetry twofold axis which is perpendicular to the plane of the drawing). The adjacent pairs of dimers are rotated by 45° relative to each other. Thus, the crystallographic axis makes a half-turn by the fifth pair; therefore, the first and the fifth pairs are related by a pseudo-translation of **c**/2 and eight pairs of dimers span the unit cell.

Exchange of the crystallographic nature of the axes in the bottom drawing of Fig. 4($b$), in which the crystallographic axes become pseudo-symmetric and *vice versa*, would result in a different structure, which is shown in Fig. 4($e$). The latter structure, however, would have the same unit-cell parameters and PSSG as the original structure. All structures related by such permutations of the crystallographic and pseudo-symmetry axes can be enumerated by considering two adjacent pairs of dimers, as the two crystallographic axes relating the subunits in these two pairs (plus the translation **a**) generate the whole SG. Two possibilities for each of the two pairs result in four possible structures belonging to two enantiomorphic SGs $P4_122$ and $P4_322$ (Figs. 4$b$–4$e$). Therefore, the presence of translational pseudo-symmetry in this example creates a potential for three different pseudo-origin MR solutions. Several tests were performed after the true structure had been determined. In particular, Table 4 presents refinement statistics for the true and pseudo-origin structures.

**Table 4**
Refinements of the crystal structure of the CodY GAF domain and three associated pseudo-origin structures belonging to two enantiomorphic SGs.

In each case, reference is made to the Hermann–Mauguin symbol and origin as in Table 2 and the corresponding panel of Fig. 4. To generate starting models, a model with PSSG symmetry ($P4_222$ with halved $c$) was obtained by MR and expanded into the four subgroups of the PSSG shown. Therefore, all four rigid-body refinements started from internally identical models ($R_{\rm cryst}$ of 0.63). The output models from rigid-body refinements were used as input models for the corresponding restrained refinements. Both rigid-body and restrained refinements clearly indicated the correct structure (Fig. 4$b$).

| | | | Rigid-body refinement | Restrained refinement | |
|---|---|---|---|---|---|
| Hermann–Mauguin symbol | Origin *versus* true origin | | $R_{\rm cryst}$ | $R_{\rm cryst}$ | $R_{\rm free}$ |
| $P4_322$ | **0** | Fig. 4($b$) | 0.44 | 0.30 | 0.38 |
| $P4_322$ | **c**/4 | Fig. 4($c$) | 0.52 | 0.40 | 0.50 |
| $P4_122$ | **0** | Fig. 4($d$) | 0.48 | 0.38 | 0.47 |
| $P4_122$ | **c**/4 | Fig. 4($e$) | 0.48 | 0.38 | 0.46 |

**3.3.2. Attempt at structure determination with a dimeric search model**. Search models for the MR were generated from the crystal structure of the CodY GAF domain in complex with isoleucine (PDB emtry 2b18; Levdikov *et al.*, 2006), which formed a crystallographic dimer. When the structure of the nonligated GAF domain was eventually determined, it was

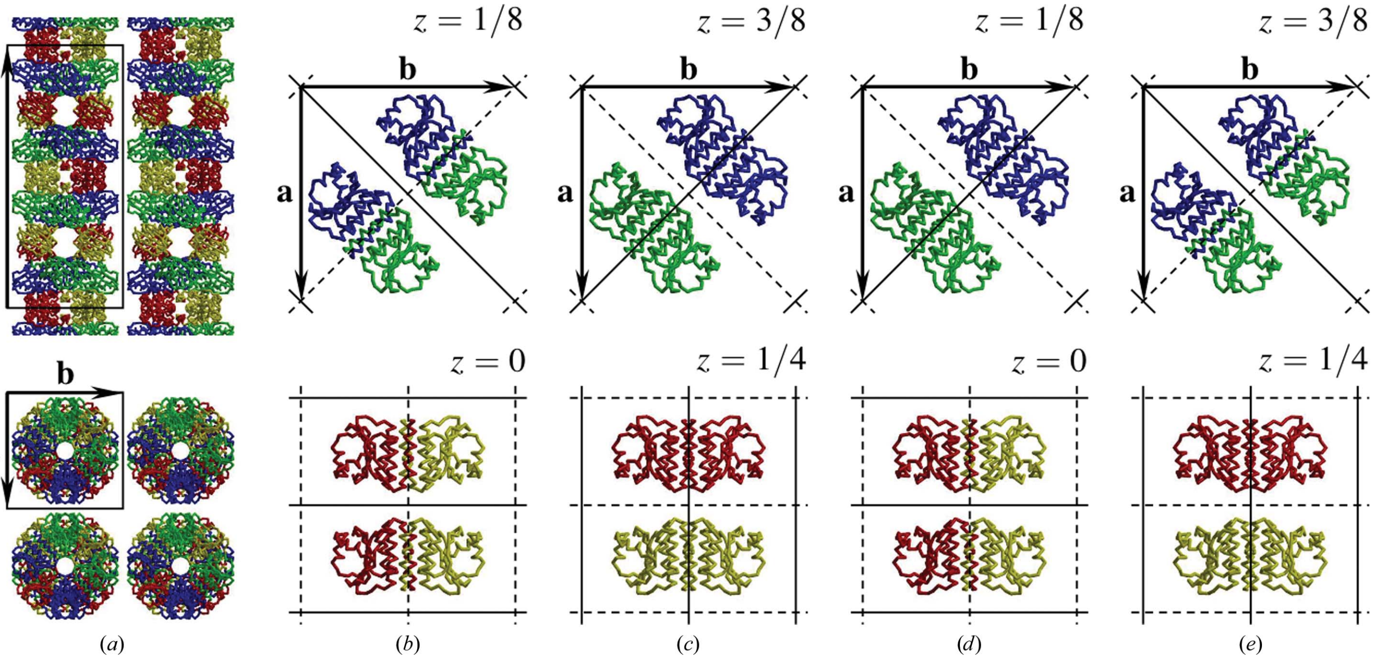


**Figure 4**
Crystal structure of the GAF domain of CodY and associated pseudo-origin structures. ($a$) Overall organization of the crystal. The unit cell is shown in thin black lines. ($b$) Two slices of the molecular cylindrical assembly, with each slice containing two dimers related by the crystallographic twofold axis (solid black lines). In addition, there is a common pseudo-symmetry axis (dashed black lines) relating monomers within these dimers. ($c$, $d$, $e$) Reassignments of crystallographic and pseudo-symmetry axes would result in three possible pseudo-origin structures. In all panels of this figure, the subunits related by crystallographic symmetry are shown in the same colour and the pseudo-translation **c**/2 relates the red substructures to the yellow substructures and the green substructures to the blue substructures. The origin for a given combination of crystallographic axes and consequently the $z$ coordinates of sections shown in ($b$), ($c$), ($d$) and ($e$) are defined by the standard setting of the corresponding SG.

found to contain topologically similar dimers, with the relative orientations of the subunits differing by 14°. As a result, an attempt to solve the crystal structure of the nonligated form using the dimeric model derived from the ligated structure failed.

Interestingly, had the MR search with a dimeric model been successful, the packing constraints would have prevented the positioning of the dimer on a crystallographic axis and the pseudo-origin MR solutions (Figs. 4*c*, 4*d* and 4*e*) would never have occurred. In this scenario the potential problem with the pseudo-origin MR solution would not even be noticed.

In contrast, had the correct configuration been any other than that in Fig. 4(*b*) the use of a dimeric search model would inevitably have led to a pseudo-origin solution. In general, an MR search with an oligomeric model should be used with caution as the asymmetric unit may contain incomplete oligomer(s). Confusion may occur when one of the molecular axes of the oligomeric model and one of the crystallographic proper axes have the same order of rotational symmetry.

**3.3.3. Structure determination with a monomeric search model**. Eventually, MR with a single subunit model was successful, although it was not a trivial task as there were significant conformational differences between the two forms of the protein. Various options of *MOLREP* were tried in both enantiomorphic SGs with different truncated versions of the monomer. One of the MR runs in $P4_122$ resulted in a structure formed by two dimers which were similar to the dimer observed in the known structure. A significant drop in $R_{\rm free}$ in the course of the initial refinement with *REFMAC*5 and interpretable electron density supported this solution. The electron density was good enough to partially rebuild the model. However, the refinement stalled at an $R_{\rm free}$ of 0.46 and validation of the SG assignment was undertaken.

To eliminate any bias towards the pseudo-origin solution, refinement in the PSSG ($P4_222$ with $\mathbf{c}' = \mathbf{c}/2$) was carried out. Experimental data were reindexed with $l' = l/2$. This led to the exclusion of reflections with $l = 2n + 1$ (mainly weak reflections). One of the monomers from the structure refined in $P4_122$ was used as a search model. *MOLREP* was used to position two monomers comprising the asymmetric unit of the $P4_222$ structure with the small cell. In this structure, all of the pseudo-symmetry axes shown in Figs. 4(*b*)–4(*e*) became crystallographic. Therefore, after refinement, this synthetic structure was expected to be equally close to any of the four possible structures with the true unit-cell dimensions. This proved to be an essential step of the protocol.

The $P4_222$ structure (with *c* halved) was expanded into *P*1 with correct unit-cell dimensions and rigid-body refinement was performed at 47–2.7 Å resolution against the original data expanded to *P*1. As the refinement started from the symmetrized model, the initial $R_{\rm cryst}$ was as high as 0.64. The refined *P*1 structure ($R_{\rm cryst}$ = 0.38) was used for the identification of crystallographic axes. The *P*1 model was rotated using *LSQKAB* (Kabsch, 1976) around twofold axes parallel to either *x* or *y* and crossing the *z* axis at either $z = 0$ or $z = 1/4$, and was then visually compared with the original *P*1 model using *Coot*. For two crystallographic axes the overlap of the structure and its copy was visually exact, while discrepancies of about 1 Å were clearly seen for two pseudo-symmetry axes. At this point, the *P*1 refinement has proved to be successful and, in the next step of the procedure, the *P*1 structure was converted to a *P*121 structure and then to a $P222_1$ structure by

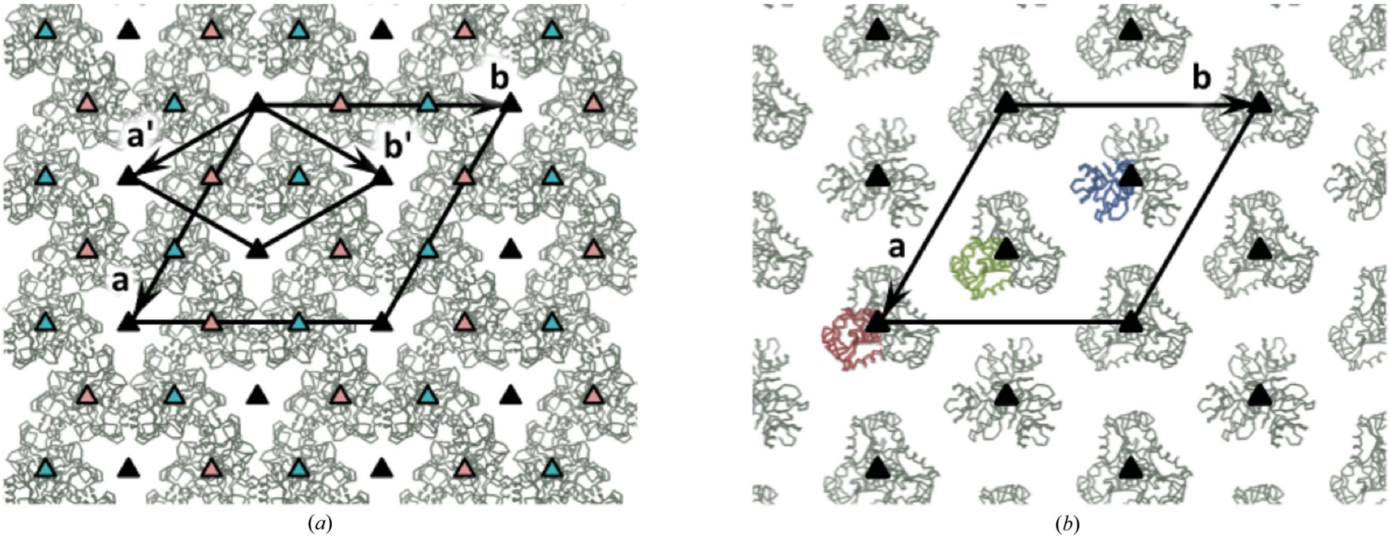


**Figure 5**
Organization of the CLEC5A protein crystal. $C^\alpha$ traces show the crystal packing for (*a*) the large substructure formed by molecules *A*–*F* and their symmetry equivalents and (*b*) the small substructure formed by molecules *H*–*I* and their symmetry equivalents. Crystallographic $3_1$ axes are indicated by black triangles. Two classes of pseudo-symmetry $3_1$ axes are indicated by orange and blue triangles. Crystallographic translations **a** and **b** and pseudo-translations $\mathbf{a}' = (\mathbf{a} - \mathbf{b})/3$ and $\mathbf{b}' = (\mathbf{a} + 2\mathbf{b})/3$ are indicated by arrows. The complete structure belongs to SG $P3_1$. The substructure in (*a*) has pseudo-symmetry $P3_121$ with translation basis $\mathbf{a}'$, $\mathbf{b}'$. In the original MR solution for the large substructure (molecules *A*–*F*) the crystallographic origin coincided with one of the pseudo-symmetry axes. The small substructure (molecules *H*–*I*) is not symmetrical relative to the rotations about the pseudo-symmetry axes and therefore it could not be solved until the position of the origin in the large substructure had been corrected.

**Table 5**
Origin correction for the pseudo-origin partial model (subunits *A*–*F*) of the CLEC5A crystal.

All subgroups of the PSSG shown in the table have experimentally observed unit-cell parameters. In each case, reference is made to the Hermann–Mauguin symbol and origin as in Table 2. Structure transformations and refinements were carried out within a single run of *Zanuda*. For each refinement, the r.m.s.d.s between the initial and the refined structure and the final $R_{\rm cryst}/R_{\rm free}$ are shown. The Hermann–Mauguin symbol $P3_1$ and the vector **0** in the column 'origin *versus* true origin' indicates the true structure. The origin shifts **a**/3 and 2**a**/3 correspond to two pseudo-origin $P3_1$ structures. The input structure, which was a partial MR solution of CLEC5A, had the origin shift 2**a**/3. This solution contained six out of nine molecules in the asymmetric unit and corresponded to Fig. 5(*a*), with the pseudo-symmetry axes shown in blue being incorrectly assigned as crystallographic axes.

| Hermann–Mauguin symbol | Origin *versus* true origin | R.m.s. shift (Å) | $R_{\rm cryst}$ | $R_{\rm free}$ |
|---|---|---|---|---|
| $P1$ | **0** | 1.24 | 0.430 | 0.466 |
| $P3_1$ | **a**/3 | 0.97 | 0.460 | 0.498 |
| $P3_1$ | 2**a**/3 | 1.09 | 0.459 | 0.495 |
| $P3_1$ | **0** | 1.24 | 0.430 | 0.466 |
| $C2$ | **0** | 1.17 | 0.441 | 0.481 |
| $P3_112$ | **0** | 1.20 | 0.455 | 0.480 |

moving it along *z* (to bring the crystallographic axes to their standard positions), changing the SG in the PDB file header and removing redundant copies of monomers. Transformations to candidate $P4_122$ and $P4_322$ structures were performed in a similar way and the latter was chosen because of the nearly exact overlap between redundant copies of monomers. Eventually, the $P4_322$ structure (Figs. 4*a* and 4*b*) was refined to an $R_{\rm cryst}/R_{\rm free}$ of 0.153/0.212 against the complete 1.74 Å resolution data set (Levdikov *et al.*, 2006). Note that the method used here has also ruled out the possibility of lower point-group symmetry and twinning.

### 3.4. Structure of human CLEC5A and its determination

The structure of CLEC5A has been described previously (Watson *et al.*, 2011; PDB entry 2yhf). Here, we focus on the critical steps of structure solution, reassignment of the origin of a substructure and the use of partial structure phases in *MOLREP*.

**3.4.1. Structure**. The complete structure belongs to SG $P3_1$ and can be presented as a combination of two substructures (Fig. 5). The asymmetric unit of the complete structure contains nine subunits; six of them belong to the large substructure (Fig. 5*a*), which has a pseudo-translation **a**/3 + 2**b**/3.

The pseudo-translation and crystallographic $3_1$ axes (filled black triangles in Figs. 5*a* and 5*b*) generate pseudo-symmetry $3_1$ axes in the large substructure (coloured triangles in Fig. 5*a*). In addition, the large substructure has twofold pseudo-symmetry axes running along **a**, **b** and **a** + **b** and therefore the PSSG is $P3_121$ with (**a**′, **b**′, **c**′) = (**a**/3 − **b**/3, **a**/3 + 2**b**/3, **c**). Table 5 shows all of the subgroups of the PSSG with experimentally observed unit-cell parameters (*i.e.* with the basis **a**, **b**, **c**). These include three $P3_1$ subgroups, with origins at 0 (the true SG of the crystal), **a**/3 and 2**a**/3, and with corresponding sets of $3_1$ axes.

The remaining three molecules from the asymmetric unit of the complete structure belong to the small substructure shown in Fig. 5(*b*). The small substructure does not satisfy the definition of pseudo-symmetry used in this article: two of the three molecules forming it are related by translation, while the third molecule has a different orientation. The pseudo-translation in the large substructure and the translational NCS in the small one generate non-origin Patterson peaks with a height of about 0.4 of the origin peaks at 4 Å resolution.

**3.4.2. Twinning**. The presence of partial twinning in the CLEC5A crystal can be established using the *H*-test (Yeates, 1988), with the twinning coefficient estimated to be in the range 0.10–0.15. Such a low fraction of domains with alternative orientation does not normally affect structure solution and refinement. However, a possible morphology of this twin is particularly interesting. The directions of the three equivalent twin axes coincide with the directions of twofold axes in the pseudo $P3_121$ SG to which the large substructure belongs. This suggests that the large substructure is continuous throughout the whole crystal, whereas the orientation of the small substructure varies and defines twin domains. Such an organization of a crystal suggests a high correlation between intensities from twin domains in alternative orientations and, therefore, reduced contrast in perfect twinning tests. This effect could be one of the reasons why the *L*-test (Padilla & Yeates, 2003) using the entire data set failed to produce a clear indication of twinning.

Not only is the large substructure continuous throughout the whole twinned crystal, but its crystallographic $3_1$ axes (black triangles in Fig. 5*a*) also follow the same pattern in the two twin orientations. A different situation is found in the alternative $P3_1$ SGs. The threefold axes in SGs $P3_1$(**a**/3) and $P3_1$(2**a**/3) (orange and blue triangles in Fig. 5*a*) are related by twofold axes from the PSSG which are collinear with the twofold twin axes. Therefore, had the SG $P3_1$(**a**/3) corresponded to the true structure, the SG $P3_1$(2**a**/3) would also represent the true structure: that of another twin individual. Therefore, although there were three alternative SGs with Hermann–Mauguin symbol $P3_1$, they corresponded to only two possible twins.

**3.4.3. Structure solution**. The three molecules *A*, *B* and *C* have very similar orientations and their self-vectors jointly contribute to the same peak of the rotation function (RF). This implies up to a three times higher RF peak compared with the unique orientation, *i.e.* we can say that the multiplicity of this peak equals three. The same applies to molecules *D*, *E* and *F*. Also, molecules *H* and *I* have similar orientations, and the height of the peak for this orientation in the RF is doubled, while orientation of *J* is unique and its RF peak has a multiplicity of one. As a result, the rotation peaks for molecules *H*, *I* and *J* could not be located in the noise and it was not possible to find these molecules by routine MR.

Had the twinning coefficient been closer to 0.5, the heights of RF peaks from dissimilar orientations would have become even more different because of the relation between twinning and pseudo-symmetry discussed above. Molecules *A*, *B*, *C* and *D*′, *E*′, *F*′ (where the primes signify another twin individual) have very close orientations and their joint RF peak would have a multiplicity of six in a perfect twin, whereas the multiplicity of the joint RF peak from $H$ and $I$ (and from $H'$ and $I'$) would remain equal to two and the multiplicity for $J$ (and $J'$) would remain one.

Four of the six monomers representing the two dominating orientations were found by conventional MR (*MOLREP*) and the remaining two were found using an MR search in the electron density calculated from the refined partial model as described in Watson *et al.* (2011). The search included three steps: spherically averaged phased translation function, phased rotation function and phased translation function (SAPTF + PRF + PTF implemented in *MOLREP*; Vagin & Isupov, 2001). Initial manual model correction and refinement indicated that this structure might have been assigned a pseudo-origin.

Correction of the origin was performed using *Zanuda*; refinement statistics in the subgroups of the PSSG are given in Table 5, where subgroups $P3_1(\mathbf{a}/3)$ and $P3_1(0)$ represent the originally assigned and the correct SGs, respectively. After the origin correction the quality of the electron density improved and $R_{\rm free}$ decreased from 0.50 to 0.46. However, the $R_{\rm free}$ remained high and continuous electron density emerged in the solvent region, indicating that the current structure was incomplete.

It is worth noting that this example is rather an exception. The input pseudo-origin model, which refined to an $R_{\rm cryst}$ and $R_{\rm free}$ of as high as 0.46 and 0.50, respectively, was nevertheless good enough for SG correction using *Zanuda*. Usually, such high values of the $R$ factors indicate a completely wrong MR solution or too many model errors, which need to be corrected before the $R$ factors become sufficiently sensitive criteria for distinguishing symmetry and pseudo-symmetry. However, in the example under consideration the model was sufficiently accurate, while the reason for the high $R$ factors was its incompleteness.

At this point, the existence of a small substructure became evident and this was solved by the SAPTF + PRF + PTF method. The use of the phased functions (SAPTF + PRF) for finding the orientations of molecules $G$, $H$ and $I$ was key to solving the small substructure. As we discussed previously, the signal from molecules $G$, $H$ and especially $I$ in the conventional RF was too weak to generate high-rank peaks. As opposed to conventional RF, which is in effect a Patterson search, the orientation search using combination of SAPTF and PRF works with electron-density maps; therefore, it is local and is not affected by dominating orientations. The full model thus built was refined to $R_{\rm cryst}/R_{\rm free}$ of 0.216/0.267 at 1.9 Å resolution (Watson *et al.*, 2011).

## 4. *Zanuda*

### 4.1. Historical perspective

Although advances in X-ray data processing and analysis help to distinguish true twinning from higher point-group symmetry in most cases, there remains a class of structures with strong pseudo-symmetry where both SG and point-group assignment may require comparative refinements in alternative space groups at the stage when the model is nearly complete. Several borderline cases were found during analysis of twinning cases in the PDB (Lebedev *et al.*, 2006). However, at the time automation of this process did not seem sufficiently important because of the low frequency of such marginal cases and the relative simplicity of the procedure involving standard MR and refinement in a couple of candidate SGs.

The first instance of a pseudo-origin structure that we came across was Anti-TRAP (Isupov & Lebedev, 2008; Shevtsov *et al.*, 2010). To our surprise, an apparently correct high-contrast MR solution could not be refined to an $R_{\rm free}$ of better than 0.43. However, subsequent MR using the refined model gave a new solution that easily refined to an $R_{\rm free}$ of 0.26. Comparison of the initial solution and the final refined structure gave us an insight into the problems that can arise in the presence of pseudo-translation and showed that the initial MR search resulted in a wrong solution that was named a pseudo-origin solution. Importantly, this example has shown that refinement of a pseudo-origin solution can be beneficial and can lead to the resolution of SG ambiguity by subsequent MR.

The next example of a pseudo-origin that we encountered, UDP-glucose 4-epimerase (PDB entry 2c20) from the SPINE project carried out in YSBL (Au *et al.*, 2006), had to be approached in a more systematic manner. The structure had a pseudo-translation basis $(\mathbf{a} - \mathbf{b})/3$, $(\mathbf{a} + 2\mathbf{b})/3$, $\mathbf{c}/2$ (the PSSG unit cell was six times smaller than the true unit cell). The initial MR solution was found with a wrong origin and the true SG had to be recovered manually. The procedure started from refinement in $P1$ and involved SG extension by addition of the best symmetry operation at each step followed by the next round of refinement. The protocol used in the CodY example (§3.3) was in effect a simplified version of the protocol used for UDP-glucose 4-epimerase, with intermediate refinements omitted. Both the UDP-glucose 4-epimerase and the CodY structures required a significant amount of time and effort to resolve them; however, it became obvious that many operations were being repeated and that automation is feasible and could be advantageous for future pseudo-origin cases. Thus, based on the protocol used for UDP-glucose 4-epimerase the program *Zanuda* has been developed, which only extends this protocol with one extra step: a preliminary refinement in all candidate SGs.

### 4.2. *Zanuda* protocol

*Zanuda* is a Python script that uses *REFMAC*5 and several *CCP*4 (Winn *et al.*, 2011) programs for handling MTZ files and one purpose-written Fortran program which is used for the determination of the PSSG and for transformations of the data and the models from one subgroup of the PSSG to another. Once the PSSG has been established, the atoms of the input model which lack one or more of their pseudo-symmetry equivalents are removed, so any two pseudo-symmetry-related molecules have the same composition. Transformation of a model from a certain group involves duplication and transformation of individual molecules in order to extend the asymmetric unit of the original group and fill the asymmetric unit of its subgroup; the experimental data are transformed accordingly. The transformation from a certain group into its supergroup (particularly into the PSSG) is carried out by reduction of the asymmetric unit. Individual molecules in this case are first transformed into a new asymmetric unit and then the coordinates of atoms, which should be equivalent in the supergroup, are averaged. The geometrical parameters of the resulting model are distorted; however, these are restored in the course of subsequent refinements.

All calculations pass through three stages. In the first stage the solvent molecules are removed, the PSSG is determined, the pseudo-symmetry-related molecules are modified to have the same composition and, optionally, the starting model is transformed into the PSSG. This option may be useful in certain cases, as discussed in the next subsection (§4.3).

Note that in the preliminary pseudo-symmetry analysis *Zanuda* imposes an upper limit of 3 Å for the $C^{\alpha}$ r.m.s.d. between the structure and its copy generated by a global operation to be included into the PSSG. Global operations with higher values of the r.m.s.d. are ignored as it is very unlikely that they are true crystallographic operations, whichever structure-solution method was used.

At the second stage, a series of refinements are conducted in the subgroups of the PSSG. The unit cell established in the course of data integration is considered to be the true unit cell. Therefore, only subgroups which do not contain extra translations relative to the input SG are taken into consideration. (For example, the PSSG $A2_122$ from the example in §3.2 would not be included in this list if the input model and data belonged to the true SG $P2_12_12_1$.) A model and data for a particular refinement are prepared from the model obtained at the first stage of the *Zanuda* procedure and the original experimental data, respectively, using appropriate transformations. The refinement is conducted in two steps: rigid-body refinement is followed by restrained refinement. This stage increases the chances of escaping from a wrong local minimum, since refinement in *P*1 does not always achieve this if the input model has been initially refined in the wrong SG. In addition, the resulting table comparing the series of refinements in subgroups of the PSSG may be useful on its own.

The model which had the lowest $R_{\rm free}$ after refinement at the second stage is passed to the final third stage. This model and the original data are transformed into *P*1 and undergo several rounds of refinement. After each such round, an attempt is made to extend the current working SG by adding one new operation from the PSSG (and all generated operations), with the r.m.s.d. between the current model and its transformed copy being the selection criterion. The $R_{\rm free}$ value only slightly changes after next round of refinement if the true symmetry operation is added and increases by several percent if the new operation is a pseudo-symmetry operation. Therefore, the procedure terminates when a steep increase in $R_{\rm free}$ is observed (an increase of up to 1% is tolerated) or when all suitable symmetry operations are already included in the current SG.

### 4.3. Possible scenarios

The first scenario, recovery from a pseudo-origin, has already been discussed in detail. One thing to emphasize here, in order to relate this scenario to the other two, is that both the input SG and the true SG have the same point-group symmetry. The technique implemented in *Zanuda*, which involves a series of refinements, is usually successful here and leads to recovery from the pseudo-origin solution. The symmetrizing of the input model by transforming it into the PSSG prior to further manipulations may or may not be beneficial and it is worth trying both options.

Another scenario is realised when the structure has been solved, intentionally or erroneously, in a lower symmetry SG. Sometimes, for example owing to suspected twinning and with a good MR model available, the structure is solved in *P*1 to avoid any initial assumption about the true SG. The asymmetry of such a model is usually trustworthy and, in order to preserve it, the option of symmetrizing the input model in the PSSG must be avoided. Usually, this is an easy case for *Zanuda* and its automatic run will clearly indicate the correct SG. Often in this scenario the true symmetry can be identified immediately from the analysis of the input model, without any refinements. The option of skipping refinements is available from the *Zanuda* task interface included in *CCP4i* (Potterton *et al.*, 2003). This protocol is fast but not automated, so the user has to analyse a table in the log file. The model transformed into the true SG should have (i) a very small r.m.s.d. from the input model and (ii) the highest symmetry among the models satisfying the first criterion.

The most challenging is the opposite scenario, when the currently assigned point group is a supergroup of the true point group. A combination of twinning and pseudo-symmetry, when the twin axis is parallel to the pseudo-symmetry axes, decreases the contrast in twinning tests and therefore can easily lead to such 'over-merging' of the data. Usually in this scenario the current wrong SG coincides with the PSSG, so the option of merging into the PSSG has no effect. An incorrect assignment of an SG corresponding to a higher point group typically results in a deep local minimum from which refinement cannot escape. Therefore, the automatic *Zanuda* run may keep the initial SG, even if it is incorrect. If doubts remain regarding the SG assignment substantial manual work is required and *Zanuda* can be used as an auxiliary tool. One possibility is to disable refinement and run *Zanuda* in the transformation-only mode. The output will consist of models belonging to different SGs. Any or all of these models can be used (i) as a reference for the *POINTLESS* and *AIMLESS* pipeline (Evans & Murshudov, 2013) to generate properly merged data sets in the required point group and (ii) as a starting model for refinement against this data set. Future plans for *Zanuda* include an optional input of unmerged data, with the merging step being performed internally for each point group involved. This will increase the chances of isolating the true structure in such a scenario.

### 4.4. Program usage

Originally, *Zanuda* was designed for the YSBL server at the University of York, England and has been recently moved to the CCP4 server (http://www.ccp4.ac.uk/BALBESSERV/), where it runs in the default mode. *Zanuda* is also included in the *CCP*4 program suite series 6.3 and later. The choice of program options is provided *via* the *CCP*4*i*.

The program input contains model and reflection data files, which must be in PDB and MTZ formats, respectively. Both files are mandatory. The input model is assumed to have already been refined against input data and therefore both must have the same SG and unit-cell parameters. A readability check is performed with *REFMAC*5.

The program has two modes. In the default mode it performs a series of refinements but outputs only the model that it considers to be the best. The model is in the PDB format. In addition, the output contains an MTZ file with *REFMAC*5 map coefficients. In the second mode no refinements are performed; instead, the input model and data are converted into SGs consistent with observed unit-cell parameters and these models and data sets are stored in a directory defined by a user.

Importantly, the transformed data in the output MTZ files are generated from already merged input data. If the initial and final SGs have different point groups, the transformed data should not be used in later stages of refinement; by no means should they be used for the PDB deposition. For these two purposes the original experimental data have to be processed again in the selected SG. In a future version of *Zanuda*, which will have the option of using unmerged input data, the necessity of reprocessing the data will be avoided.

## 5. Conclusions

Problems in macromolecular structure solution and refinement usually manifest themselves with stubbornly high values of $R_{\rm cryst}$ and $R_{\rm free}$. The possible causes range from a wrong MR solution to crystal disorder. Misinterpretation of pseudo-symmetry operations as the true crystallographic operations at the data-reduction stage is one of the most confusing mistakes, because the structure still might be 'solved' in the wrong space group with good initial progress in model rebuilding and refinement. For structures with pseudo-translation, a mistake of the same nature may happen further downstream in the structure-determination process, at the stage of phasing, especially when phasing is performed using MR. The pseudo-translation, if present, and the true crystallographic axes generate pseudo-symmetry axes of the same order and orientation. A misinterpretation of the axis types occurs if the phasing program assigns the pseudo-origin as the true crystallographic origin. In this paper, the geometry and symptoms of the pseudo-origin solutions as well as methods for their correction are discussed using five real examples in which the pseudo-origin problem was encountered during structure determination. It should be highlighted that a wrong choice of crystallographic origin is a gross mistake and the pseudo-origin structure is an incorrect solution, not a different interpretation of the true structure.

This paper introduces the program *Zanuda*, which is intended to automatically restore the correct space group in structures with misinterpreted pseudo-symmetry. In particular, *Zanuda* successfully corrects the input pseudo-origin models in all of the examples in this paper. The automatic procedure involves a series of refinements in the candidate space groups and uses relative values of $R_{\rm free}$ after refinement as a selection criterion. Absolute values of overall refinement statistics are not taken into consideration because the input data and model may vary in quality; in addition, *Zanuda* removes solvent molecules from the input model and trims (pseudo)symmetry-related macromolecules in order to equalize their composition. In particular, in the examples provided the final $R_{\rm free}$ for the corrected output model varies from 0.32 to 0.47 and the difference in $R_{\rm free}$ between the pseudo-origin and corrected models varies from 0.03 to 0.14, with the lower $R_{\rm free}$ corresponding to the higher difference. Although examples of genuine pseudo-symmetry with this difference being less than 0.03 do exist, such a small value usually indicates either that the PSSG coincides with the true crystal space group, that the input model is not yet good enough or that *Zanuda* has failed to escape from an incorrect local minimum. In such cases *Zanuda* should be considered as an auxiliary tool and its results used as a guideline for further data reprocessing, structure solution and refinement. For example, rebuilding and refinement of the model, even in an incorrect SG, usually improves contrast in the subsequent *Zanuda* run. In conclusion, it is important to highlight that the interpretability of electron density, particularly ligand density, is the ultimate criterion for macromolecular refinement or any procedure that uses it.

The authors would like to thank Dr Paul Young and Dr Garib Murshudov for useful discussions and help with development of the *Zanuda* program and its deployment on the YSBL server, Professor Jennifer Littlechild, Dr Christopher Sayer and Aaron Westlake working on the AT project and Dr Vladimir Levdikov, and Dr Elena Blagova and Dr Aleksandra Watson for their very interesting examples of CodY and CLEC5A. The authors thank the Diamond Light Source, UK for access to beamline I24 (proposal No. MX6851) and the beamline staff scientists. AL is grateful to STFC UK and CCP4 for funding. MI is grateful to the University of Exeter and for the BBSRC-funded ERA-IB grant BB/L002035/1.

## References

Au, K. *et al.* (2006). *Acta Cryst.* D**62**, 1267–1275.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

Carter, C. W. & Sweet, R. M. (1997). *Methods Enzymol.* **276**, 286–494.

Cowtan, K. (2010). *Acta Cryst.* D**66**, 470–478.

Crowther, R. A. & Blow, D. M. (1967). *Acta Cryst.* **23**, 544–548.

Dauter, Z., Botos, I., LaRonde-LeBlanc, N. & Wlodawer, A. (2005). *Acta Cryst.* D**61**, 967–975.

DeLano, W. L. (2002). *PyMOL*. http://www.pymol.org.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.

Evans, P. R. (2011). *Acta Cryst.* D**67**, 282–292.

Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst.* D**69**, 1204–1214.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proc. R. Soc. Lond. A*, **225**, 287–307.

Isupov, M. N. & Lebedev, A. A. (2008). *Acta Cryst.* D**64**, 90–98.

Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.

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

Lebedev, A. A. & Isupov, M. N. (2012). *CCP4 Newsl. Protein Crystallogr.* **48**, contribution 11.

Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2006). *Acta Cryst.* D**62**, 83–95.

Lee, S., Sawaya, M. R. & Eisenberg, D. (2003). *Acta Cryst.* D**59**, 2191–2199.

Leslie, A. G. W. & Powell, H. R. (2007). *Evolving Methods for Macromolecular Crystallography*, edited by R. J. Read & J. L. Sussman, pp. 41–51. Dordrecht: Springer.

Levdikov, V. M., Blagova, E., Colledge, V. L., Lebedev, A. A., Williamson, D. C., Sonenshein, A. L. & Wilkinson, A. J. (2009). *J. Mol. Biol.* **390**, 1007–1018.

Levdikov, V. M., Blagova, E., Joseph, P., Sonenshein, A. L. & Wilkinson, A. J. (2006). *J. Biol. Chem.* **281**, 11366–11373.

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.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Padilla, J. E. & Yeates, T. O. (2003). *Acta Cryst.* D**59**, 1124–1130.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst.* D**54**, 1285–1294.

Pletnev, S., Morozova, K. S., Verkhusha, V. V. & Dauter, Z. (2009). *Acta Cryst.* D**65**, 906–912.

Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. (2003). *Acta Cryst.* D**59**, 1131–1137.

Powell, H. R., Johnson, O. & Leslie, A. G. W. (2013). *Acta Cryst.* D**69**, 1195–1203.

Rossman, M. G. (1972). Editor. *The Molecular Replacement Method.* New York: Gordon & Breach.

Rye, C. A., Isupov, M. N., Lebedev, A. A. & Littlechild, J. A. (2007). *Acta Cryst.* D**63**, 926–930.

Sayer, C., Isupov, M. N., Westlake, A. & Littlechild, J. A. (2013). *Acta Cryst.* D**69**, 564–576.

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

Shevtsov, M. B., Chen, Y., Isupov, M. N., Leech, A., Gollnick, P. & Antson, A. A. (2010). *J. Struct. Biol.* **170**, 127–133.

Skubák, P. & Pannu, N. S. (2013). *Nature Commun.* **4**, 2777.

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

Vagin, A. A. & Isupov, M. N. (2001). *Acta Cryst.* D**57**, 1451–1456.

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

Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* D**66**, 22–25.

Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. (2007). *Methods Mol. Biol.* **364**, 215–230.

Watson, A. A., Lebedev, A. A., Hall, B. A., Fenton-May, A. E., Vagin, A. A., Dejnirattisai, W., Felce, J., Mongkolsapaya, J., Palma, A. S., Liu, Y., Feizi, T., Screaton, G. R., Murshudov, G. N. & O'Callaghan, C. A. (2011). *J. Biol. Chem.* **286**, 24208–24218.

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

Winter, G. (2010). *J. Appl. Cryst.* **43**, 186–190.

Yeates, T. O. (1988). *Acta Cryst.* A**44**, 142–144.