BIOLOGICAL CRYSTALLOGRAPHY

ISSN 1399-0047

Combining phase information in reciprocal space for molecular replacement with partial models

Claudia Millán,[a] Massimo Sammito,[a] Irene Garcia-Ferrer,[a] Theodoros Goulas,[a] George M. Sheldrick[b] and Isabel Usón[c]*

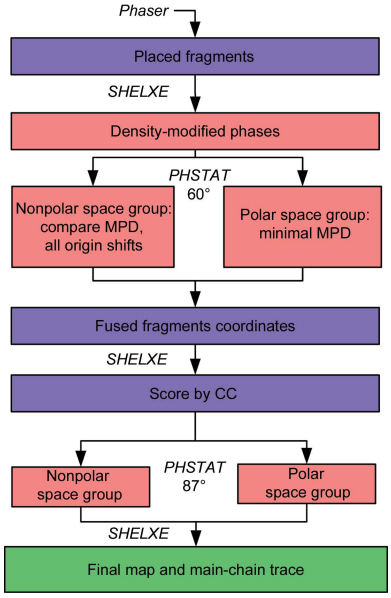
[a]Structural Biology, Instituto de Biologia Molecular de Barcelona, Carrer Baldiri Reixac 15, 3 A17, 08028 Barcelona, Spain, [b]Structural Chemistry, Institut für Anorganische Chemie, University of Göttingen, Tammannstrasse 4, 37077 Göttingen, Germany, and [c]Structural Biology, ICREA at IBMB–CSIC, Carrer Baldiri Reixac 13–15, 08028 Barcelona, Spain. *Correspondence e-mail: uson@ibmb.csic.es

Received 30 January 2015
Accepted 8 July 2015

Edited by R. J. Read, University of Cambridge, England

Keywords: phasing; *ARCIMBOLDO*; clustering; *ab initio*; small fragments; molecular replacement.

ARCIMBOLDO allows *ab initio* phasing of macromolecular structures below atomic resolution by exploiting the location of small model fragments combined with density modification in a multisolution frame. The model fragments can be either secondary-structure elements predicted from the sequence or tertiary-structure fragments. The latter can be derived from libraries of typical local folds or from related structures, such as a low-homology model that is unsuccessful in molecular replacement. In all *ARCIMBOLDO* applications, fragments are searched for sequentially. Correct partial solutions obtained after each fragment-search stage but lacking the necessary phasing power can, if combined, succeed. Here, an analysis is presented of the clustering of partial solutions in reciprocal space and of its application to a set of different cases. In practice, the task of combining model fragments from an *ARCIMBOLDO* run requires their referral to a common origin and is complicated by the presence of correct and incorrect solutions as well as by their not being independent. The *F*-weighted mean phase difference has been used as a figure of merit. Clustering perfect, non-overlapping fragments dismembered from test structures in polar and nonpolar space groups shows that density modification before determining the relative origin shift enhances its discrimination. In the case of nonpolar space groups, clustering of *ARCIMBOLDO* solutions from secondary-structure models is feasible. The use of partially overlapping search fragments provides a more favourable circumstance and was assessed on a test case. Applying the devised strategy, a previously unknown structure was solved from clustered correct partial solutions.



1. Introduction

ARCIMBOLDO (Rodríguez *et al.*, 2009) implements a phasing method based on the location of model fragments with *Phaser* (McCoy *et al.*, 2007) and density modification with *SHELXE* (Sheldrick, 2002) in a multisolution frame, exploiting the computational power of a grid. A modern multicore machine may be used to solve simple cases with *ARCIMBOLDO_LITE* (Sammito *et al.*, 2015). Typically, the most successful model fragments used in *ARCIMBOLDO* are secondary-structure elements such as polyalanine α-helices (Rodríguez *et al.*, 2012) predicted from the sequence. This constitutes an *ab initio* phasing approach, since no specific structural knowledge about the target structure is required. The presence of helices can be predicted from the experimental data intensity distribution (Morris *et al.*, 2004) and the Patterson function (Caliandro *et al.*, 2012). Other previous phasing approaches have made use of model helices (Glykos & Kokkinidis, 2003) or nucleic acid bases (Robertson & Scott, 2008; Robertson *et al.*, 2010; Pröpper *et al.*, 2014). The

rationale underlying the method is the substitution of the atomicity constraint central to dual-space recycling methods (Sheldrick *et al.*, 2011) by the enforcement of secondary-structure or tertiary-structure constraints. A correctly located fraction of the structure, as small as 10% of the main chain, that is very accurately placed (r.m.s.d. of around 0.5 Å) provides sufficient information to phase the whole structure at a resolution of up to 2 Å when further constrained by density modification (Yao *et al.*, 2005, 2006; Burla *et al.*, 2010, 2012). The same approach allows the exploitation of the prior knowledge derived from having a low-homology model, which even if unsuccessful in molecular replacement may contain a fold or folds that are significantly similar to the target structure. Fragments from such a low-homology model can be selected by evaluating the LLG improvement in the rotation function in *Phaser* (Storoni *et al.*, 2004) reached by systematically omitting different parts of the structure. The template is accordingly trimmed and used in an *ARCIMBOLDO_SHREDDER* run (Sammito *et al.*, 2014). A further extension to the use of small folds in an *ab initio* approach is found in *ARCIMBOLDO_BORGES* (Sammito *et al.*, 2013), which takes advantage of the fact that there is a wealth of structural knowledge in the >100 000 structures deposited in the PDB (Bernstein *et al.*, 1977; Berman *et al.*, 2000). Given sufficiently small fragments, such as a few helices or strands, similar fragments to the folds present in our unknown structure should be represented in other structures deposited in the PDB. If a library of similar local folds is generated, it can be exhaustively evaluated and refined through the results of a rotation search. The best-scoring models are then used as *ARCIMBOLDO* search fragments. All of these programs run on a single multicore machine or from a workstation sending jobs to a local or remote grid or supercomputer and are available from http://chango.ibmb.csic.es (Millán *et al.*, 2015).

The secondary-structure model fragments used for *ab initio* phasing purposes typically amount to 3–4% of the main chain, so that two or three, or sometimes four, need to be correctly placed in order to provide a substructure that makes up 10% of the main chain and is susceptible to being expanded into the complete solution. Currently, this is performed by sequentially locating each fragment and combining the structural hypotheses resulting from each fragment search with the next round. With each additional fragment, the number of solutions to pursue increases exponentially. Instead, a set of single, correct partial solutions obtained after a first fragment-search run could succeed if combined, even though when isolated they would lack the minimum phasing power. This would avoid the need to compute subsequent searches, which exponentially increase the number of calculations to be performed. Another appealing scenario is the use of partial models from close to remote homologues representing a reduced fraction of the total scattering mass. In this case, even though the models are more complete than bare local folds or secondary-structure elements, their structure may not match the target as accurately as required for phasing and a larger percentage of the structure may be needed. Also, structural differences may accumulate, so a combination of slightly differently placed fragments from derived phase sets may model the geometrical deviation. A previous feasibility study on cluster analysis of molecular-replacement solutions (Buehler *et al.*, 2009) established how an average phase set derived from multiple partial solutions may be more precise than the individual phase sets. Similar to *ARCIMBOLDO*, implementations aiming to solve the phase problem from modelled structural hypotheses, such as *AMPLE* (Rigden *et al.*, 2008; Bibby *et al.*, 2012) and *mr_rosetta* (DiMaio *et al.*, 2011), could profit from combining the results of correct solutions as far as they can be identified within a large pool. The fragmentation and combination in real space of *Rosetta ab initio* models (Qian *et al.*, 2007) used as search models has been proposed (Shrestha *et al.*, 2011, 2015). In turn, *phaser.MRage* combines molecular-replacement models in real space (Bunkóczi *et al.*, 2013). Assessment of the agreement among electron-density maps from the same or different models has been discussed by Urzhumtsev *et al.* (2014).

Within the cases aiming to establish a partial structure to be expanded by density modification, the task is to extract a weak signal from a very noisy landscape. Even combinations of phases derived from correct fragments will yield mean phase differences (MPDs) that are barely a few degrees better than random (90°). Recognition is enhanced for models with a partial overlap, as a result of lower MPDs.

Combining phases correctly implies relating them to the same crystallographic origin, since solutions do not have an absolute reference. This generates a space-group dependency in the way that relative origins are constrained. For polar space groups the number of possible origins is unlimited, but for the case of nonpolar space groups with limited origins all possible origins can be exhaustively tested, in attempts either to recognize the common origin or to phase blindly from combined solutions. This implies that regardless of whether the correct origin can be identified, for nonpolar space groups it should still be possible to match fragment pairs by trying

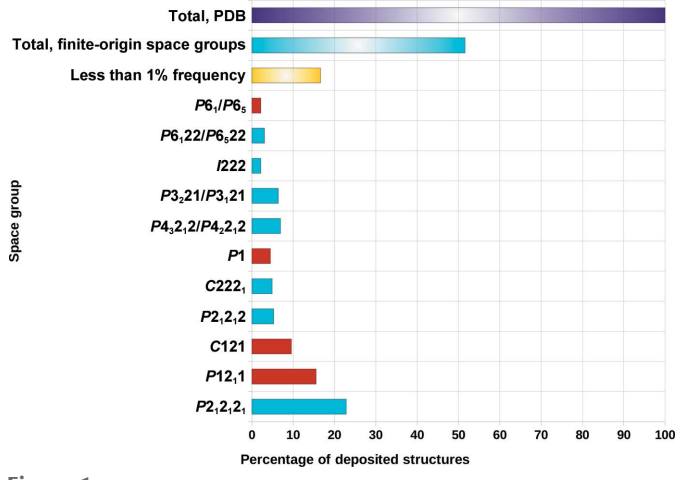


Figure 1
Space-group distribution within the approximately 105 000 crystallographic entries deposited in the PDB as of January 2015. Nonpolar space groups are shown in blue and polar space groups in red. Nonpolar space groups make up half of the total. $P2_12_12_1$ is the most common space group, occurring in almost one quarter of the deposited entries.

Table 1
X-ray data-collection and refinement statistics for all structures used.

Values in parentheses are for the highest resolution 0.1 Å shell. Statistics were calculated using *XPREP* (Sheldrick, 2008). Refinement information was obtained from the deposited PDB entries.

	3gwh	2iu1	In-house xylose isomerase (1mnz)	3o55	3jvl	2y8p	PPAD
Data collection							
Space group	$P2_1$	$P2_12_12_1$	$I222$	$C222_1$	$P2_12_12$	$C222_1$	$P2_12_12_1$
Unit-cell parameters							
a (Å)	37.39	32.14	92.89	50.85	52.06	123.32	58.631
b (Å)	65.75	70.25	98.46	76.57	73.05	183.93	60.357
c (Å)	38.19	81.70	102.68	62.31	32.30	35.29	113.884
$\alpha = \gamma$ (°)	90.0	90.0	90.0	90.0	90.0	90.0	90.0
β (°)	109.6	90.0	90.0	90.0	90.0	90.0	90.0
Resolution (Å)	24.00–1.95	35.32–1.78	21.16–1.54	38.29–1.90	19.53–1.20	40.12–1.99	53.33–1.50
$\langle I/\sigma(I)\rangle$	18.9 (2.5)	20.7 (7.9)	30.6 (8.8)	34.53 (4.89)	40.87 (9.22)	12.90 (3.59)	31.62 (5.51)
Completeness (%)	99.8 (98.8)	97.6 (88.3)	99.8 (99.2)	98.4 (92.3)	99.7 (99.9)	99.6 (97.6)	99.1 (95.6)
Refinement							
No. of reflections	12832 (893)	18464 (961)	61001 (3196)	9758 (1272)	39262 (8263)	27781 (3313)	64785 (10707)
R_{work}†/R_{free} (%)	19.5/24.1 (31.8/25.4)	24.3/28.8 (25.7/30.4)	19.0/22.6 (19.7/23.5)	18.2/26.7 (27.4/34.0)	11.6/4.5 (15.3/20.49)	20.30/25.18 (23.5/29.56)	15.69/17.71 (17.83/24.69)
No. of protein atoms	1680	1473	3017	1170	987	2915	3381
B factors (Å^2)							
Protein main chain	27.0	21.8	21.2	16.23	11.38	20.88	20.39
Protein side chain	25.3	20.3	19.8	21.37	18.18	24.95	22.78

† $R_{work} = \sum_{hkl} \big||F_{obs}| - |F_{calc}|\big| / \sum_{hkl} |F_{obs}|$, where F_{obs} and F_{calc} are the structure-factor amplitudes from the data and the model, respectively. R_{free} is the same as R_{work} but calculated with a 5% test set.

all possible origin shifts. More than 50% of the structures deposited in the PDB belong to nonpolar space groups, with the most frequent space group for proteins, $P2_12_12_1$, also being nonpolar. Fig. 1 displays the space-group distribution in the PDB.

The main goal of the work presented in this study is to assist structure solution by the combination in reciprocal space of the phase information from partial solutions generated by multisolution *ARCIMBOLDO* procedures. Partial solutions constitute approximations to the complete structure and their phases tend to the true phases. Based on this, the identification of relative origin shifts among fragments is tested in ideal and real cases, evaluating their relative MPDs calculated as *F*-weighted mean phase errors (MPEs) against the average (Lunin & Woolfson, 1993). The results of this analysis are exploited in the recognition at an early stage or in difficult cases of combinations of phase sets that successfully develop into a solution. Finally, the solution of a previously unknown structure relying on the proposed procedure is described.

2. Experimental

2.1. Computing setup

Structure solutions and tests were run on a local Condor (Tannenbaum *et al.*, 2002) grid made up of 120 nodes totalling 175 gigaflops. A workstation with an Intel Core i7-2600 processor and 12 GB RAM running Debian 6 was used for local computations and as a submitter to the Condor grid.

2.2. Software versions

Analyses were performed by *ALIXE*, programmed in Python, and *PHSTAT*, programmed in Fortran 77. *SHELXE* (Sheldrick, 2010) is required to provide density modification based on the sphere-of-influence algorithm (Sheldrick, 2002) and for phase extension. The program allows appropriate values to be specified or defaults to be accepted for a series of parameters including the number of cycles of density modification, the resolution cutoff for starting phases or data and the solvent content. In this work, all parameters have been left at their default values except for the solvent content particular to each structure and the number of cycles of density modification. *PHSTAT* performs clustering of phase sets by a cyclical procedure. It takes a set of phase files in `.phs` format as input and sets one of them as a reference. Then, for nonpolar space groups, it applies all of the allowed origin shifts to the phases and calculates the *E*- or *F*-weighted mean phase error (MPE) for each case. For polar space groups, the allowed discrete origin shifts are tested and an initial origin shift is estimated in the polar direction using the layer of index 1 and is refined against all reflections (Lunin & Lunina, 1996). Keeping the shifts with the lowest MPE, weights for each phase set are adjusted to minimize the MPE to the combined set until convergence. Customizable parameters are the selection of amplitudes or normalized amplitudes, the number of cycles (the default is three), the reference file for clustering (the default is the highest syntheses correlation coefficient), the resolution limit for the phase sets (the default is 2 Å) and the tolerance in degrees for the MPD between the sets of phases to be clustered. If a file containing reference phases in fcf format (Sheldrick, 2008) is provided, it also calculates the MPE *versus* this reference, for example the final refined structure.

ARCIMBOLDO (Millán *et al.*, 2015) was used relying on *Phaser* versions from 2.5 to the current 2.7 used through either the *CCP*4 (Winn *et al.*, 2011) or *PHENIX* (Adams *et al.*, 2010)

distributions. Default *Phaser* thresholds were used to select partial solutions. *Coot* (Emsley *et al.*, 2010) and *PyMOL* (v.1.5.0.4; Schrödinger) were used for graphical examinations and figures.

2.3. Test data

The characteristics of all of the test data used in this study are summarized below and relevant statistics are given in Table 1.

2.3.1. PRDII data. PRDII (PDB entry 3gwh) is a transcriptional antiterminator of the BglG family from *Bacillus subtilis*, which was solved *ab initio* with *ARCIMBOLDO* (Rodríguez *et al.*, 2009). The crystals belonged to space group $P2_1$, with unit-cell parameters $a = 37.39$, $b = 65.75$, $c = 38.19$ Å, $\beta = 109.58^\circ$. The asymmetric unit contains two copies of the monomer with 111 residues and 40% solvent content, although the solvent content was deliberately increased to 45% in *SHELXE* runs. The data resolution is 1.95 Å. The structure comprises ten α-helices ranging from 11 to 20 residues.

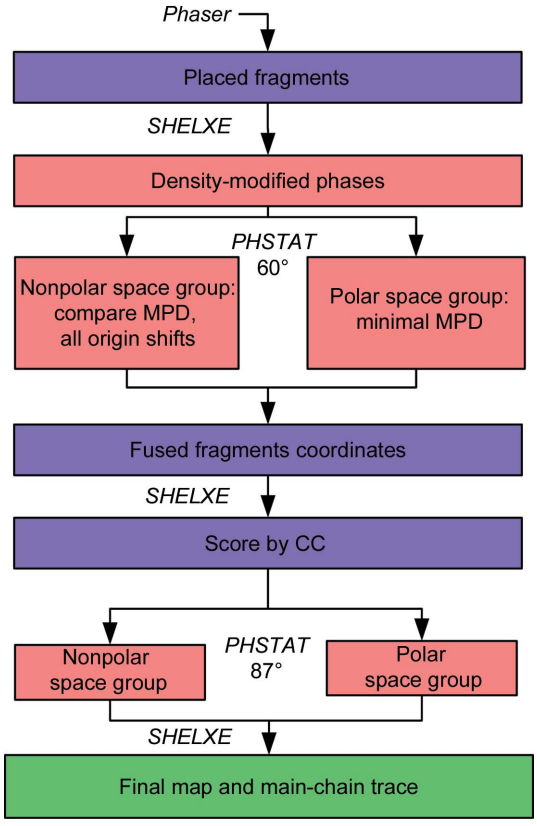


Figure 2
Flow of the *ALIXE* procedure for the combination of fragment solutions. Phase sets derived from the placed fragments with *Phaser* (or tests with perfect fragments) are subjected to density modification with *SHELXE*. A first clustering round with a low MPD threshold (<60°) allows similar solutions to be fused with partial overlap. For nonpolar space groups differences among all possible origin shifts are computed as an additional figure of merit. New phase sets are computed and modified from the fused coordinates. The solutions are sorted by CC and a second round of MPD evaluation with a threshold of 87° is run in order to identify complementary solutions. The resulting clusters are further subjected to density modification and autotracing to solve the structure.

2.3.2. EIF5 data. Crystals of the C-terminal part (residues 232–431) of eukaryotic translation factor 5 (EIF5) were obtained in space group $P2_12_12_1$, with unit-cell parameters $a = 32.23$, $b = 71.08$, $c = 80.64$ Å. The asymmetric unit contains one monomer of 185 residues and 42% solvent content, which was set to 45% in *SHELXE*. Data to 1.67 Å resolution were available. The structure (PDB entry 2iu1) was originally solved by experimental phasing (Bieniossek *et al.*, 2006) and contains ten α-helices ranging from nine to 19 residues.

2.3.3. Xylose isomerase data. The xylose isomerase from *Streptomyces rubiginosus* is a TIM-barrel protein for which in-house data were available to 1.54 Å resolution. The space group is $I222$ and the unit-cell parameters are $a = 92.89$, $b = 98.46$, $c = 102.68$ Å. The asymmetric unit contains a monomer of 388 residues, with 50% solvent content. A structure of the same crystal form with data to 0.99 Å resolution was deposited in the PDB as entry 1mnz (Carrell *et al.*, 1994) and contains 15 α-helices ranging from five to 27 residues.

2.3.4. Brd4 data. The structure of the P-TEFb-activating protein Brd4 from *Mus musculus* (Vollmuth *et al.*, 2009) has been deposited in the PDB as entry 3jvl and data are available to 1.2 Å resolution. The space group is $P2_12_12$ and the unit-cell parameters are $a = 52.06$, $b = 73.05$, $c = 32.30$ Å. The asymmetric unit contains a monomer of 120 residues, with 44% solvent content.

2.3.5. ALR-MIA40 data. Crystals of the human augmenter of liver regeneration protein (Banci *et al.*, 2011), for which data are available to 1.9 Å resolution (PDB entry 3o55), belonged to space group $C222_1$, with unit-cell parameters $a = 50.85$, $b = 76.57$, $c = 62.31$ Å. The asymmetric unit contains a monomer of 125 residues, with 39% solvent content.

2.3.6. MltE data. MltE (PDB entry 2y8p) is a bacterial outer membrane-anchored endolytic peptidoglycan lytic transglycosylase (Artola-Recolons *et al.*, 2011). Diffraction data to 2.0 Å resolution were available. The crystals belonged to space group $C222_1$, with unit-cell parameters $a = 123.32$, $b = 183.93$, $c = 35.29$ Å. They contained two copies of the 194 amino-acid MltE monomer in the asymmetric unit, corresponding to a solvent content of 45%.

2.3.7. PPAD data. PPAD is a peptidylarginine deiminase from *Pseudomonas gingivalis* (Goulas *et al.*, 2015). 20 diffraction data sets from different crystals were available, ranging from 2.97 to 1.5 Å resolution. 16 of these, with unit cells of similar dimensions and rendering an average $R_{\rm int}$ of 0.37 and R_σ of 0.02, were combined. The crystals belonged to space group $P2_12_12_1$ and contained one copy of the 432-amino-acid monomer in the asymmetric unit, corresponding to a solvent content of 40%, which was set to 50% in *SHELXE* runs.

3. Results and discussion

Combining phase information from partial solutions in reciprocal space has been studied in four different scenarios that will be described and discussed separately. A study on perfect fragments extracted from the final deposited structure illus-

trates the intrinsic difficulties even in this best-case scenario. This is followed by the use of model helices located in an *ARCIMBOLDO* run on several test cases. A test case on the combination of larger fragments with errors derived from a low-homology model is then examined and, finally, the solution of a previously unknown structure that required the combination of fragments from a structure with a similar fold is discussed.

Calculations were implemented in a procedure named *ALIXE*, which was developed for this purpose. The general workflow in *ALIXE* is illustrated in Fig. 2 and can be outlined as follows.

(i) The placement of search fragments with *Phaser*. The tests described for perfect fragments start by extracting them from the final deposited structure.

(ii) *SHELXE* generation of phase sets from any set of roto-translated fragments and the application of density modification to enhance origin recognition in step (iii).

(iii) The use of *PHSTAT* for iterative clustering of phase sets with relative mean phase differences (MPDs) below 60°. This depends on the correct determination of the relative origin shifts between phase sets.

(iv) Application of the corresponding shifts to the fragments and real-space merging of coordinates.

(v) Calculation of a correlation coefficient for the merged fragments as a figure of merit.

(vi) Selection of the top merged coordinates around each rotation peak.

(vii) Clustering within the 87° MPD threshold using the phases from selected sets against all other rotation phase clusters as a reference.

(viii) Density modification and auto-tracing with *SHELXE* to expand from the combined phases.

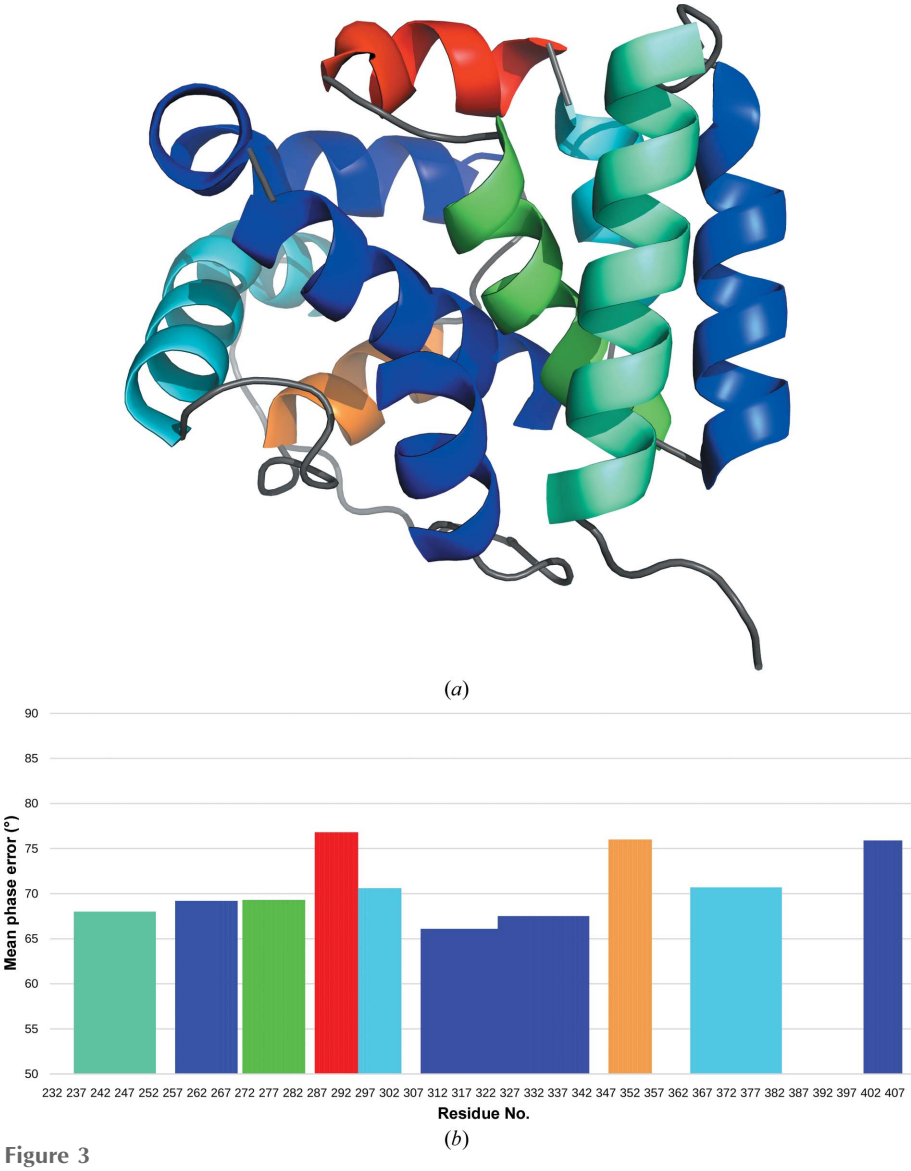


Figure 3
Characterization of the ten helices in the structure of PDB entry 2iu1. (*a*) Cartoon representation, with a rainbow colour gradient representing the main-chain average *B* factor of the helices (red for highest and blue for lowest *B*). (*b*) MPE of the phase set obtained from each helix *versus* residue count, colour coded according to the *B* values as in (*a*). The results are comparable except for two of the smallest helices.

3.1. Perfect helices from test structures

Main-chain α-helices are the most successful search fragments for *ab initio* phasing with *ARCIMBOLDO*, as their presence can be predicted from the sequence and they are rigid, constant and nearly ubiquitous. In order to explore the combination of perfect, yet partial, solutions, all helical fragments were extracted from three test structures (PDB entries 3gwh, 2iu1 and 1mnz). 2iu1 belongs to the most frequent among all Sohncke space groups (nonpolar $P2_12_12_1$), while 3gwh adopts the most common polar space group for proteins, $P2_1$. 1mnz, a TIM barrel in the nonpolar space group $I222$, does not have an all-helical composition as in the other two cases but contains a substantial fraction of β-strand.

In the tests, phases were calculated from each of the helices in the structure after truncating residues to alanine and setting the *B* factors to a common value. Phase sets were subjected to a variable number of density-modification cycles with *SHELXE*: 0, 5, 10, 15, 20 and 30 cycles. All of the remaining parameters were set to constant values, adopting the program defaults, apart for the solvent content, which was particular to each structure. Subsequently, all possible phase combinations derived from two, three and four fragments were calculated, determining the origin shifts leading to minimal phase differences.

As all fragments are extracted from the final structures, a relative origin shift signals a failure in the clustering process. For nonpolar space groups, with a limited number of origin choices, the MPD can be calculated for phase combinations at all possible allowed shifts. This allows the difference between the lowest and second-lowest MPD obtained to be examined in order to assess whether a large difference would indicate a more reliable discrimination of the origin shift.

The phase sets resulting from these combinations were subjected to further density-modification cycles (0, 5, 10, 15, 20 and 30). Evidently, at this stage the relative origin shift has been fixed, but density modification might still improve the *a posteriori* discrimination of correct *versus* incorrect shifts.

3.1.1. An all-helical structure in the most frequent nonpolar Sohnke space group: 2iu1. The structure of the carboxy-terminal domain of human translation initiation factor EIF5 (Bieniossek *et al.*, 2006) is displayed in Fig. 3(*a*). It contains ten helices, the extension, average *B* values and mean phase errors (MPEs) of which relative to the deposited structure are represented in Fig. 3(*b*).

For all possible binary combinations of helical fragments, adopting either fragment in the pair as the reference, the relative *F*-weighted MPDs considering the eight origin shifts allowed in the space group were computed. The values ranged from 81.3 to 89.9°. A syntheses correlation coefficient (CC; Fujinaga & Read, 1987) was calculated for each file resulting from fusing the fragments after applying the selected origin shift. The fragment selected as a reference may influence the outcome, but in general the results are consistent within each pair of fragments and are thus displayed for a sparse set of all possible combinations. The effect of density modification on the discrimination between the two lowest MPDs for each pair of fragments is illustrated in Fig. 4. Negative bars indicate those cases where the correct origin shift would be missed as it does not correspond to the minimum MPD.

The correct origin shift would be unequivocally chosen in 39 of the 45 pairs, and would be chosen in 42 if the fragment in the pair characterized by the higher CC against the native data, usually the largest, was trusted as a reference in case of discrepancy. This would seem to be the natural choice and solutions from fragment search are sorted according to this FOM. The use of density modification enhanced

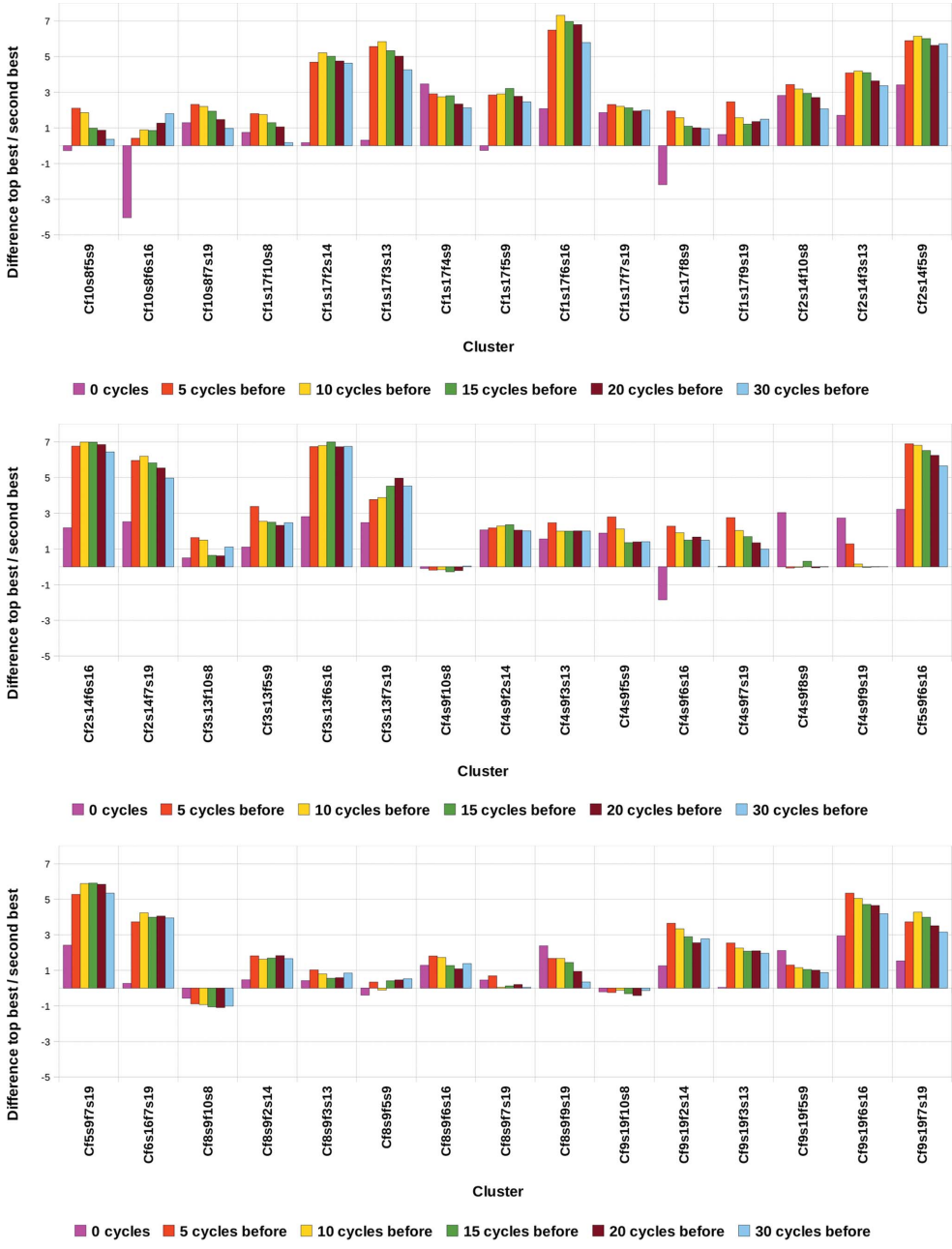


Figure 4
MPD difference between clusters of perfect helices from 2iu1 for the two origin shifts corresponding to the lowest and second-lowest MPD. Fragments are labelled by their order in Fig. 3 and the number of amino acids. The bars show the difference in degrees between the MPD corresponding to the correct origin shift and the MPD for the best-scoring wrong origin shift. Negative bars represent cases where a wrong origin shift yields the minimal MPD and would have been selected instead of the correct one. Density modification assists in the selection of the correct origin shift in six cases and generally improves the discrimination of the correct origin choice (pink bars *versus* the rest).

discrimination; four cases actually required it in order to avoid selecting an incorrect origin shift, but the number of density-modification cycles did not determine the outcome. The three cases where the correct origin shift was missed involve the smallest helix in the structure as a common fragment in combination with two other small helices and a distorted helix. Still, there are some clusters that even if correctly matched present absolute MPD differences among different origin choices of below 1°, indicating that the correct shift could have been accidentally chosen but would hardly be trusted if the structure was unknown. This occurs in other matches involving the smallest helices. Also, those cases where discrimination is clearer tend to contain common fragments, such as the long sixth helix, yielding the lowest MPE (Fig. 3*b*).

As density modification helped to reveal the common origin, the effect of the number of cycles on the resulting MPE *versus* the final structure was assessed (Fig. 5). The general trend, even when the phase information derived from fragment pairs is not yet sufficient to solve the structure, is that the application of five cycles of density modification prior to phase combination renders the lowest MPE (43 of 45 cases). Exceptions correspond to a very small MPD difference value and involve helices of barely ten residues in length. Alternatively, using the *E*-weighted, rather than the *F*-weighted, MPDs to cluster these perfect fragments led to the correct origin shift being identified in all cases.

In the case of ternary combinations, two relative origin shifts need to be determined. Clustering all possible combinations, considering each of the three fragments in a set as the reference in turn, generates 360 sets. Density modification enhances the correct origin discrimination from 254 to 304 sets, representing 60 and 80% of the cases, respectively. Failure occurs only for clusters containing at least one of the three smallest helices of nine amino acids. Fig. 3(*b*) shows that these fragments render the phase sets with the highest MPEs. In practice, such small fragments are of limited use within *ARCIMBOLDO*, as their correct location would fail. Clustering using the best fragments in each triplet as a reference increases the success rate. Sorting fragments according to their CC and taking the top CC for a triplet as a reference would appear to be advisable, as fragments characterized by better FOMs (LLG or CC) are more likely to be correct. Consistent clustering of a triplet when alternative references are used and discrimination between the top and second MPD obtained for all allowed origin shifts can be used to evaluate confidence in the origin choice. Fig. 6 plots graphical results for combinations of phases derived from three fragments subjected or not subjected to density modification. Helix 8, which was found to fail in most clustering tests, yields a negative CC (−0.14) and a high MPE and is omitted from the figure. In all of the remaining cases, applying density modification prior to clustering enhances origin-shift recognition.

In the case of quaternary combinations, origin recognition succeeds every time and for all conditions the MPEs decrease. As more correct fragments are added, the phases improve and the determination of the correct origin becomes easier; as expected, the structure can be solved from so many correct fragments. The general trend of optimal results when applying five cycles of density modification prior to clustering is maintained for helices of a minimum length of ten. Possibly, the early discrimination of protein and solvent regions brought about by density modification enhances correct clustering.

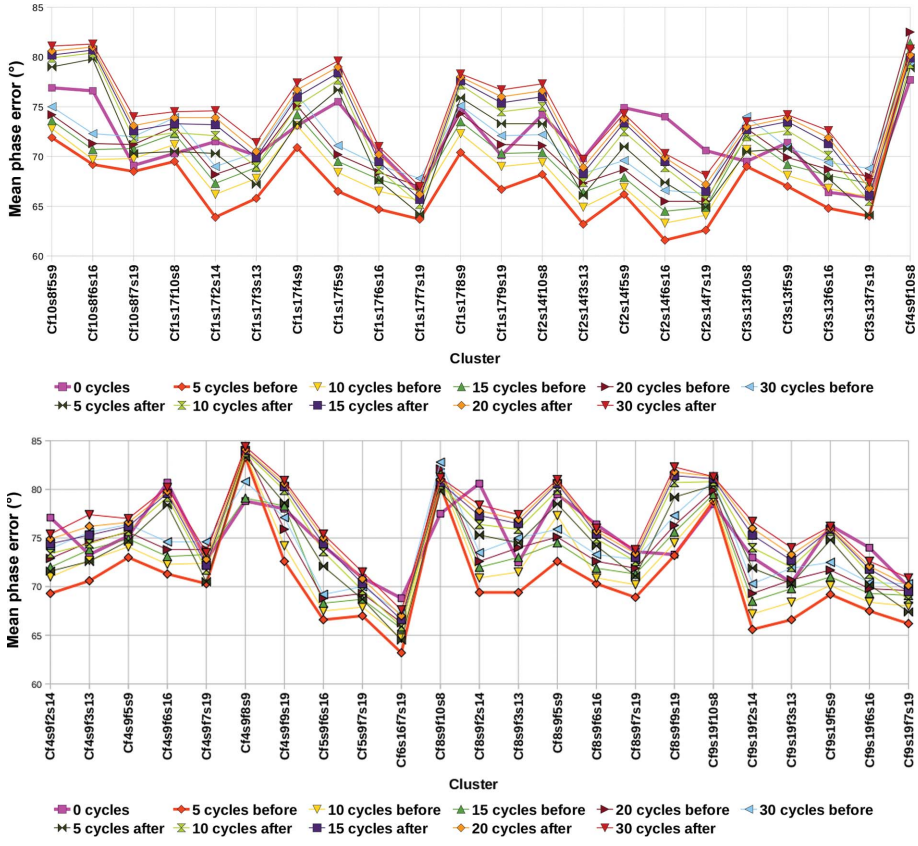


Figure 5
Effect of density modification applied at different stages on the MPE of clusters combining perfect helices extracted from PDB entry 2iu1 with respect to the true phases as calculated from the deposited structure. The orange line, representing five cycles of density modification, always leads to the best phases with the lowest MPE.

3.1.2. The TIM-barrel structure of xylose isomerase: 1mnz. This test case uses an in-house data set to 1.5 Å resolution, which does not correspond to the deposited structure, which is associated with 0.99 Å resolution data. Xylose isomerase is made up of 387 residues and the crystals belonged to the nonpolar space group *I*222, where four different origins may be chosen. The structure, *B* factors and MPE of the

phases derived from each of its helical stretches are shown in Fig. 7. It contains 14 helices of sizes ranging from five to 27 residues, so that 182 binary combinations, corresponding to 91 pairs of helices, may be evaluated. Fig. 8 shows the origin-shift discrimination for all unique pairs.

Again, density modification enhances the origin-shift discrimination, leading to the correct value in 77 pairs. Three more correct cases are unmatched when selecting the smallest fragment in the pair as a reference. As in the previous case, incorrect identification of the origin shift is only seen in the case of small helices of below ten amino acids. Wrong matches cluster towards the end of the table as ordered by the CC of the joint fragments, MPD or discrimination by MPD differences for the best/second-best scoring shifts. The relative origin is clearest for the largest helices and improves upon density modification.

Results for the MPE to the final structure are shown in Fig. 9. In the majority of the clusters, five cycles of density modification before clustering yield the lowest MPE value. There are nine cases for which none of the density-modification conditions improve the MPE, and all of them contained small helices.

In the case of ternary and quaternary combinations of helices, recognition of the correct origin succeeds every time, presumably as a result of the greater information content. Indeed, for all conditions the MPEs decrease. The general trend of optimal results upon the application of five cycles before clustering except for small helices is maintained.

As this structure contains a substantial percentage of β-strands, four partially overlapping polyalanine three-stranded fragments of 14–17 amino acids were extracted as well (Fig. 8*a*). Single-stranded stretches are too short to be of practical use in *ARCIMBOLDO*, but libraries of three-stranded fragments have successfully been used in *ARCIMBOLDO_BORGES* (Sammito *et al.*, 2013). Phases derived from the four perfect fragments were subjected to five cycles of density modification and combinations of two fragments were tested. As the sheet fragments are overlapping, they are readily clustered, with MPDs ranging from 46 to 70°. Particularly interesting is the clustering of binary combinations

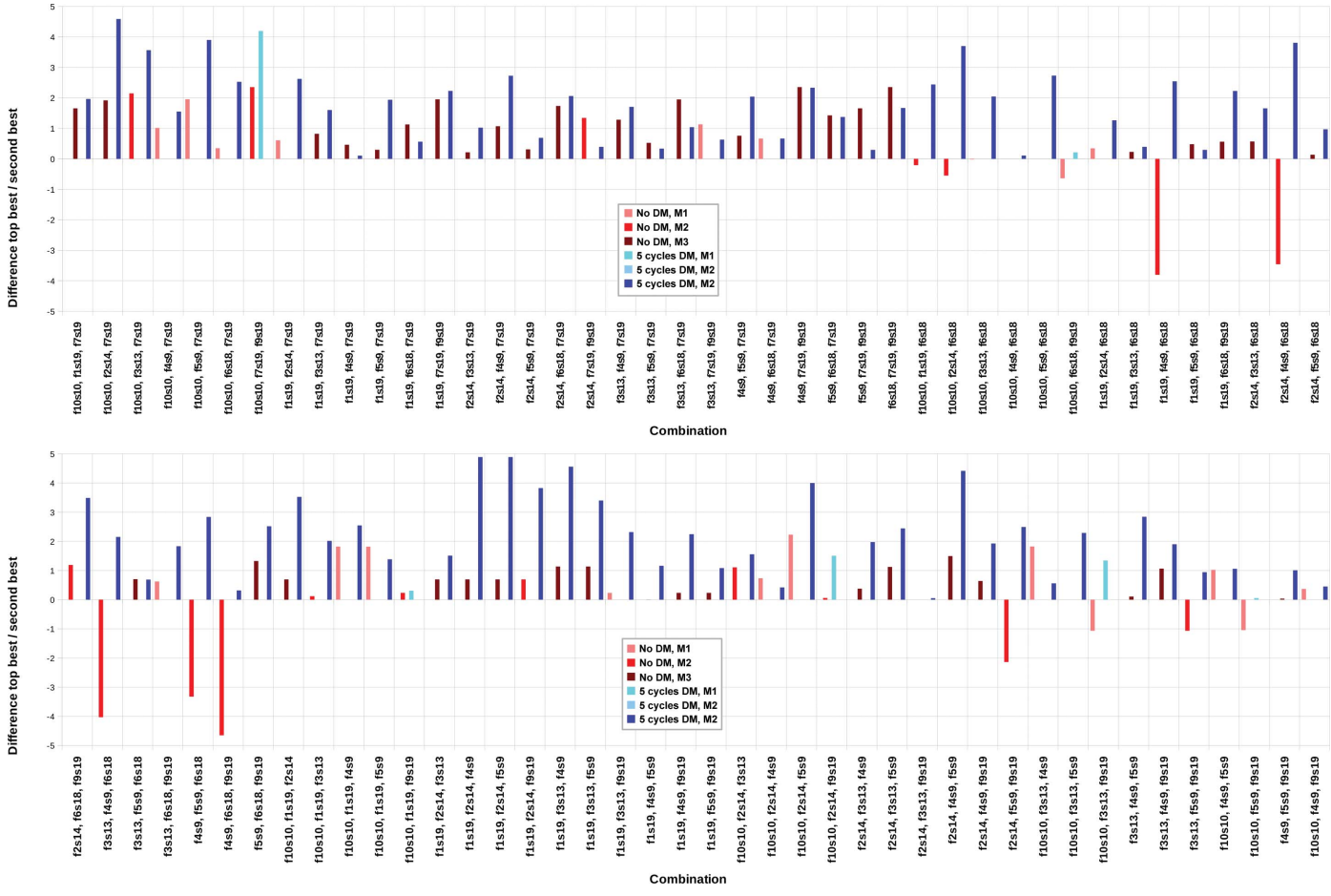


Figure 6
MPD difference between clusters combining three perfect helices from 2iu1 for the two origin shifts corresponding to the lowest and second-lowest MPD. Fragments are labelled in the order indicated in Fig. 3 and by the number of amino acids. The bars quantify the difference in degrees between the MPD corresponding to the correct origin shift and the MPD for the best-scoring wrong origin shift; shades of red represent no density modification and shades of blue represent five cycles of density modification; the darker the colour, the higher the consistency among different reference choices. Negative bars represent cases where a wrong origin shift yields the minimal MPD and would have been selected instead of the correct one. With density modification (blue bars) the correct origin is always identified.

from the eight largest helical and four sheet fragments, as in a real search case they would be independent and necessarily complementary. The correct origin shift leads to the lowest MPD in 29 of the 32 unique cases, although the MPDs are high (86–89.6°). The structure can be phased from the combination characterized by the highest CC and MPD discrimination among origins.

3.1.3. The transcription antiterminator PRDII (3gwh) in a polar space group. The structure is composed of 220 residues; the crystals diffracted to 1.95 Å resolution and belonged to the most common polar space group in the PDB, $P2_1$. Ten helical fragments ranging from 11 to 20 residues were extracted from the deposited structure. A cartoon characterizing the helical fragments and a graph showing their characteristics and the MPEs for the phases that originate from them are displayed in Fig. 10.

The same procedure as described for the previous cases was followed for this structure, except that the possible origin shift along the *y* direction is not constrained. This precludes the differences among possible origins as a criterion, but emphasizes the role of consistency of the results when using the alternative fragments as references. When calculating the CC for a structure composed of the fused fragments, their positions were allowed to refine locally. The correct origin is selected in over one quarter of the combinations (26 out of 90 possible combinations). The model selected as a seed to reference the origin influences the outcome, so that the resulting shift may differ within a pair. Sorting according to the MPD and selecting the 20 pairs where the origin shift was equivalent regardless of the reference chosen allowed ten reliable clusters to be identified. Correct matches are furthermore characterized by relatively lower MPDs and higher CCs, although the difference is not clear-cut. The fact that the origin shift is unconstrained in one of the directions was expected to hinder discrimination, and even with perfect fragments the correct combinations are often missed. All preliminary tests with this structure using placed model fragments in real searches failed to recognize the correct origin shift relating correct solutions. Therefore, the subsequent results and discussion are limited to nonpolar space groups.

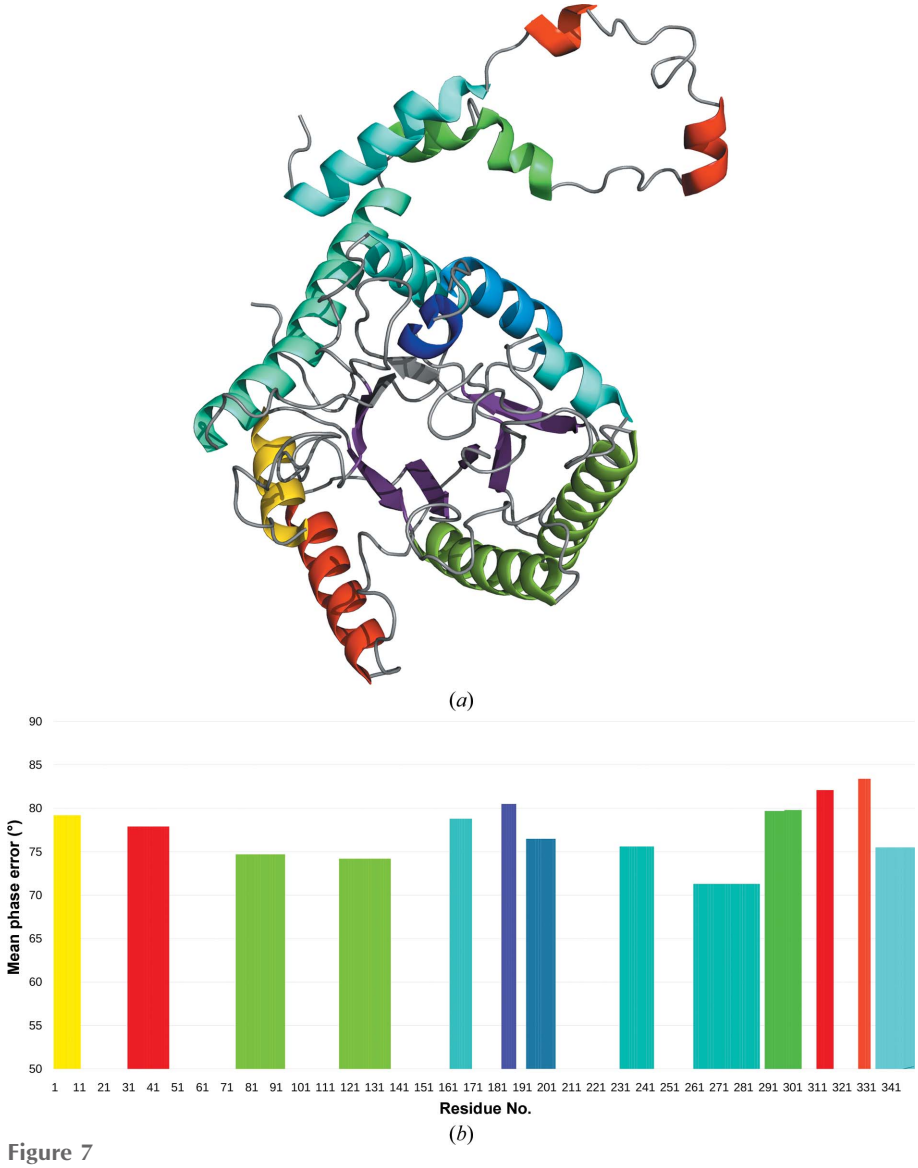


Figure 7
Characterization of the 14 helices and six strands used in the structure of 1mnz. (*a*) Cartoon representation, with a rainbow colour gradient representing the main-chain average *B* factor of the helices (red for highest and blue for lowest *B*). (*b*) MPE of the phase set obtained from each helix *versus* residue count, colour-coded according to the *B* values as in (*a*). The four helices presenting higher *B* factors are either very small or exposed at the surface, and as in the case of 2iu1 the results are comparable except for the smaller helices.

3.2. Model helices located by *Phaser* in an *ARCIMBOLDO* run

A model helix can fit a structure at different, partially overlapping positions, but a real case will yield incorrect as well as correct locations, so that search solutions can be wrong or right, related or independent. In the following test cases, partial solutions from an *ARCIMBOLDO* run with default parameters using model helices of 14 alanine residues as search fragments were clustered in reciprocal space. The conclusions derived from the tests on perfect fragments were assumed; that is, five cycles of density modification were applied to single fragments prior to phase combination and discrimination among the lowest/second-lowest MPD was assessed for nonpolar space groups. The aim was to reach a solution after just one round of fragment placement in

cases where the structure cannot be solved by expanding from a single fragment.

ARCIMBOLDO computes a *Phaser* rotation search and clusters all selected rotation peaks within 15°, taking symmetry into account. While the same rotation cluster may lead to different solutions for the same location of a helix or for roughly parallel helices, different clusters should mark different helices. This is taken into account when clustering phase sets.

Figure 8
MPD difference between clusters for the two origin shifts corresponding to the lowest and second-lowest MPD in the case of xylose isomerase. Fragments are labelled by their order in Fig. 7 and the number of amino acids. The bars show the difference in degrees between the MPD corresponding to the correct origin shift and the MPD for the best-scoring wrong origin shift. Negative bars represent cases where a wrong origin shift yields the minimal MPD and would have been selected instead of the correct one. In 15 cases the use of density modification (not used in the pink bars) is required to select the correct origin shift. It often improves the difference between a wrong and the correct shift.

3.2.1. The all-helical structure in a nonpolar space group: 2iu1. The run yielded 128 different fragment locations originating from four distinct rotation clusters. Three of them are correct. Clustering the phase sets produced after five cycles of density modification within 60° MPD reduces the pool to 113, identifying positions that are related by an elongation of the helix regardless of their correctness. The extent of the elongation is typically 1–3 residues, and the MPD indicates the value. At this stage, although it is possible to proceed from the resulting phase sets, combining matching solutions into a single PDB entry after taking the appropriate origin shift into account gives MPEs that are lower by a few degrees. The PDB entries yielding the highest CC within each cluster were selected as a reference to identify solutions originating from the remaining rotation clusters that would blend within 87° MPD. In this case, this produces just four clusters that can be sent to *SHELXE* expansion. One of them corresponds to the combination of correct solutions, and ten cycles of density modification and autotracing render a solution characterized by an MPE of 47.8° and a CC of 47.70%.

3.2.2. 3jvl from a standard *ARCIMBOLDO_LITE* standalone run. This 120-amino-acid structure, with data to a resolution of 1.2 Å, was used to test the effect of limiting the resolution of the data used in solution clustering. A default limit of 2 Å has generally been used in order to assess the suitability of the method at least up to this resolution limit. On the other hand, including higher resolution data might highlight differences and thus be detrimental. Initially, resolution limits of 2.0, 1.8, 1.5 and 1.2 Å were compared. The *ARCIMBOLDO* run yielded 98 different fragment locations originating from four distinct rotation clusters. Two of the solutions, characterized by weighted mean phase errors computed at full resolution of 73.7 and 75.4°, were clearly correct, and a third one corresponded to one of the previous helices backtraced

(MPE of 83.7°). Two nonrandom solutions, with a partial overlap with correct fragments, were characterized by an MPE of around 86°. Clustering the phase sets produced after five cycles of density modification within 60° MPD reduced the pool to 90, identifying positions related by elongation and one corresponding to a slight turn of a helix. These elongated solutions are consistent but wrong. Indeed, given the small number of correct solutions and the fact that they are not independent, consistency is not an indication of correctness, which is only revealed at the expansion stage in the case of a successful solution. The two helical fragments placed at the same position in reversed directions are not clustered as their mutual MPD is around 68° (depending on the resolution), but this value is far from those produced by independent helices. After combination of matching solutions into single coordinate files, all fragments were sorted by CC within each rotation cluster and sequentially used as references to identify solutions originating from the remaining rotation clusters that would blend within 87° MPD. Evaluating all possible 90 clusters through *SHELXE* expansion, those corresponding to the combination of both correct solutions, that placed backwards and a few other solutions including the nonrandom ones succeed in solving the structure after ten cycles of density modification and autotracing (MPE of 19° and CC of 46%). These solutions are joined whenever one of the nonrandom fragments is used as a reference, although the spurious fragments also included may vary. Nevertheless, correct solutions are distinguished since their MPD differences for best/second-best origin shift are markedly higher. The same results with respect to fragment selection and numerical differences for MPDs are observed at each of the four resolution limits tested; thus, the default of 2 Å is adopted. Systematically testing the effect of the resolution limit from 2.0 to 10 Å in 0.1 Å steps led to a correct origin shift and a gradual lowering of the MPD from 85 to 77°, preserving the correct clustering, until 5.4 Å resolution. Below this resolution the correct shift is always missed. The disadvantage is that as the resolution is lowered an increasing number of wrong solutions are drawn into the cluster, preventing successful expansion. In contrast, at 2.0 Å resolution the reverse helix stands out through its MPD discrimination.

3.2.3. 3o55 from a standard *ARCIMBOLDO_LITE* standalone run. The 125-amino-acid structure of the human FAD-linked augmenter of liver regeneration (ALR; Banci *et al.*, 2011) is composed of a bundle of roughly parallel helices. Data to 1.9 Å resolution in space group $C222_1$ were available (PDB entry 3o55).

The run yielded 79 different fragment locations originating from two distinct rotation clusters matching true rotations. After the translation search, two positions were correct. As the second was an elongation of the first, they belonged to the same rotation cluster. Clustering the phase sets produced after five cycles of density modification within 60° MPD reduced the pool to 57 clusters, identifying positions related by elongation or slight translation. In this case, the fused solution from the 60° clustering, characterized by a CC among the top five, may already be pushed to a solution after 27 cycles of density modification and autotracing (MPE of 57° and CC of 30%), whereas none of the single fragments succeed. As these are the only two correctly placed fragments, further clustering with a high MPD threshold does not yield any additional solutions.

The default practice of limiting the second clustering step to combine solutions from different rotation clusters saves computing time but may be detrimental if different helices in the structure accidentally share the same rotation.

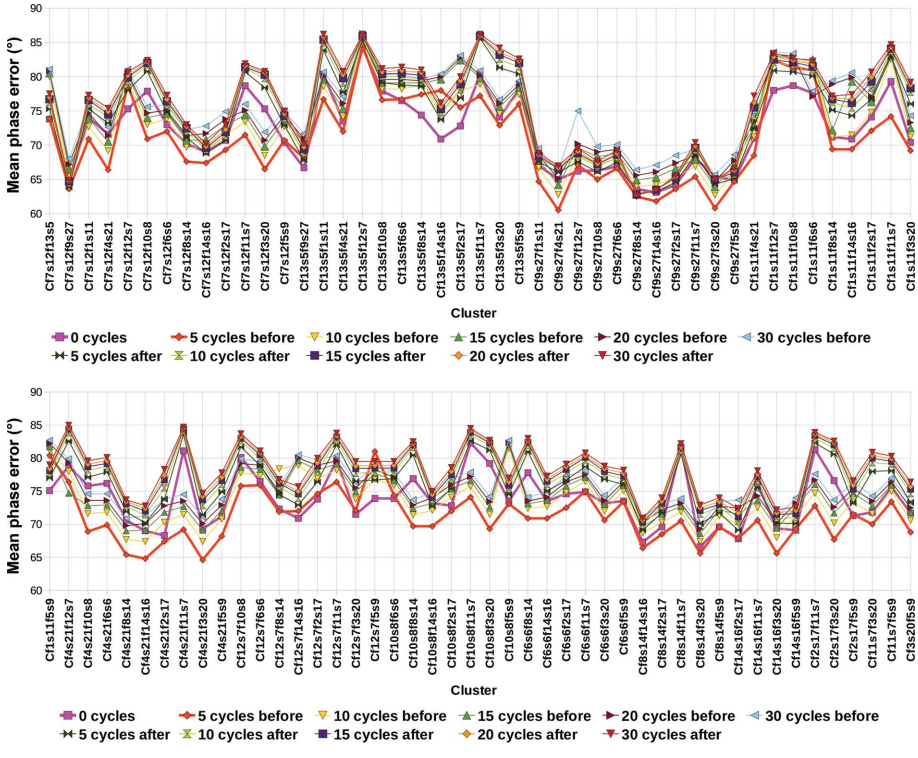


Figure 9
Effect of density modification applied at different stages on the MPE of clusters combining perfect helices extracted from xylose isomerase with respect to true phases as calculated from the refined structure. The orange line, representing five cycles of density modification, leads in most cases to the best phases with the lowest MPE.

3.3. Clustering larger, overlapping fragments from distant homologues: MltE

MltE (Artola-Recolons *et al.*, 2011; PDB entry 2y8p) is a bacterial endolytic peptidoglycan lytic transglycosylase from *Escherichia coli*. The crystals belonged to space group $C222_1$ and contained two copies of the mainly helical, 194-amino-acid MltE monomer in the asymmetric unit. This structure was first solved with 2 Å resolution data

using the *ARCIMBOLDO_SHREDDER* approach (Sammito *et al.*, 2014) with the structure of Slt70 (PDB entry 1qte; van Asselt *et al.*, 1999) as a template. Traditional molecular replacement had failed using this closest homologue, which displayed an r.m.s.d. of 3.1 Å over 160 C^{α} atoms, to extract a search fragment. Instead, smaller partial fragments were required for successful placement. *ARCIMBOLDO* succeeded using, as an alternative to single helices, a set of partially overlapping, small models generated by systematically omitting parts of the loop-trimmed 1qte structure.

Originally, the placement of two copies and trimming the solution amino acid by amino acid were needed to solve the structure. Alternatively, combining information from partial structures after searching for one monomer should be a suitable strategy, cutting down the computing effort associated with searching for a second copy to complete each of the putative solutions yielded for each model. For this study, 90 polyalanine models were created by omitting chunks of 50 residues from the initial 140 in windows of one residue. From these 90 models, 5929 solutions were generated with *Phaser*, but most of them were incorrect, which is unsurprising given the high r.m.s.d. of the models from the target structure. The phases generated from these models are random except for four sets, depicted in Fig. 11(*a*), characterized by mean phase errors of 70.5, 71.1, 71.8 and 73.2°. They do not correspond to top solutions within their runs, as they occupy positions 25/25, 35/50, 17/95 and 20/35 in the *Phaser* LLG rank of solutions. This highlights how the central problem is to extract extremely weak signals from the dominating noise. Combining these four sets succeeds in solving the structure while no isolated solution does, as displayed in Fig. 11(*b*), which shows one of the solutions and the map obtained with the cluster. Clustering was accomplished in a two-step procedure. Firstly, similar solutions from overlapping models placed on top of each other are identified and clustered. Then, complementary clusters corresponding to very different models or placed on different monomers are built. In this case where partial solutions with errors are combined, emphasizing the high-resolution data is adverse and the use of the *E*-weighted MPD rather than the *F*-weighted MPD leads to higher values by a small margin (up to 2°) and the differences between top and second best become smaller. This has no practical consequences when identifying similar solutions, but compromises the detection of complementary fragments.

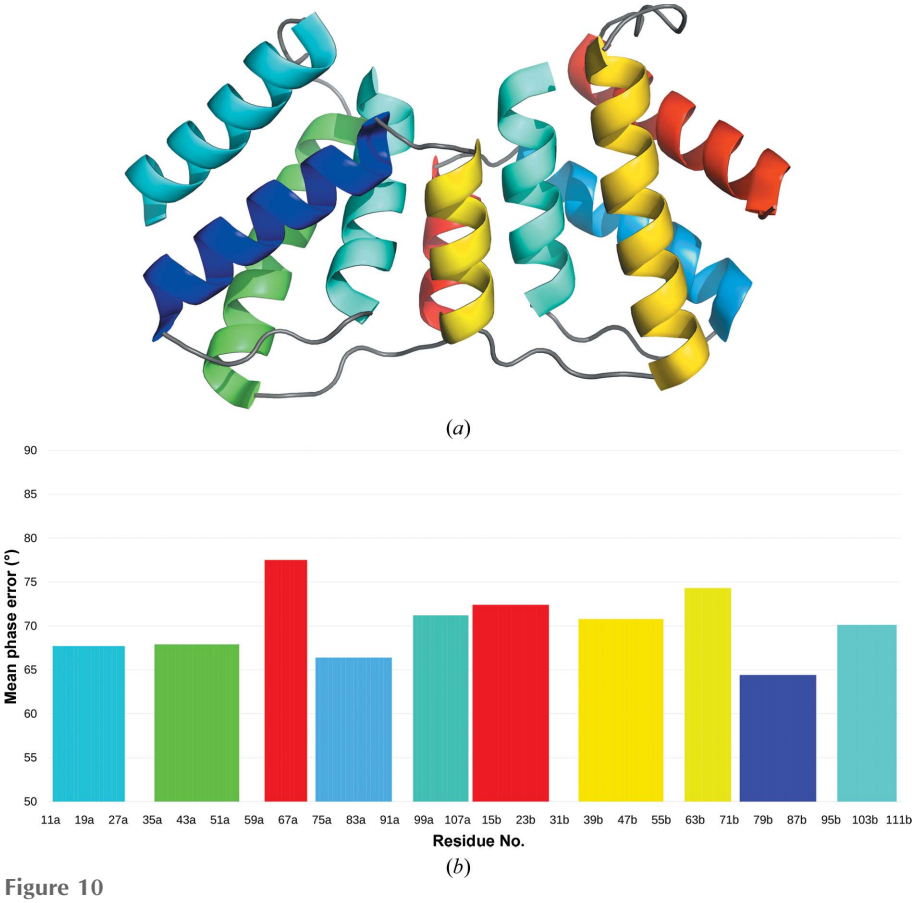


Figure 10
Characterization of the ten helices in the structure 3gwh. (*a*) Cartoon representation, with a rainbow colour gradient representing the main-chain average *B* factor of the helices (red for highest and blue for lowest *B*). (*b*) MPE of the phase set obtained from each helix *versus* residue count, colour-coded according to the *B* values as in (*a*). The monomers in the asymmetric unit presents different *B* factors; for equivalent helices, the lower the *B*-factor average the better the MPE.

3.3.1. Identifying similar solutions from different models. The procedure followed started by ranking solutions according to the CC yielded by the fragment against the experimental data and generating phase sets from each coordinate file, subjecting them to five cycles of density modification. Then, for the top 10% of the solutions the phases were clustered against all other files within 60° MPD. 550 clusters were obtained, each grouping 3–6 solutions. In practice, the 60° threshold is generous, as in general clustered solutions are characterized by mean phase differences of below 40°. They involve mainly contiguous models, differing in a few residues. At the end of this round, three of the correct solutions are joined in one of the 550 clusters formed, while the fourth one remains single.

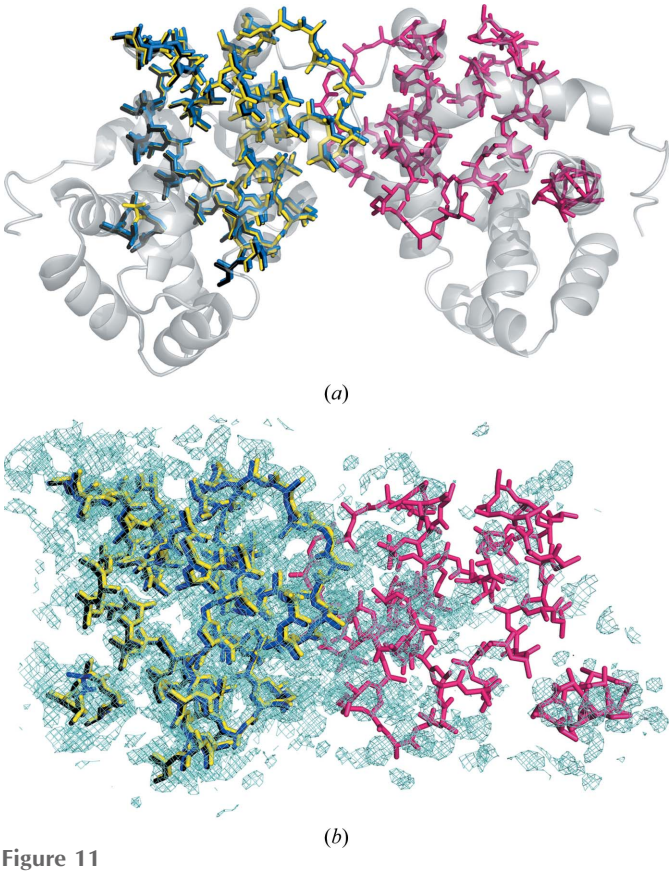
3.3.2. Combining complementary information. The clustered phase sets generated were used as seeds for a further clustering round within 87° MPD in order to combine sets containing complementary information, such as the four models represented in Fig. 11(*a*), covering parts of different monomers in the structure. One of the clusters obtained at this step, characterized by a difference between correct and second-best origin shifts of 3.1°, contained a combination of a previous cluster of three of the correct solutions described and the fourth one. Using the clustered phases as input,

three cycles of *SHELXE* density modification and autotracing were able to discriminate a solution characterized by a trace of 165 residues with a CC of 24.27%. Further cycles can be applied to improve this solution to a main-chain trace of 304 residues characterized by a CC above 45% after nine iterations.

3.4. Solution of a previously unknown structure clustering fragments from homologues: PPAD

PPAD is a peptidylarginine deiminase from *P. gingivalis*. Over 20 data sets were collected from different crystals at the ESRF and ALBA synchrotrons and combined into a highly redundant data set (Giordano *et al.*, 2012). Table 1 shows the statistics of the merged data set used for phasing. The crystals belonged to space group $P2_12_12_1$ and contained one copy of the 432-amino-acid monomer in the asymmetric unit, corresponding to a solvent content of 40%.

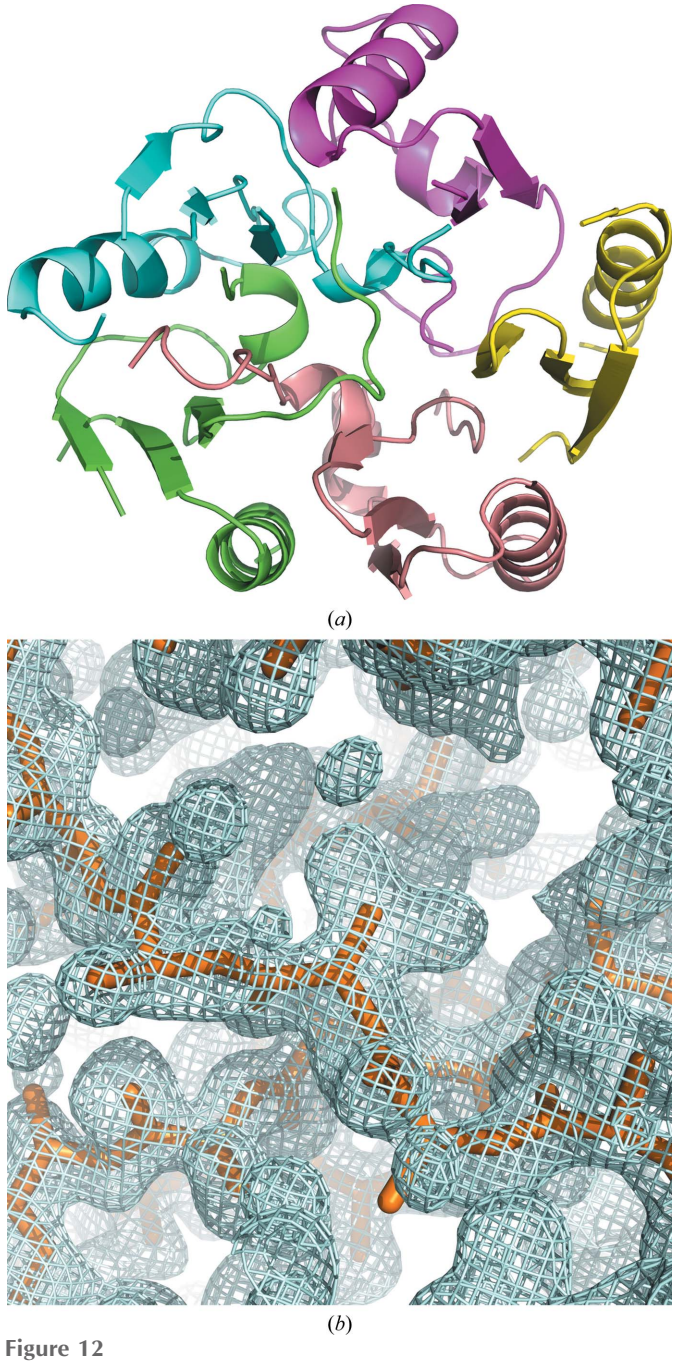
A protein–protein *BLAST* (Altschul *et al.*, 1997) search with the PPAD sequence showed that the closest sequence available corresponds to chain *A* of PDB entry 3hvm (Jones *et al.*, 2010), with an *E*-value of 2×10^{-5} and 23% sequence identity. However, a search in *HHpred* (Söding, 2005) broadened the choice of templates to a set of six structures characterized by 100% probability based on the real-world score distribution for negative and homologous domain pairs in an all-against-all comparison using *SCOP* (Murzin *et al.*, 1995). Apart from 3hvm, the following structures were found to have a homologous relationship: 1zbr, 1xkn, 2jer, 3h7c and 2evo. The common fold in all of these structures is a pentein β/α propeller composed of five α–β–β–α–β units arranged around a pseudo-fivefold axis, as depicted in Fig. 12, which shows 1zbr (Northeast Structural Genomics Consortium, unpublished work). Standard molecular replacement did not



(*a*)

(*b*)

Figure 11
Structure of MltE. (*a*) Cartoon representation displaying with the same origin the four partial fragments combined in blue, black, yellow and pink on the final structure drawn as a grey cartoon. (*b*) Electron-density map generated from the phases of the three overlapping models (blue, black and yellow), showing some features of the missing monomer in pink.



(*a*)

(*b*)

Figure 12
Solution of PPAD. (*a*) Cartoon representation of the search models extracted from 1zbr. The figure shows five subunits. The best models were combinations of two contiguous subunits. The search and the target structures present an r.m.s.d. of 1.53 Å over a core of 246 residues. (*b*) Electron-density map.

succeed with any of these six models. Instead, a variety of fragments were generated from all of these templates. The structures were decomposed into the five pseudo-repeats and the models were cut in a number of ways: with and without the helices, with and without side chains, further trimmed from loops and partially overlapping. The resulting models and libraries were used as search fragments in *ARCIMBOLDO* runs with various parameterizations, systematically varying the resolution for *Phaser* rotation and translation (McCoy *et al.*, 2005) searches as well as the r.m.s.d. estimation. One of the models cut out from the 1zbr template, composed of the poly-Ala-trimmed fifth and first repeats, stood out as producing a unique rotation cluster and a lower number of solutions with a higher LLG than any other trial for a rotation resolution cutoff of 2.1 Å and a translation cutoff of 1.7 Å and r.m.s.d. set to 0.8 Å (Oeffner *et al.*, 2013). Still, its expansion did not yield a solution. The top LLG solution for this model was used as a reference to cluster phases from all of the other 350 solutions produced by the pool of five models. From the 350 phase sets derived from these models after five cycles of density modification, one solution produced by a model derived from the fourth and fifth repeats matched within a tolerance of 60°. While other clusters could be generated from other placed fragments, they did not solve the structure, as no other correct partial solutions were produced. The aforementioned solutions, once merged, succeeded with *SHELXE* (Thorn & Sheldrick, 2013), iterating 20 cycles of density modification with autotracing, using data extrapolation to 1.0 Å (Usón *et al.*, 2007) in rendering a trace of 368 residues with a CC of 40.14%.

4. Conclusions and outlook

From the studies with perfect fragments extracted from test cases, it became evident that applying a few cycles of density modification prior to clustering enhances both recognition of the origin shift and improvement of the clustered phases to be expanded in autotracing. As the partial solutions constitute approximations to the complete structure, the improvement brought about by solvent flattening possibly enhances the approximation to the correct phases and aids recognition. In the case of nonpolar space groups, evaluating the *F*-weighted MPD among fragments for all possible origin shifts allows the discrimination between the lowest and second-lowest MPD to be relied on. Fragment combinations characterized by larger differences can be trusted to be correct for fragments that are large enough to be located in a search. For polar space groups identification of the correct origin shift is more uncertain, and although tests have established it to be possible with perfect fragments, no success has been achieved with solutions from a run. Therefore, the option of clustering partial *ARCIMBOLDO* solutions from placed helices is only advised for nonpolar space groups. In any case, the approach within *ARCIMBOLDO* remains a multi-solution one: clustering and expansion from the combined phases may be attempted after each fragment-placement round. If unsuccessful, the next round is initiated.

The clustering of partial solutions from helices placed in an *ARCIMBOLDO* round is performed in two steps. Related, equivalent or partially overlapping helices are reduced to a common solution by clustering MPDs with a tolerance of 60°. A second round set with a high tolerance of 87° picks complementary, non-overlapping solutions, building both correct and incorrect clusters.

Larger tertiary-structure fragments as search models offer a more favourable scenario than isolated secondary-structure fragments. In this case, the use of partially overlapping fragments enhances origin-shift recognition. Therefore, we are next going to include clustering as a default within *ARCIMBOLDO_SHREDDER*, where the associated CPU increase is in the range of minutes. A test case containing two copies in the asymmetric unit was solved with fragments cut out from a low-homology model. Only four of the 6000 partial solutions were correct, but the same strategy as established for model fragments, consisting of density modification and two rounds of clustering with different MPD thresholds, allowed them to be combined. From the resulting phase set, the structure was solved. We have successfully applied reciprocal-space clustering to solve the previously unknown structure of a 432-amino-acid protein from partial, overlapping models derived from a homologous protein for which standard molecular replacement did not render a solution.

Acknowledgements

This work was supported by grants BFU2012-35367 and BIO2013-49604-EXP from the Spanish Ministry of Economy and Competitiveness and 2014SGR-997 from Generalitat de Catalunya. We thank the ESRF and ALBA for provision of synchrotron-radiation facilities as well as Alexander Popov for assistance with beamline ID23 and Jordi Juanhuix for assistance with beamline XALOC (Juanhuix *et al.*, 2014). We thank Randy Read and Airlie McCoy for helpful discussions and corrections.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.
Artola-Recolons, C., Carrasco-López, C., Llarrull, L. I., Kumarasiri, M., Lastochkin, E., Martínez de Ilarduya, I., Meindl, K., Usón, I., Mobashery, S. & Hermoso, J. A. (2011). *Biochemistry*, **50**, 2384–2386.
Asselt, E. J. van, Thunnissen, A.-M. W. H. & Dijkstra, B. W. (1999). *J. Mol. Biol.* **291**, 877–898.
Banci, L., Bertini, I., Calderone, V., Cefaro, C., Ciofi-Baffoni, S., Gallo, A., Kallergi, E., Lionaki, E., Pozidis, C. & Tokatlidis, K. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 4811–4816.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2012). *Acta Cryst.* D**68**, 1622–1631.
Bieniossek, C., Schütz, P., Bumann, M., Limacher, A., Usón, I. & Baumann, U. (2006). *J. Mol. Biol.* **360**, 457–465.

Buehler, A., Urzhumtseva, L., Lunin, V. Y. & Urzhumtsev, A. (2009). *Acta Cryst.* D**65**, 644–650.
Bunkóczi, G., Echols, N., McCoy, A. J., Oeffner, R. D., Adams, P. D. & Read, R. J. (2013). *Acta Cryst.* D**69**, 2276–2286.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2012). *J. Appl. Cryst.* **45**, 1287–1294.
Burla, M. C., Giacovazzo, C. & Polidori, G. (2010). *J. Appl. Cryst.* **43**, 825–836.
Caliandro, R., Dibenedetto, D., Cascarano, G. L., Mazzone, A. & Nico, G. (2012). *Acta Cryst.* D**68**, 1–12.
Carrell, H. L., Hoier, H. & Glusker, J. P. (1994). *Acta Cryst.* D**50**, 113–123.
DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H. L., Das, D., Vorobiev, S. M., Iwaï, H., Pokkuluri, P. R. & Baker, D. (2011). *Nature (London)*, **473**, 540–543.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Fujinaga, M. & Read, R. J. (1987). *J. Appl. Cryst.* **20**, 517–521.
Giordano, R., Leal, R. M. F., Bourenkov, G. P., McSweeney, S. & Popov, A. N. (2012). *Acta Cryst.* D**68**, 649–658.
Glykos, N. M. & Kokkinidis, M. (2003). *Acta Cryst.* D**59**, 709–718.
Goulas, T., Mizgalska, D., Garcia-Ferrer, I., Kantyka, T., Guevara, T., Szmigielski, B., Sroka, A., Millán, C., Usón, I., Veillard, F., Potempa, B., Mydel, P., Solà, M., Potempa, J. & Gomis-Rüth, F. X. (2015). *Sci. Rep.* **5**, 11969.
Jones, J. E., Causey, C. P., Lovelace, L., Knuckley, B., Flick, H., Lebioda, L. & Thompson, P. R. (2010). *Bioorg. Chem.* **38**, 62–73.
Juanhuix, J., Gil-Ortiz, F., Cuní, G., Colldelram, C., Nicolás, J., Lidón, J., Boter, E., Ruget, C., Ferrer, S. & Benach, J. (2014). *J. Synchrotron Rad.* **21**, 679–689.
Lunin, V. Y. & Lunina, N. L. (1996). *Acta Cryst.* A**52**, 365–368.
Lunin, V. Y. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.
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.
McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Cryst.* D**61**, 458–464.
Millán, C., Sammito, M. & Usón, I. (2015). *IUCrJ*, **2**, 95–105.
Morris, R. J., Blanc, E. & Bricogne, G. (2004). *Acta Cryst.* D**60**, 227–240.
Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.
Oeffner, R. D., Bunkóczi, G., McCoy, A. J. & Read, R. J. (2013). *Acta Cryst.* D**69**, 2209–2215.
Pröpper, K., Meindl, K., Sammito, M., Dittrich, B., Sheldrick, G. M., Pohl, E. & Usón, I. (2014). *Acta Cryst.* D**70**, 1743–1757.
Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J. & Baker, D. (2007). *Nature (London)*, **450**, 259–264.
Rigden, D. J., Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* D**64**, 1288–1291.
Robertson, M. P., Chi, Y.-I. & Scott, W. G. (2010). *Methods*, **52**, 168–172.
Robertson, M. P. & Scott, W. G. (2008). *Acta Cryst.* D**64**, 738–744.
Rodríguez, D. D., Grosse, C., Himmel, S., González, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Nature Methods*, **6**, 651–653.
Rodríguez, D., Sammito, M., Meindl, K., de Ilarduya, I. M., Potratz, M., Sheldrick, G. M. & Usón, I. (2012). *Acta Cryst.* D**68**, 336–343.
Sammito, M., Meindl, K., de Ilarduya, I. M., Millán, C., Artola-Recolons, C., Hermoso, J. A. & Usón, I. (2014). *FEBS J.* **281**, 4029–4045.
Sammito, M., Millán, C., Frieske, D., Rodríguez-Freire, E., Borges, R. J. & Usón, I. (2015). *Acta Cryst.* D**71**, 1921–1930.
Sammito, M., Millán, C., Rodríguez, D. D., de Ilarduya, I. M., Meindl, K., De Marino, I., Petrillo, G., Buey, R. M., de Pereda, J. M., Zeth, K., Sheldrick, G. M. & Usón, I. (2013). *Nature Methods*, **10**, 1099–1101.
Sheldrick, G. M. (2002). *Z. Kristallogr.* **217**, 644–650.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.
Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.
Sheldrick, G. M., Gilmore, C. J., Hauptman, H. A., Weeks, C. M., Miller, R. & Usón, I. (2011). *International Tables for Crystallography*, Vol. *F*, 2nd online ed., edited by E. Arnold, D. M. Himmel & M. G. Rossmann, pp. 413–429. Chester: International Union of Crystallography.
Shrestha, R., Berenger, F. & Zhang, K. Y. J. (2011). *Acta Cryst.* D**67**, 804–812.
Shrestha, R. & Zhang, K. Y. J. (2015). *Acta Cryst.* D**71**, 304–312.
Söding, J. (2005). *Bioinformatics*, **21**, 951–960.
Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Cryst.* D**60**, 432–438.
Tannenbaum, T., Wright, D., Miller, K. & Livny, M. (2002). *Beowulf Cluster Computing with Linux*, edited by T. Sterling, pp. 307–350. Cambridge: The MIT Press.
Thorn, A. & Sheldrick, G. M. (2013). *Acta Cryst.* D**69**, 2251–2256.
Urzhumtsev, A., Afonine, P. V., Lunin, V. Y., Terwilliger, T. C. & Adams, P. D. (2014). *Acta Cryst.* D**70**, 2593–2606.
Usón, I., Stevenson, C. E. M., Lawson, D. M. & Sheldrick, G. M. (2007). *Acta Cryst.* D**63**, 1069–1074.
Vollmuth, F., Blankenfeldt, W. & Geyer, M. (2009). *J. Biol. Chem.* **284**, 36547–36556.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Yao, J.-X., Dodson, E. J., Wilson, K. S. & Woolfson, M. M. (2006). *Acta Cryst.* D**62**, 901–908.
Yao, J., Woolfson, M. M., Wilson, K. S. & Dodson, E. J. (2005). *Acta Cryst.* D**61**, 1465–1475.