Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Type II dehydroquinase: molecular replacement with many copies

Kirsty Anne Stewart,[a] David Alexander Robinson[a,b] and Adrian Jonathan Lapthorn[a]*

[a]Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland, and [b]Structural Sciences, Vernalis (R&D) Ltd, Granta Park, Cambridge CB21 6GB, England

Correspondence e-mail: adrian@chem.gla.ac.uk

Received 7 May 2007
Accepted 31 October 2007

Type II dehydroquinase is a small (150-amino-acid) protein which in solution packs together to form a dodecamer with 23 cubic symmetry. In crystals of this protein the symmetry of the biological unit can be coincident with the crystallographic symmetry, giving rise to cubic crystal forms with a single monomer in the asymmetric unit. In crystals where this is not the case, multiple copies of the monomer are present, giving rise to significant and often confusing noncrystallographic symmetry in low-symmetry crystal systems. These different crystal forms pose a variety of challenges for solution by molecular replacement. Three examples of structure solutions, including a highly unusual triclinic crystal form with 16 dodecamers (192 monomers) in the unit cell, are described. Four commonly used molecular-replacement packages are assessed against two of these examples, one of high symmetry and the other of low symmetry; this study highlights how program performance can vary significantly depending on the given problem. In addition, the final refined structure of the 16-dodecamer triclinic crystal form is analysed and shown not to be a superlattice structure, but rather an *F*-centred cubic crystal with frustrated crystallographic symmetry.

1. Introduction

Molecular replacement where multiple molecules are searched for is not unusual owing to the frequency of non-crystallographic symmetry in protein crystal structures (Kleywegt, 1996; Vonrhein & Schulz, 1999). Many-body searches in molecular replacement are usually carried out sequentially: an individual copy of the molecule is located, its contribution is fixed and the next molecule is then searched for (Navaza, 1994). Each step is dependent on the information gained in the previous rotation and translation searches; therefore, if many molecules are to be found, the accuracy of the initial solutions is critical for success. Crystal structures containing many molecules can prove to be difficult molecular-replacement problems. In part, this is a consequence of the fraction of the total scattering contributed by the individual search model being small and hence the signal being searched for being weak. This problem will be exacerbated if the search model differs significantly from the target structure owing to low sequence similarity, disordered regions and systematic deviations introduced by domain movements. A solution to the weak signal being searched for is the application of automated search procedures, which evaluate many peaks in the rotation and translation searches; success still relies on the ability to identify correct solutions *via* a given scoring function. More recently, the introduction of maximum-likelihood molecular-replacement functions have made a

significant impact on this problem (McCoy, 2007). A number of solutions to the problem of many molecules have been proposed; for example, when the macromolecular assembly of the protein is known with simple point-group symmetry then a locked rotation function can be used (Tong & Rossmann, 1990). Some molecular-replacement procedures attempt to locate many bodies through simultaneous multi-dimensional searches (Kissinger *et al.*, 1999; Glykos & Kokkinidis, 2000), while a method permitting the construction of a multi-copy search model from properly oriented monomers using a special translation function has also been proposed (Vagin & Teplyakov, 2000). In cases where there is significant translational symmetry, which can be identified from strong peaks in the Patterson function of the X-ray data, a subset of the data can be considered, reducing the number of independent bodies to be located and significantly reducing the computational time (Navaza *et al.*, 1998). Overall, these methods have allowed the solution of complex multiple-copy problems such as 15 copies of the aggregation-prone V_κ antibody fibre (McCoy, 2007), 16 copies of the ribonuclease inhibitor barstar (Navaza *et al.*, 1998) and 24 copies of superoxide dismutase with twinning (Lee *et al.*, 2003).

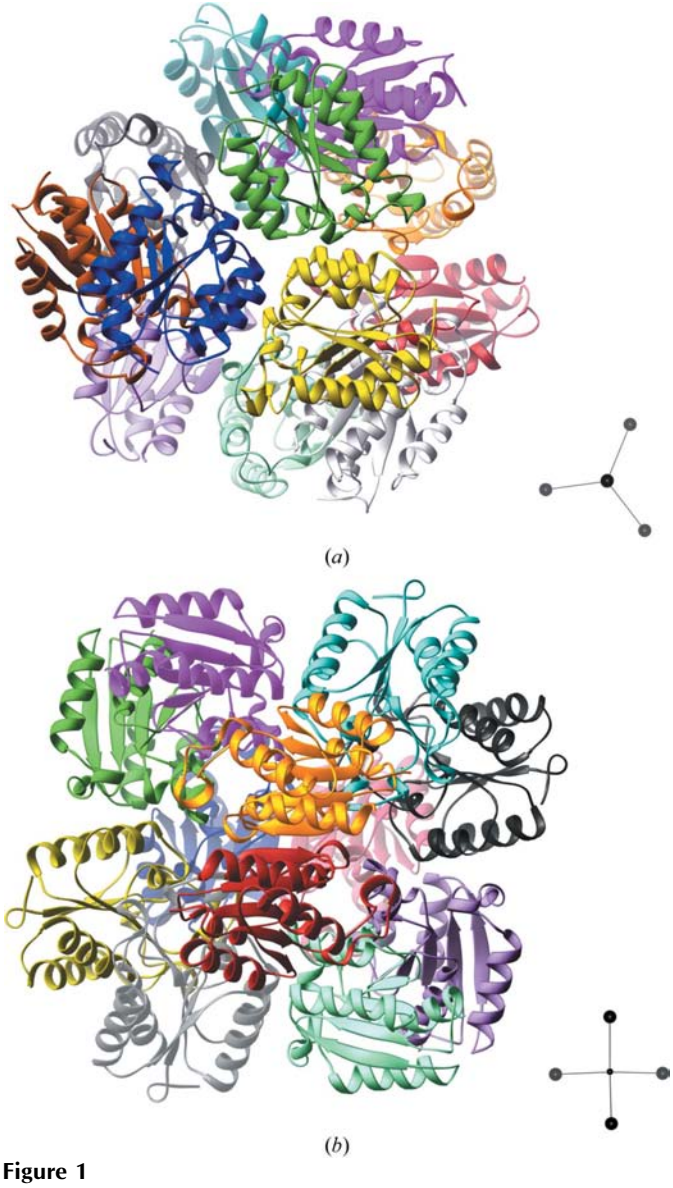

Figure 1
A ribbon representation (Carson, 1991) of the dodecamer structure of type II DHQase from *S. coelicolor* viewed down the threefold axis (*a*) and down a twofold axis (*b*). To the right of each dodecamer view is a five-sphere tetrahedral-shaped model of the structure. A small central sphere represents the centroid of the dodecamer; the four surrounding spheres (which are joined to the centroid by lines) represent the position of the threefold axes.

In this paper, we will consider a number of structures of type II dehydroquinase (DHQase), which catalyses the dehydration of dehydroquinate to dehydroshikimate, the third step in the shikimate pathway. The importance of the shikimate pathway for the synthesis of aromatic compounds in microorganisms and its absence in animals make the enzymes attractive targets for the development of antimicrobials (Abell, 1999). Type II DHQase is a small (~150 amino-acid residue) protein with a flavodoxin-like fold and a dodecameric quaternary structure in which the monomers are related by cubic 23 symmetry (Fig. 1). The structure has been solved in a number of different crystal forms, including a single monomer in the asymmetric unit of *Mycobacterium tuberculosis* DHQase in the cubic space group $F23$ (Gourley *et al.*, 1999), a dodecamer of the *Streptomyces coelicolor* enzyme in space groups $P2_12_12_1$ and $P2_1$ (Roszak *et al.*, 2002) and the related protein from *Bacillus subtilis* in space group $P2_1$ with two dodecamers (24 monomers) in the asymmetric unit (Robinson *et al.*, unpublished work). A transition-state analogue 2,3-anhydroquinate ($K_i = 30\ \mu M$), designed to mimic the flattened enolate intermediate, provided a starting inhibitor for structure-based drug design. The crystal structure of this inhibitor with the *S. coelicolor* enzyme (Roszak *et al.*, 2002) not only identified the binding mode of the substrate but also showed glycerol and tartrate molecules (derived from the crystallization mother liquor) in close proximity. These additional ligand-binding sites have been the focus of inhibitor design by several groups (Sanchez-Sixto *et al.*, 2005; Toscano *et al.*, 2003; Payne *et al.*, 2007; Prazeres *et al.*, 2007), which has resulted in the structure-based design of nanomolar inhibitors. The high-resolution structure determination of DHQase–inhibitor complexes is critical for the understanding of ligand binding and to identify new directions for inhibitor design. New crystal forms are encountered if cocrystallization of inhibitors is performed; in addition, new crystallization conditions must be found for enzymes from different species. These provide interesting and sometimes challenging cases for structure solution by molecular replacement. Several cases involving type II DHQase will be discussed here, including a large $P1$ structure containing 16 dodecamers in the asymmetric unit.

2. Molecular-replacement problems

2.1. *Helicobacter pylori* DHQase–AH9095 complex: a high-symmetry example

The crystal structure of *H. pylori* DHQase with inhibitor AH9095 (Robinson *et al.*, 2006) proved to be a difficult

molecular-replacement problem owing to the cubic symmetry of the crystal and the low sequence identity with previously solved DHQase structures. X-ray data were collected to 1.5 Å resolution at the Daresbury SRS; the unit-cell parameters were $a = b = c = 131.4$ Å and the space group was $F23$. The space group has 48 symmetry operators and from calculation of the Matthews number (Tronrud *et al.*, 1987) it seemed likely that the asymmetric unit would contain a single DHQase monomer, resulting in a solvent content of 50%. The available DHQase structures which could be used as search models were from *M. tuberculosis*, *B. subtilis* and *S. coelicolor*, which had 33.5, 37.5 and 34.1% sequence identity, respectively, to *H. pylori* DHQase. Various attempts at molecular replacement using the three search models with the stand-alone version of *AMoRe* (Navaza, 1994) were unsuccessful; the absence of a strong single peak in the rotation function was a noticeable feature of these attempts. From the known quaternary structure of the enzyme (Fig. 1), it can be seen that the monomer makes extensive contacts within the dodecamer in addition to any crystal contacts; therefore, the separation of self-vectors and cross-vectors is poor in this case. In addition, the high symmetry of the structure means that self-vectors from all the molecules which are rotationally distinct in the unit cell (12 in this case) are present, further confusing the data to be searched. The low sequence identity of the search model is also a significant problem. The structure was eventually solved with a polyalanine model of the *M. tuberculosis* enzyme (PDB code 2dhq) using the program *EPMR* (Kissinger *et al.*, 1999), which we have found to be useful for problematic molecular replacements such as this.

2.2. *H. pylori* DHQase: the wrong space group

Crystallization trials of *H. pylori* DHQase mixed with a fivefold stoichiometric concentration of the non-substrate-like inhibitor GR12160X resulted in crystals with hexagonal bipyramid morphology in conditions containing 20% 1,4-butanediol, 0.1 *M* sodium acetate pH 4.5. The crystals diffracted to 2.9 Å resolution on beamline 9.6 at SRS Daresbury, where 60° of data were collected. Indexing the data with *DENZO* (Otwinowski & Minor, 1997) indicated a rhombohedral space group, with unit-cell parameters $a = b = 182.9$, $c = 658.8$ Å, $\gamma = 120°$ (in the hexagonal setting). Scaling the data with *SCALEPACK* (Otwinowski & Minor, 1997) showed the space group to be $R32$, with an $R_{\rm merge}$ of 10.2%, 99.4% completeness and an average redundancy of 3.4. The Matthews number was consistent with two dodecamers (24 monomers) in the asymmetric unit, so a dodecamer created from the 1.5 Å *H. pylori* DHQase–AH9095 inhibitor complex structure should have made an excellent search model. All attempts to solve this structure using *AMoRe* or *EPMR* gave no hint of a solution, even when using a smaller search model such as a trimer. Despite the sensible merging *R*-factor statistics for data in $R32$, the data were reprocessed in lower symmetry space groups (with no significant improvement in processing statistics) and molecular replacement was repeated using the dodecamer as a search model. Using data processed in $C2$ (65% complete), with unit-cell parameters $a = 316.78$, $b = 181.90$, $c = 243.66$ Å, $\beta = 115.67°$ and presumably six dodecamers in the asymmetric unit (based on the Matthews number), strong peaks were seen in the rotation function, resulting in a good molecular-replacement solution for four dodecamers using *AMoRe* with data between 12.0 and 4 Å resolution. This solution gave an *R* factor of 41.7% and a correlation coefficient of 64.3% after rigid-body refinement. Analysis of the self-rotation function at this point highlighted 60° and 180° rotations coincident with the *y* axis at 50% of the correlation expected for crystallographic symmetry, confirming the correct choice of space group. Refinement using *REFMAC*5 (Murshudov *et al.*, 1996), all data to 2.9 Å and tight noncrystallographic symmetry (NCS) restraints (at dodecamer level) gave an $R_{\rm work}$ of 32.1%, an $R_{\rm free}$ of 41.7% and a correlation coefficient of 76%. Inspection of the structure using *Coot* (Emsley & Cowtan, 2004) revealed electron density corresponding to a missing dodecamer which was coincident with a crystallographic twofold axis of symmetry. A model of half a dodecamer was produced and a rotation followed by a phased translation search using weighted difference map Fourier coefficients from *REFMAC*5 was performed with *MOLREP* (Vagin & Isupov, 2001). This completed the structure solution, giving 4.5 dodecamers in the asymmetric unit (54 monomers). Further refinement with *REFMAC*5 lowered $R_{\rm work}$ to 29.3% and $R_{\rm free}$ to 39.1%, with a correlation coefficient of 87.5%. Inspection of the structure using *Coot* revealed electron density for two β-strands at several crystal contacts between dodecamers; this density arose from an ordered portion of the N-terminal polyhistidine tag. No electron density consistent with the inhibitor GR12160X was present in any of the 54 NCS-related active sites and therefore in light of this and the incomplete data, refinement of the structure was not continued. This example serves as a reminder that the presence of NCS in the structure can confuse the interpretation of the correct space group and in this case the crystal system.

2.3. *S. coelicolor* DHQase–CA1 complex: 192 molecules

As only approximately 1 mg of the CA1 bifunctional inhibitor was available for cocrystallization with *S. coelicolor* DHQase, extensive crystallization screening was not possible in this case. Cocrystallizations were set up with an in-house PEG/ion screen which had previously been successful in producing a well diffracting orthorhombic crystal form of this enzyme (Roszak *et al.*, 2002). A dozen crystals were screened for suitably high-resolution diffraction at the Daresbury SRS as the crystal quality was variable and the crystals suffered from anisotropic diffraction and high mosaicity (frequently 2°). One of the crystals diffracted to 1.7 Å resolution with low mosaicity, allowing 250° of oscillation data to be collected. Indexing the diffraction pattern with *DENZO* gave a large *I*-centred tetragonal cell with unit-cell parameters $a = b = 196.6$, $c = 393.6$ Å; however, the data did not merge with this symmetry or as *I*-centred orthorhombic. Indexing the data in lower symmetry such as *C*-centred monoclinic with unit-cell

Table 1
Results of 16-body molecular replacement of the *S. coelicolor* DHQase–CA1 inhibitor complex from *AMoRe*.

Euler angles (°)			Translations (fractional coordinates)					
φ	θ	ψ	x	y	z	CC	R factor	Translation rank by CC
286.6	86.8	119.3	0	0	0	15.9	59.1	
2.4	106.6	30.2	0.8775	0.8705	0.7577	28.7	57.0	1st
286.6	86.8	119.3	0.1273	0.1204	0.2416	31.4	57.7	4th
286.6	86.8	119.3	0.7332	0.2493	0.5002	34.3	57.4	7th
163.5	93.1	299.8	0.7422	0.7484	0.5004	36.9	55.0	7th
163.5	93.1	299.8	0.6118	0.6234	0.258	40.0	54.0	7th
16.2	87	119.6	0.9881	0.4972	0.9983	42.1	52.8	7th
182.4	106	29.8	0.6084	0.1232	0.2593	45.1	51.9	5th
16.2	87	119.6	0.8626	0.3741	0.7565	47.8	51.1	1st
16.2	87	119.6	0.2323	0.2573	0.4996	49.5	50.5	=1st
196.2	86.1	119.5	0.5015	0.0075	0.9993	52.0	49.6	1st
106.5	86.2	118.1	0.2395	0.7564	0.4983	53.8	48.8	=1st
73.9	93.4	298.1	0.489	0.5085	0.9992	55.4	48.1	=2nd
73.9	93.4	298.1	0.3634	0.3843	0.757	57.0	47.4	=1st
196.2	86.1	119.5	0.3719	0.8827	0.7559	58.6	46.9	=1st
106.5	86.2	118.1	0.1175	0.6313	0.2606	59.6	46.5	1st
After rigid-body refinement						71.1	39.3	

parameters a = 280.4, b = 280.8, c = 242.7 Å, β = 125.2° was equally unsuccessful; however, the data were not mis-indexed as they merged acceptably in space group $P1$. This large triclinic crystal form had been encountered previously (Roszak *et al.*, 2002) and had been abandoned as being impractical. In the absence of any other more amenable crystals of this DHQase–inhibitor complex, a structure solution was attempted using these data. The data were processed in $P1$ with unit-cell parameters a = 196.61, b = 196.49, c = 240.62 Å, α = 65.9, β = 65.9, γ = 90.1° using the *HKL* suite of programs dimensioned for viruses. The data were 93% complete with 3 042 000 unique reflections to ∼1.7 Å and an average redundancy of 2.4. The merging R factor was 17% owing to the low redundancy and the weakness of the data beyond 2.0 Å, which was a consequence of the data being collected in one pass. Visual inspection of the merged diffraction data with *HKLVIEW* (Collaborative Computational Project, Number 4, 1994) showed no pattern of alternating strong and weak reflections as is observed in superlattice structures where an inexact lattice translation has resulted in a doubling of a cell dimension; instead, a curious striped pattern involving blocks of data was seen (Fig. 2). Analysis of the self-rotation function showed three perpendicular twofold axes, in addition to other threefold and fourfold rotations at a maximum of 83% of that expected for crystallographic symmetry. The native Patterson map showed a peak half the size of the origin at 0.127, 0.126, 0.240, indicating translational NCS. A Matthews coefficient of 2.4 Å^3 Da^{-1} can be derived for 16 dodecamers in the unit cell; this assumes 50% solvent content for this crystal form as is found in other *S. coelicolor* DHQase structures. Given such a large number of molecules in the $P1$ cell, it was presumed that the structure was a superlattice structure of some type. A search of the Protein Data Bank (PDB; Sussman *et al.*, 1998) based on the size of the asymmetric unit (Fig. 3) reveals that this structure is not unique in its size, as structures with a complete virus as the asymmetric unit have been solved. However, for its resolution, this structure is a significant outlier some 4–5 times larger in size than anything previously deposited at this or higher resolution.

Molecular replacement was attempted using the standalone version of *AMoRe* with a single dodecamer of the *S. coelicolor* DHQase (PDB code 1gu1) as the search model, data from 12.0 to 4.0 Å resolution and a 25 Å centre-of-mass exclusion between solutions. Contrary to our expectations, the

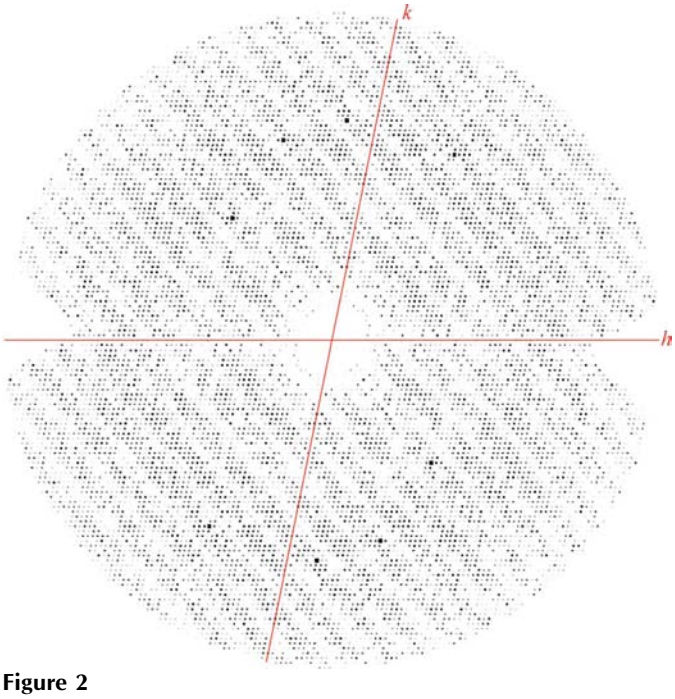

Figure 2
l = 0 section of X-ray data from the *S. coelicolor* DHQase–CA1 inhibitor structure, illustrating the striped nature of strong and weak data in the diffraction data.

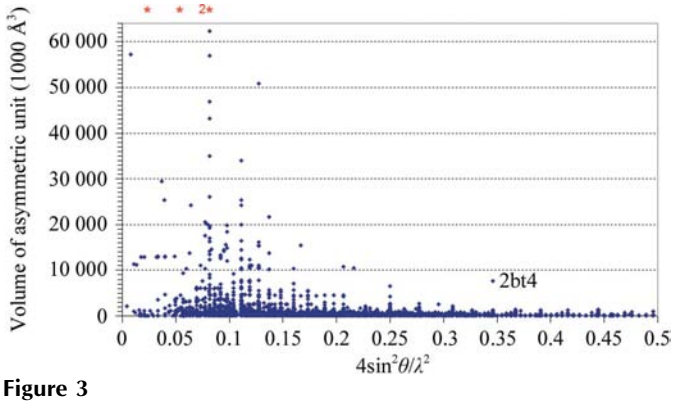

Figure 3
A plot of asymmetric unit volume against resolution for the X-ray crystal structures in the PDB. Entries with very small unit cells have been omitted, as have structures determined at higher than 1.5 Å resolution. Four structures are represented by red stars as they do not fit onto the plot; namely (from left to right) entries 1wce (birnavirus, 234 × 10^6 Å^3), 1w8x (bacteriophage Prd1, 192.5 × 10^6 Å^3), 1ohg (Hk97 bacteriophage capsid, 143 × 10^6 Å^3) and 2btv (bluetongue virus core, 123 × 10^6 Å^3). The type II dehydroquinase structure reported here has an asymmetric unit volume of 7.6 × 10^6 Å^3 and is labelled 2bt4.

Table 2
Comparison of the performance of four common molecular-replacement programs on two test cases of data.

	AMoRe	*MOLREP*	*EPMR*	*Phaser*
H. pylori DHQase (*F*23), $Z' = 1$				
Correct solution	Yes	Yes	No	Yes
R factor/CC†	0.46/0.42	0.51/0.33 (0.48/0.38‡)	0.54/0.20	0.46/0.43
Computation time	54 min	1 h 13 min	23 min§	4 min
S. coelicolor DHQase (*P*1), $Z' = 192$				
Correct solutions	Yes (16)	Yes (14)	Yes (11)	Yes (16)
R factor/CC†	0.37/0.77	0.55/0.51 (0.41/0.74‡)	0.51/0.54	0.33/0.88
Computation time	57 min	5 h 12 min	2 d 16 h§	30 d 6 h

† All *R* factors and correlation coefficients were calculated from output coordinates using *REFMAC*5 for consistency. ‡ After six cycles of rigid-body refinement using *REFMAC*5 with data to 4 Å. § Using all four processors of the computer.

16-body molecular replacement was both rapid and successful. The rotation function gave only 24 significant solutions and from these each of the *n*-body translation solutions gave a steady increase in correlation and decrease in *R* factor (Table 1). The rigid-body refinement of the final six solutions took twice as long as the molecular replacement itself and resulted in a large increase in the correlation coefficient and decrease in the *R* factor consistent with a correct solution (Table 1). The choice of packing function is critical to the success of the molecular replacement, as the solutions chosen for the third through to the eighth fixed solution were not ranked the best solutions as judged by the correlation coefficient. Analysis of the rigid-body refinement of the solutions showed that the third fixed solution was the furthest away from the correct position, needing a 4.1° rotation in θ and 1 Å translation in *x*. The other fixed solutions were not changed by rigid-body refinement by more than a 1.0° rotation and a translation of 0.5 Å. The packing of the final solution (Fig. 4) was seen to be correct using *RASMOL* (Sayle & Milnerwhite, 1995) as the structure was too large to be viewed in a number of other graphics viewers available at that time.

3. Comparison of molecular-replacement packages

The *H. pylori* DHQase–AH9095 complex and the *S. coelicolor* DHQase–CA1 complex represent two extreme examples for molecular replacement: that of high symmetry and multiple copies, respectively. Four molecular-replacement packages were assessed to compare their performances using these two examples. The programs assessed were a recent stand-alone version of *AMoRe* (Navaza, 1994), *MOLREP* v.9.4.09 (Vagin & Teplyakov, 1997), *EPMR* v.3.1 (Kissinger *et al.*, 1999) and *Phaser* v.1.3.3 (McCoy *et al.*, 2005). The four packages are similar in that they are highly automated, requiring relatively little user input. *AMoRe*, *MOLREP* and *Phaser* split the molecular-replacement problem into a rotation and a translation search, but differ substantially in treatment of data and their scoring of solutions. In contrast, *EPMR* is a six-dimensional search program which uses a stochastic search algorithm. This iteratively optimizes populations of trial solutions, choosing the best trial solutions of one population to generate a new population of solutions by introducing small random variations in parameters. Current versions of the four programs were used and calculations were performed on a dual-processor 2.8 GHz Intel Xeon Unix workstation with 4 GB memory. As each program uses its own scoring function, for consistency the coordinates of the output solutions were refined using *REFMAC*5 to generate an *R* factor and correlation coefficient for all data to 4.0 Å. For the *H. pylori* DHQase molecular replacement the complete protein model of a 1.5 Å structure of *M. tuberculosis* DHQase (PDB code 1h05) was used, while for the *S. coelicolor* DHQase case a complete dodecamer of the *S. coelicolor* enzyme determined at 1.8 Å (PDB code 1gu1) was used. X-ray data were treated with a high-resolution cutoff at 4.0 Å, while depending on the program used a low-resolution cutoff was applied at 15 Å, 12 Å or not at all.

In this first test case, the *H. pylori* DHQase–AH9095 complex, three programs successfully found a solution: *Phaser* and *MOLREP* using all data to 4.0 Å with default settings and *AMoRe* using a 12 Å low-resolution cutoff. In contrast, *EPMR* did not find a correct solution with several runs of the program and an increased population size of up to 600 (Table 2). On comparing the solutions found by *Phaser*, *AMoRe* and *MOLREP*, it appears from the statistics that the *MOLREP* solution is worse than those found by the other two programs. *MOLREP* as implemented in *CCP*4 sets all *B* values to 20 Å^2 and does not perform rigid-body refinement of the final solutions as standard, which accounts for the difference. Treating the solutions to five cycles of restrained refinement using *REFMAC*5 applied to all data to 2.0 Å, the *Phaser* and *AMoRe* solutions gave a comparable $R_{\rm work}$ of 0.420 and an $R_{\rm free}$ of 0.485, while the *MOLREP* solution had an $R_{\rm work}$ of 0.461 and an $R_{\rm free}$ of 0.537; the resultant weighted Fourier maps reflected this difference. In terms of computational time, Phaser was 12 times faster than either *AMoRE* or *MOLREP*.

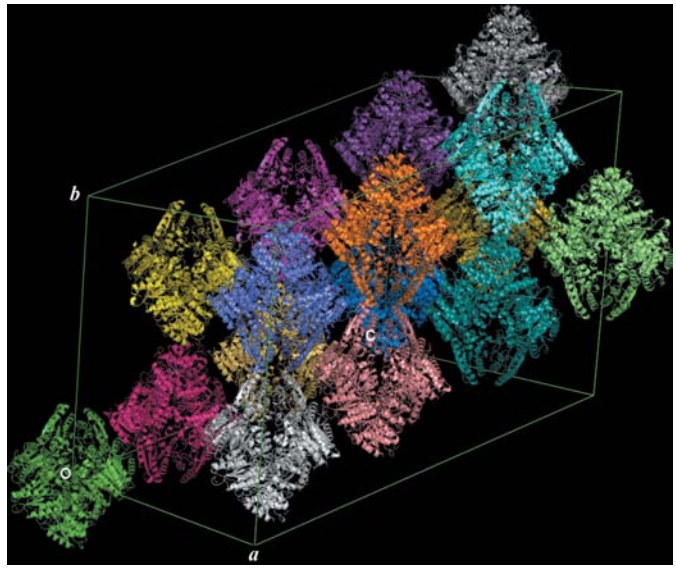

Figure 4
Packing of the 16 dodecamers within the *P*1 unit cell of the *S. coelicolor* DHQase–CA1 structure.

In the second test case (the *S. coelicolor* DHQase–CA1 complex), assessment of the programs is complicated by the number of molecules to be found. A successful solution need not find all 16 dodecamers; if a sufficient proportion of the overall structure is correctly positioned then this can be refined and used to locate the rest of the structure, *e.g.* by using a phased translation search as in the example in §2.2. In this way, multi-copy structure solution is potentially easier than the single-molecule case as it is tolerant to partial solutions and it is possible to use averaging over multiple copies of the protein to enhance the quality of the electron density to be interpreted. To illustrate this point, Fig. 5 shows the effect of two incorrectly positioned dodecamers in a related non-isomorphous *P*1 *S. coelicolor* DHQase structure. The difference density is sufficiently unambiguous to identify that these dodecamers are incorrect and to permit manual repositioning of both molecules. Therefore, by this definition all four programs were able to find at least a partial solution to this molecular-replacement problem. *AMoRe* and *Phaser* found all 16 copies of the search model, with the *Phaser* solution giving better statistics than the *AMoRe* solution (Table 2). *MOLREP* only found 14 dodecamers using data between 12 and 4 Å resolution, while a run using all data to 4 Å resulted in only eight copies being located. The *R* factor and correlation coefficient for the 14-copy solution were far worse than those from either *AMoRe* or *Phaser*. Ten cycles of rigid-body refinement with *REFMAC*5 using all data to 4.0 Å resolution improved the *R* factor and correlation coefficient to acceptable values, confirming that the solution was of comparable quality. Inspection of the resultant electron-density map showed sufficient positive difference density that the remaining two dodecamers could be positioned with a phased translation search. Finally, in the case of *EPMR* 16 copies of the search model were found in the solution; however, five dodecamers almost exactly overlapped existing solutions, despite a bump radius of 25 Å being applied. It is reasonable to assume that if these overlapped copies had been avoided, a more complete solution may have been obtained. Removal of the five overlapping dodecamers followed by rigid-body refinement of the resultant 11 copies gave a value of 50% for both the *R* factor and the correlation coefficient. The electron-density maps gave some indication that at least one of 11 dodecamers was incorrectly placed and some indication of the missing five copies. This must therefore be considered as a borderline molecular-replacement solution which would not be profitable to pursue. If we consider computation time as in test case 1 (Table 2), *AMoRE*, with a solution in less than 1 h,

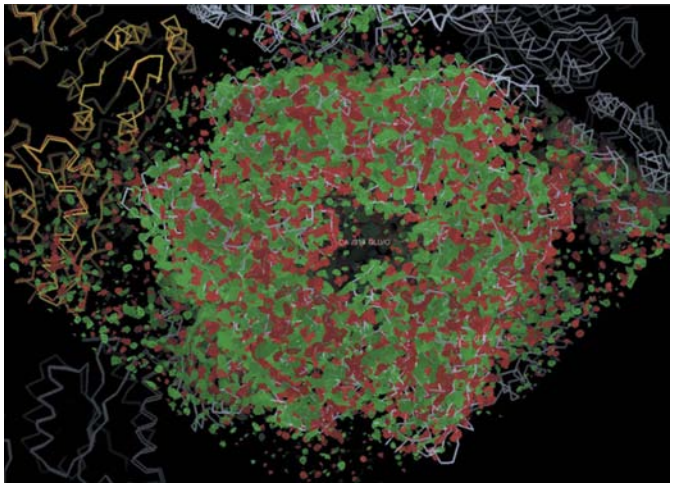

Figure 5
Difference electron density (green, positive; red, negative) from a molecular-replacement solution of a related *P*1 *S. coelicolor* DHQase inhibitor complex showing one of two incorrectly positioned dodecamers.

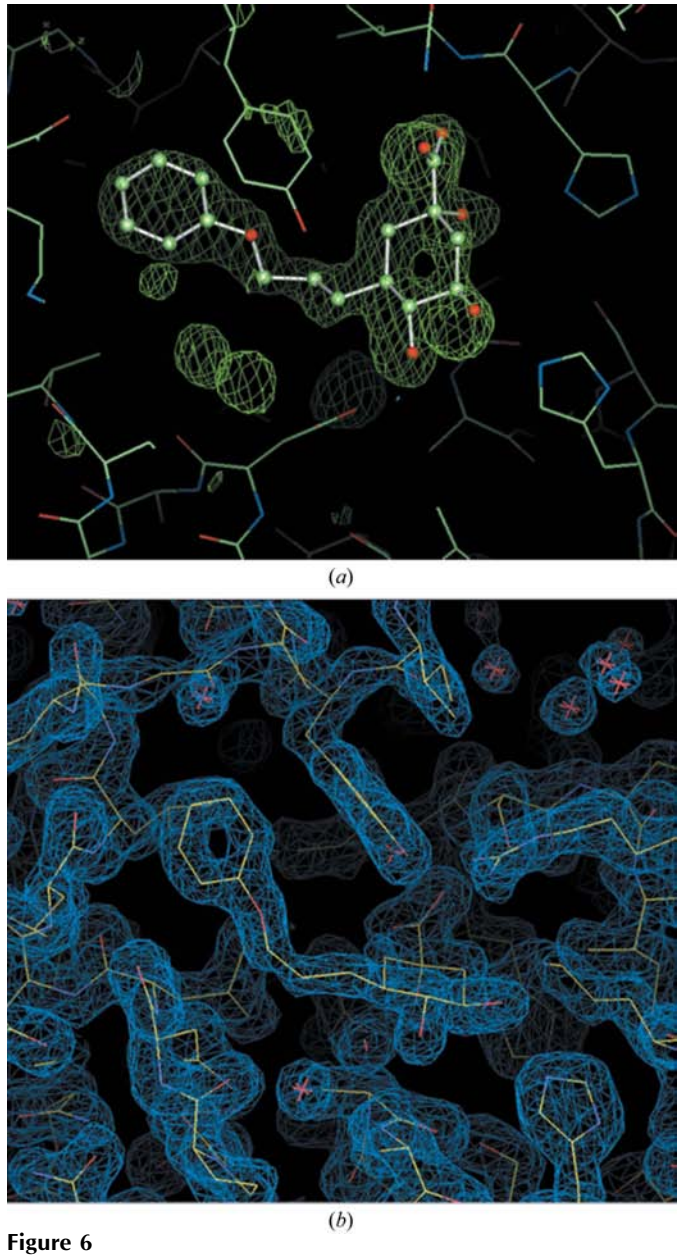

Figure 6
(*a*) The active site of *S. coelicolor* DHQase with 16-fold averaged weighted difference electron density corresponding to the inhibitor CA1 (the position of the unrefined model is shown in ball-and-stick representation), clearly showing the stereochemistry of the bifunctional inhibitor. Additional density corresponds to ordered water molecules in the active site of the enzyme. (*b*) The final unaveraged electron density contoured at 2σ for the active site of a representative monomer within the structure.

is five times faster than the partial solution obtained with *MOLREP* (without considering the additional time required for rigid-body refinement). *Phaser*, whilst producing the most accurate solution, is impractically slow in this test example.

In conclusion, the aim of this comparison was not to highlight inadequacies of any of the four programs, but rather to assess how well the various programs available perform on two very different difficult molecular-replacement problems. From the results shown in Table 2, it can be seen that the programs can take significantly different amounts of time to solve the same problem and do not all give the same quality of solution.

4. Completing the *S. coelicolor* DHQase–CA1 complex structure

The subsequent manipulation of the data set proved problematic as the current PDB format is not compatible with 192 individually labelled chains. To circumvent this problem, each dodecamer was given a unique chain identifier *A*–*P* and the monomers within it were numbered residues 1–150, 201–350, 401–550 . . . 2201–2350. This way, the structure could be refined using *REFMAC*5 with 16-fold averaging. The NCS operators relating chain *A* to the other chains within the asymmetric unit were calculated using *LSQKAB* (Collaborative Computational Project, Number 4, 1994) and a mask corresponding to the position of chain *A* was created using *NCSMASK* (Collaborative Computational Project, Number 4, 1994). The program *MAPROT* (Collaborative Computational Project, Number 4, 1994) was then used to superimpose the electron density from each monomer onto chain *A* using the density mask. Modifications to the model were initially performed using *X-AUTOFIT* within *QUANTA* (Accelrys) on the chain *A* dodecamer with 16-fold averaged $2F_o - F_c$ and $F_o - F_c$ maps. From the initial averaged difference map it was clear that the inhibitor CA1 was present within the structure (Fig. 6*a*). The ligand (CA1) was built and minimized using *INSIGHT*II (Accelerys) and fitted into the averaged $F_o - F_c$ density within the active sites of each monomer in dodecamer *A* with *QUANTA* (Accelrys). Water molecules were added manually using *X-SOLVATE* within *QUANTA* and automatically in the later stages of refinement using *ARP* (Lamzin & Wilson, 1997). After two rounds of refinement and model building using averaged maps, model building proceeded for a further three rounds using all the dodecamers and 16 chains of water molecules using *Coot*. The final model had an R_{work} of 0.197 and an R_{free} of 0.248 and comprises 216 514 protein atoms, 28 394 water molecules and 6208 ligand atoms. 91.8% of residues in the structure were in the most favoured regions of the Ramachandran plot, with the remaining 8.2% in allowed regions, while the molecular geometry was either as good as or better than expected values as evaluated by *PROCHECK* (Laskowski *et al.*, 1993).

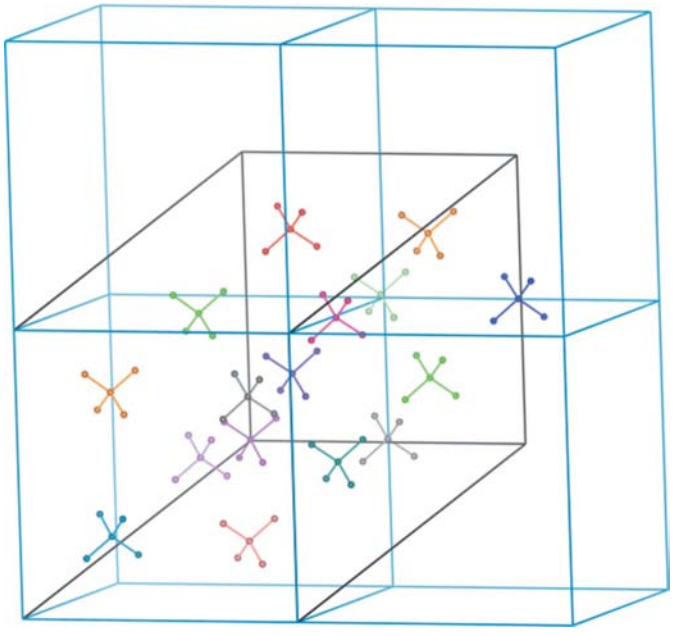

Figure 7
The relationship between the *P*1 unit cell of the *S. coelicolor* DHQase structure reported here (in black) and an idealized *F*-centred cubic lattice with four unit cells shown in blue. The dodecamers in the *P*1 cell are coloured uniquely and represented schematically by four atoms along the four threefold axes connected to a central centroid atom (see Fig. 1).

4.1. Overall packing

The symmetry of a crystal structure is normally reflected by the symmetry of the X-ray data and is usually resolved at the processing stage of the structure determination. The refined solution for the *P*1 structure indicates regular packing of dodecamers within the unit cell, suggesting that there is a higher symmetry present and that this higher symmetry is somehow not exact and therefore noncrystallographic. If we consider a single dodecamer, it is surrounded by eight other dodecamers in what crudely approximates a body-centred cubic structure with a unit cell with $a \simeq 98.3$ Å and $Z = 36$ (three dodecamers). The threefold axes of the central dodecamer are nearly coincident with the corners of the basic cube structure as would be expected. However, the dodecamers at the corners of this simple cell are not related by translation but by a twofold rotation; therefore, the unit cell must be twice as large, with $a \simeq 196.7$ Å and $Z = 192$ (16 dodecamers). The relationship between the unit cell of the *P*1 structure and an *F*-centred cubic cell is shown in Fig. 7, with the dodecamers represented schematically for clarity.

It is easiest to consider the simplest *F*-centered cubic space group, *F*23 (48 symmetry operators), which when applied to a monomer in an idealized *F*-centred unit cell with $a = 196.6$ Å generates dodecamers at each of the lattice points within the cell (Fig. 8*a*, black molecules). To generate the remaining dodecamers which are related by the twofold rotations, a second monomer related to the first by a twofold rotation around (1/4, *y*, 0) or equivalent is required, making an open network of molecules (Fig. 8*a*, red molecules). So far, this accounts for half the molecules present in the unit cell (*i.e.* 96 monomers or eight dodecamers). If we consider the higher symmetry space groups, it is possible to establish that $F4_132$,

which has 96 symmetry operators if applied to the two independent monomers, will generate all molecules in the actual unit cell (Fig. 8*b*). Comparison with the actual crystal structure superficially shows a good agreement between the idealized dodecamer positions and the true dodecamer coordinates (Fig. 8*c*). This rules out the NCS-related molecule alone being responsible for frustrating the true crystallographic symmetry by, for example, the dodecamer symmetry being 90° out of alignment with the cubic symmetry. Instead, there are relatively small deviations from ideal positions which are at a maximum of around 3.6 Å (measured by eye).

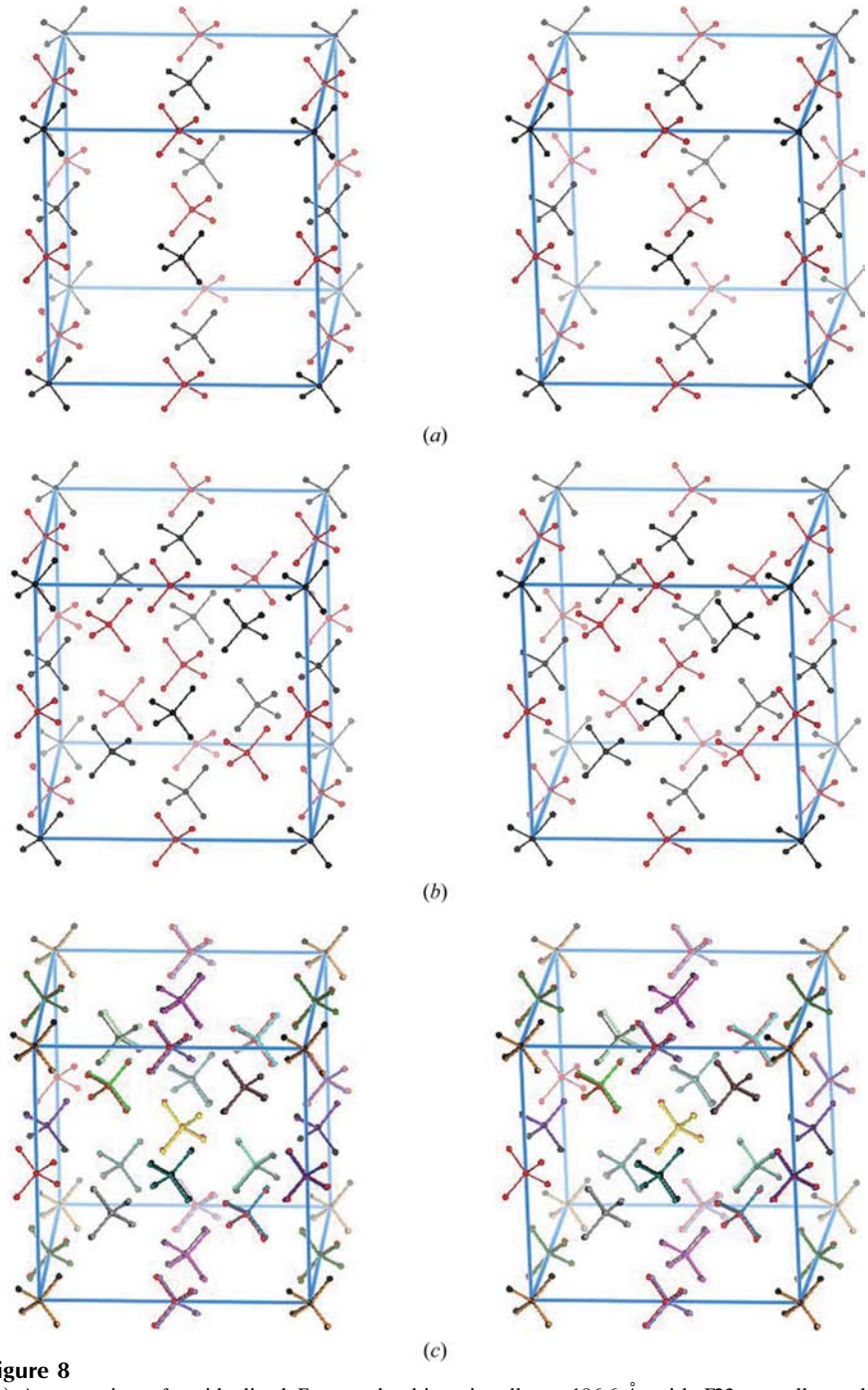

Figure 8
(*a*) A stereoview of an idealized *F*-centred cubic unit cell, a = 196.6 Å, with *F*23 crystallographic symmetry (48 symmetry operators) applied to two schematic dodecamers coloured black and red. This symmetry represents the basic arrangement of dodecamers in the *P*1 structure, with crystal contacts formed at the twofold axes within the each dodecamer so that, for example, the central dodecamer (red) forms crystal contacts with six other dodecamers (black) all at 90° from each other. (*b*) The packing in (*a*) leaves space for four dodecamers within the unit cell, which form a second lattice of dodecamers equivalent to the first but involving a translation of (1/4, 1/4, 1/4) from the origin. This arrangement is equivalent to $F4_132$ crystallographic symmetry (96 symmetry operators), which when applied to two monomers generates the entire 192 molecules found in the *P*1 crystal structure. (*c*) A comparison of the 16 NCS dodecamers in the *P*1 structure (coloured uniquely) with the idealized $F4_132$ cubic structure as shown in (*b*). This highlights the general agreement between the observed and idealized structures, while also illustrating the small differences in orientation in individual dodecamers which break the higher symmetry.

That the symmetry of this pseudo-cubic crystal structure is subtly broken such that it can only be treated in *P*1 runs counter to our preconceptions about molecules in the crystalline state. To understand the cause of these deviations, it is necessary to examine the crystal contacts between dodecamers in the crystal. Fortunately, these are very limited and are dominated by the interaction between dodecamers related by the twofold rotation between NCS monomers and seen in the *F*23 symmetry example (Fig. 8*a*). The crystal contacts involve residues 42–48 from helix α1 of the structure, which is rich in alanine residues and makes a symmetrical contact, with Ala45 packing against Lys42 of the symmetry-related dodecamer (Fig. 9). This small crystal contact region does not involve any hydrogen bonds and is dominated by van der Waals interactions which partially bury the two hydrophobic surfaces. The amine N atom of Lys42 is in proximity to the main-chain O atoms of Ala45 and Ala46 so that a hydrogen bond might be expected together with an electrostatic interaction with the end of the α-helix (residues 38–46). However, the conformation of this Lys42 in almost all monomers suggests that no significant interaction is present and that the amine interacts with solvent.

Thus, a pair of twofold crystal contacts is formed at six equivalent sites on the dodecamer, forming a network of interactions extending ∼90° in each direction and generating dodecamers centred at all permutations of coordinates with values 0 and 1/2 (Fig. 8*a*; chains *A*, *C*, *E*, *H*, *I*, *L*, *N* and *P*). There is sufficient space within this *F*23 lattice of dodecamers to pack one dodecamer

at the centre of a cube of eight dodecamers. These dodecamers are defined by the symmetry operators described by the higher symmetry $F4_132$. These dodecamers again form interactions at the twofold axes, involving residues 42–48 from helix $\alpha1$ forming a separate network of interactions, centred at all permutations of coordinates with values 1/4 and 3/4 (Fig. 8*b*; chains *B*, *D*, *F*, *G*, *J*, *K*, *M* and *O*).

The only question remaining is how these equivalent lattices interact with each other *via* crystal contacts. As the twofold ends of the dodecamers are all involved in crystal contacts, the accessible areas on the molecule that remain are around the dodecamer threefold axes. The first dodecamer view shown in Fig. 1 is down the threefold axis of the dodecamer, highlighting the flat compact nature of this region (represented by a sphere in the schematic), while the reverse face of the dodecamer is less suited to form crystal contacts owing to its open nature. Crystal contacts are seen between the compact threefold surfaces of the dodecamer (Fig. 9) and, not unexpectedly, involve three equivalent contact points. Each contact point involves residues from the N-terminus (residues 2–8), the end of helix $\alpha2$ (residues 65–73) and part of strand $\beta2$ (residues 50–54), with between six and 12 water molecules at each contact point. The interactions are either hydrophilic or van der Waals in nature, with several water-mediated hydrogen bonds (*e.g.* Thr50 OG1 to Asn72 OD1, Glu68 OE2 to Glu68 OE2 and Arg2 NH2 to Thr50 OG) and one direct hydrogen bond between the side chain Arg2 NH1 and main chain Asn6 O. From the interaction between dodecamers *E* and *G* (Fig. 9), it can be seen that the threefold crystal contact is not symmetrical but off-centre. Furthermore, if we consider that there are only five of these threefold contacts within the unit cell rather than potentially 32, the source of the breakdown in symmetry becomes clearer. With reference to Fig. 8(*b*), if we compare the environments of the idealized dodecamers formed by the first monomer (black) with those formed by the second NCS-related monomer (red), the dodecamers formed by the first monomer can make threefold-to-threefold crystal contacts, while those of the second monomer cannot. Therefore, only eight of the dodecamers can be involved in the threefold crystal contacts: four from each network of dodecamers described in the previous paragraph, *i.e.* *A*, *C*, *E* and *P* at positions 0 and 1/2 and *G*, *J*, *K* and *M* at positions 1/4 and 3/4. Owing to the ubiquitous twofold interactions between dodecamers involving the longest axis of the molecule, the remaining space is too large for any one dodecamer to form four symmetrical threefold interactions. In fact, only chains *E* and *M* form three threefold interactions, one with each other and two with two other dodecamers (*E* to *M*, *E* to *G* and *E* to *K*, *M* to *A* and *M* to *P*). The space is sufficiently large that dodecamers *C* and *J* make no threefold interactions, *e.g.* between *E* and *J* (Fig. 9).

In conclusion, a combination of factors are responsible for the breaking of the $F4_132$ cubic symmetry in this case, namely the predominant NCS twofold crystal contacts involving the longest dimension of the dodecamer structure producing slightly larger spaces which cannot be occupied symmetrically by another dodecamer. It is tempting to speculate that in the absence of any threefold interactions, the crystal would have *F*23 cubic symmetry with two monomers in the asymmetric unit (Fig. 8*a*). The presence of the threefold interactions breaks the cubic symmetry and distorts both networks of twofold interactions. With relatively few threefold interactions (five out of a possible 16), any imperfections within the packing of the crystal may result in twinning or disordering of one of the lattices, both of which have been observed in the crystal form (unpublished observations).

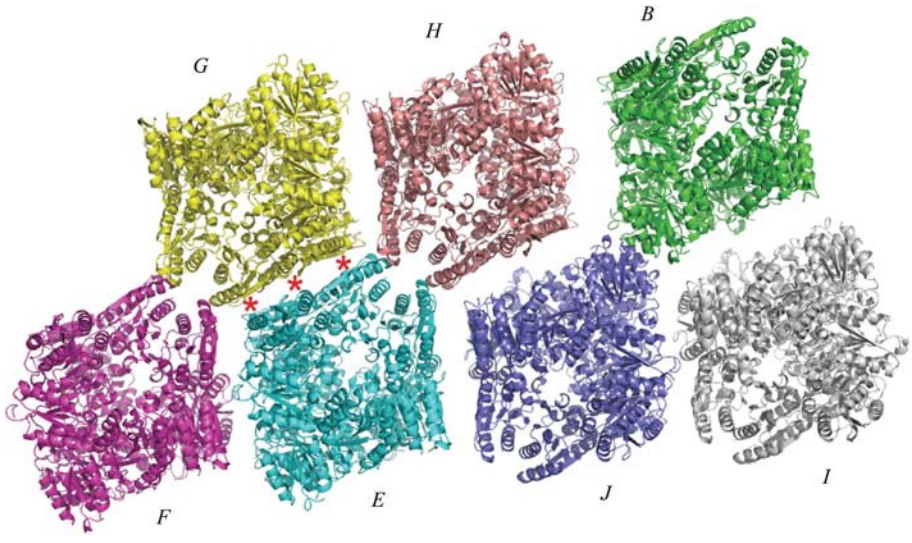

Figure 9
Seven dodecamers from the *S. coelicolor* DHQase–CA1 structure represented as ribbons and coloured by chain. This view of the structure illustrates the two types of crystal contacts seen within the structure. The main interaction, which is made at each of the twofold axes of each dodecamer in the structure, can be seen between chains *F* and *G* and between chains *E* and *H* and a perpendicular view between chains *B* and *F*. With these interactions alone chains *F*, *G*, *B* and *J* (shown here) together with chains *D*, *K*, *M* and *O* form a network of interactions. Chains *E*, *H* and *I* (shown here) together with chains *A*, *C*, *L*, *N* and *P* form a similar network of interactions within the spaces left by the first group of dodecamers. The interactions made between these two groups of dodecamers are *via* the flat threefold face of only certain dodecamers. In this diagram the only interaction is between chain *E* and *G* and is highlighted with three red stars roughly indicating the three interaction sites; all other potential interactions are too distant, *e.g.* between chain *E* and chain *J*. This figure was produced using *PyMOL* (DeLano, 2002).

4.2. Multiple models compared

The refined structure of the *S. coelicolor* DHQase–CA1 inhibitor complex results in 192 independent monomers which can be structurally compared. At 1.7 Å resolution, it is generally considered there are enough observations of the parameters such that each monomer can be considered independently. The NCS-related molecules exist in chemically very similar but not identical

environments; therefore, 16-fold NCS restraints were applied throughout the refinement process, as applying a complete set of NCS restraints was not possible with the current implementation of *REFMAC*5. It would be reasonable to expect very similar values for the geometry of most monomers, while individual chemical environments, for example at the three-fold crystal contacts, could lead to some variations.

It is normal to consider the geometry of a protein structure as a whole compared with ideal values derived from small-molecule databases. However, in this case there are sufficient copies of the protein monomer that individual bond distances and angles can be compared. Fig. 10 shows selected bond distances and angles calculated for two of the three cysteine residues in each monomer using *PROCHECK*. Each distance and angle forms a Gaussian distribution of values centred on the ideal value in most cases. This spread in values is somewhat surprising; however, it is unlikely that it represents true variation between the structures within the crystal rather than the coordinate error [Cruickshank DPI = 0.16 Å (Cruickshank, 1999), 0.09 Å maximum-likelihood (ML) positional parameter, 5.6 $Å^2$ ML thermal parameter (Murshudov & Dodson, 1997)], which is a function of the resolution of the data. The distributions of values do contain overall structural information; for example, the distribution of the Cys40 CB—CA—C bond angles away from the ideal value. Effective removal of the bond-distance and angle restraints while maintaining 16-fold NCS results in a further broadening of the distributions while the overall trend is preserved. The structure is interesting in the light of recent discussions regarding the use of multiple models to more accurately describe either the uncertainty or conformational heterogeneity of protein X-ray structures (DePristo *et al.*, 2004; Furnham *et al.*, 2006). However, here we have multiple copies of the same structure fitted to different regions of the electron density. The errors associated with the X-ray data are hopefully nonsystematic and therefore an enhancement in the signal-to-noise ratio and hence the accuracy of the structure would be expected by considering the ensemble of all 192 structures. Unfortunately, the manipulation of so many NCS-related molecules and the interpretation of the individual models is too laborious at this time to make this a realistic proposition, given the current tools available.

5. Summary

From crystallographic studies of the type II DHQase it is apparent that the molecular symmetry of the quaternary structure is rarely coincident with the crystallographic symmetry. As a result 12, 24, 56 and 192 copies of the molecule have needed to be found by molecular replacement, which fortunately can be carried out with the biological unit of the protein, the dodecamer, vastly simplifying the problem. Molecular replacement with many copies can be problematic when there is translational symmetry. Artificially high correlation coefficients for the first molecule searched for are often a symptom of this, as this molecule satisfies in part the contributions of both the translationally related copies in the structure (Navaza *et al.*, 1998). There is no foolproof solution

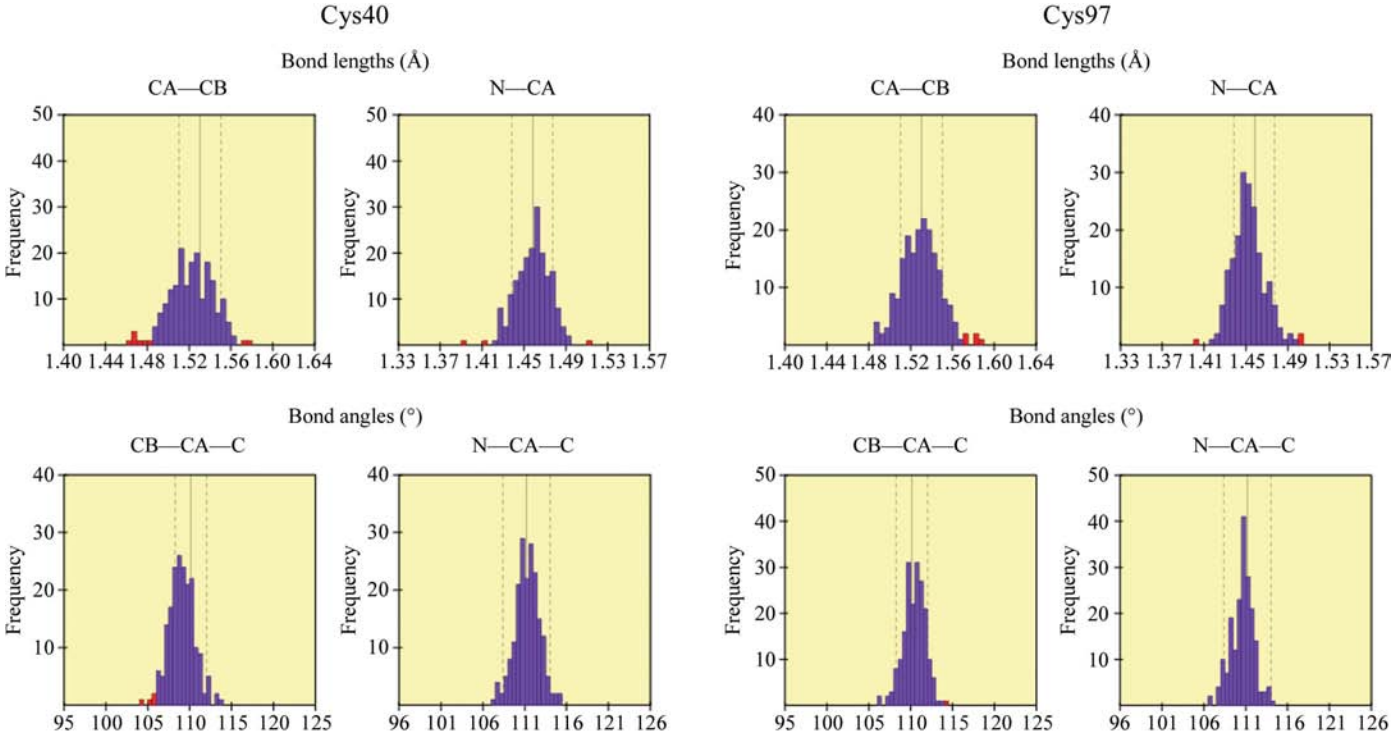

Figure 10
A comparison of specific bond angles and distances for all copies of two cysteine residues present in the *S. coelicolor* DHQase–CA1 structure using *PROCHECK* (Laskowski *et al.*, 1993). The solid and dashed lines represent the mean and standard deviation values as per Engh and Huber small-molecule data.

to counteract this problem, but identification of the translational component in a self Patterson map may permit searching for pairs of molecules or, if the translation results in a doubling of the unit-cell axis, treating the data as if it was in a smaller cell. If there is an indication of the frustrated symmetry from the X-ray data then molecular replacement can in some cases be carried out in the higher symmetry despite poor merging *R*-factor statistics. In these cases, it is hoped that generation of the full structure from this crystallographic symmetry and the application of rigid-body refinement will be sufficient to produce a reasonable starting model. Molecular replacement with many copies is not necessarily problematic, as refinement with NCS and the use of averaged electron density means that a partial solution may be sufficient to solve the structure. However, not all molecular-replacement packages handle multiple models well and, as is shown here, the time taken to obtain a solution can be a significant issue. Therefore, when considering choices about the search model and the resolution range of the X-ray data to use it is also important to think carefully about the choice of molecular-replacement program.

The large *P*1 *S. coelicolor* DHQase–CA1 inhibitor complex described here with so many copies in the asymmetric unit is an oddity which we have shown represents an extreme example of frustrated $F4_132$ symmetry. It is highly likely that other examples of structures such as this have been encountered in the past and have been discarded as intractable owing to their size. The P1 *S. coelicolor* DHQase–CA1 inhibitor complex proves that problems such as this are tractable, if at the limit of current molecular-replacement methods, and that useful information can be extracted from these structures.

We thank Jorge Navaza for useful discussions and the Biotechnology and Biological Sciences Research Council for studentships to DAR and KAS. DAR received support from the GlaxoSmithKline Action TB programme for this work.

References

Abell, C. (1999). *Comprehensive Natural Products Chemistry*, edited by U. Sankawa, pp. 573–607. Amsterdam: Elsevier.
Carson, M. (1991). *J. Appl. Cryst.* **24**, 958–961.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.
DeLano, W. L. (2002). *The PyMOL Molecular Graphics System.* http://www.pymol.org.
DePristo, M. A., de Bakker, P. I. W. & Blundell, T. L. (2004). *Structure*, **12**, 831–838.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Furnham, N., Blundell, T. L., DePristo, M. A. & Terwilliger, T. C. (2006). *Nature Struct. Mol. Biol.* **13**, 184–185.
Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.
Gourley, D. G., Shrive, A. K., Polikarpov, I., Krell, T., Coggins, J. R., Hawkins, A. R., Isaacs, N. W. & Sawyer, L. (1999). *Nature Struct. Biol.* **6**, 521–525.
Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.
Kleywegt, G. J. (1996). *Acta Cryst.* D**52**, 842–857.
Lamzin, V. S. & Wilson, K. S. (1997). *Methods Enzymol.* **277**, 269–305.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Lee, S., Sawaya, M. R. & Eisenberg, D. (2003). *Acta Cryst.* D**59**, 2191–2199.
McCoy, A. J. (2007). *Acta Cryst.* D**63**, 32–41.
McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Cryst.* D**61**, 458–464.
Murshudov, G. N. & Dodson, E. A. (1997). *CCP4 Newsl.* **33**, 31–39.
Murshudov, G. N., Dodson, E. J. & Vagin, A. A. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore & S. Bailey, pp. 93–104. Warrington: Daresbury Laboratory.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Navaza, J., Panepucci, E. H. & Martin, C. (1998). *Acta Cryst.* D**54**, 817–821.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Payne, R. J., Peyrot, F., Kerbarh, O., Abell, A. D. & Abell, C. (2007). *ChemMedChem*, **2**, 1015–1029.
Prazeres, V. F. V., Sanchez-Sixto, C., Castedo, L., Lamb, H., Hawkins, A. R., Riboldi-Tunnicliffe, A., Coggins, J. R., Lapthorn, A. J. & Gonzalez-Bello, C. (2007). *ChemMedChem*, **2**, 194–207.
Robinson, D. A., Stewart, K. A., Price, N. C., Chalk, P. A., Coggins, J. R. & Lapthorn, A. J. (2006). *J. Med. Chem.* **49**, 1282–1290.
Roszak, A. W., Robinson, D. A., Krell, T., Hunter, I. S., Frederickson, M., Abell, C., Coggins, J. R. & Lapthorn, A. J. (2002). *Structure*, **10**, 493–503.
Sanchez-Sixto, C., Prazeres, V. F. V., Castedo, L., Lamb, H., Hawkins, A. R. & Gonzalez-Bello, C. (2005). *J. Med. Chem.* **48**, 4871–4881.
Sayle, R. A. & Milnerwhite, E. J. (1995). *Trends Biochem. Sci.* **20**, 374–376.
Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O. & Abola, E. E. (1998). *Acta Cryst.* D**54**, 1078–1084.
Tong, L. & Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 783–792.
Toscano, M. D., Frederickson, M., Evans, D. P., Coggins, J. R., Abell, C. & Gonzalez-Bello, C. (2003). *Org. Biomol. Chem.* **1**, 2075–2083.
Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.
Vagin, A. A. & Isupov, M. N. (2001). *Acta Cryst.* D**57**, 1451–1456.
Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.
Vagin, A. & Teplyakov, A. (2000). *Acta Cryst.* D**56**, 1622–1624.
Vonrhein, C. & Schulz, G. E. (1999). *Acta Cryst.* D**55**, 225–229.