Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

A fragmentation and reassembly method for *ab initio* phasing

Rojan Shrestha[a,b] and Kam Y. J. Zhang[a,b]*

[a]Structural Bioinformatics Team, Division of Structural and Synthetic Biology, Center for Life Science Technologies, RIKEN, Yokohama, Kanagawa 230-0045, Japan, and [b]Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, Kashiwa, Chiba, Japan

Correspondence e-mail: kamzhang@riken.jp

Received 24 July 2014
Accepted 20 November 2014

Ab initio phasing with *de novo* models has become a viable approach for structural solution from protein crystallographic diffraction data. This approach takes advantage of the known protein sequence information, predicts *de novo* models and uses them for structure determination by molecular replacement. However, even the current state-of-the-art *de novo* modelling method has a limit as to the accuracy of the model predicted, which is sometimes insufficient to be used as a template for successful molecular replacement. A fragment-assembly phasing method has been developed that starts from an ensemble of low-accuracy *de novo* models, disassembles them into fragments, places them independently in the crystallographic unit cell by molecular replacement and then reassembles them into a whole structure that can provide sufficient phase information to enable complete structure determination by automated model building. Tests on ten protein targets showed that the method could solve structures for eight of these targets, although the predicted *de novo* models cannot be used as templates for successful molecular replacement since the best model for each target is on average more than 4.0 Å away from the native structure. The method has extended the applicability of the *ab initio* phasing by *de novo* models approach. The method can be used to solve structures when the best *de novo* models are still of low accuracy.

1. Introduction

The most widely used computational method for phasing protein X-ray diffraction data is molecular replacement (MR; Rossmann & Blow, 1962). This is made possible owing to the ever-increasing number of protein structures deposited in the Protein Data Bank (Berman *et al.*, 2002). MR attempts to find the placement of a template in the unit cell and then uses it to provide estimated phases to solve the target structure. This template model should bear a high degree of structural similarity to the target structure for MR to succeed. Advances in sequence alignment (Altschul *et al.*, 1997) and the development of robust comparative structure modelling have provided tools to identify and even to construct suitable templates for MR (Martí-Renom *et al.*, 2000). However, there are numerous sequences that do not have homologous structures and MR cannot be used to solve their structures. Under these circumstances, computationally predicted models can be used as templates for MR.

The *de novo* models can be obtained in principle by searching for the lowest energy conformation given the protein sequence (Baker & Sali, 2001). *Rosetta* (Rohl *et al.*,

2004) was the first method that predicted all-atom models at atomic-level accuracy using a fragment-assembly approach (Bradley *et al.*, 2005). Subsequently, the fragment-assembly approach has been widely exploited in many methods such as *QUARK* (Xu & Zhang, 2012), *I-TASSER* (Roy *et al.*, 2010), *EdaFold* (Simoncini *et al.*, 2012), *NEFILIM* (Shrestha & Zhang, 2014) and others for *de novo* modelling. The accuracy of the predicted *de novo* structures has reached the quality required for search models to achieve successful MR solutions (Qian *et al.*, 2007).

The initial success of the *ab initio* phasing with *de novo* models approach has opened up new opportunities (Qian *et al.*, 2007). This approach was subsequently tested with a large data set (Das & Baker, 2009). This showed that increased conformational sampling substantially improved the success rate in MR. The computational time incurred with increased sampling has been reduced significantly by incorporating MR in the course of *de novo* structure prediction (Shrestha *et al.*, 2011). Crowd-sourced model refinement has also been shown to enable successful MR (Khatib *et al.*, 2011). Many studies have been carried out to increase the success rate in MR. These include using manual intervention in computationally predicted models to check for wrong regions (Rigden *et al.*, 2008), rebuilding models with increased sampling of error-prone residues (Shrestha *et al.*, 2012), using an ensemble of models and trimming potentially wrong regions (Bibby *et al.*, 2012).

The introduction of maximum-likelihood target functions has increased the sensitivity of MR searches (McCoy *et al.*, 2007). This has enabled the placement of small protein fragments correctly in the unit cell. The phase information from correctly placed ideal α-helices has been shown to be sufficient to solve a complete structure in combination with partial structure expansion and automated model building (Rodríguez *et al.*, 2009). This method was further expanded to identify, retrieve, refine and exploit the general tertiary-structural information from small fragments available in the Protein Data Bank (Sammito *et al.*, 2013). Similarly, the potential invariant regions after conformational sampling using elastic network models have also been used for phasing (McCoy *et al.*, 2013).

Phasing by MR with either homologous proteins or *de novo* models requires these templates to be sufficiently close to the target structure. The search model should be within 2.0 Å C^{α}

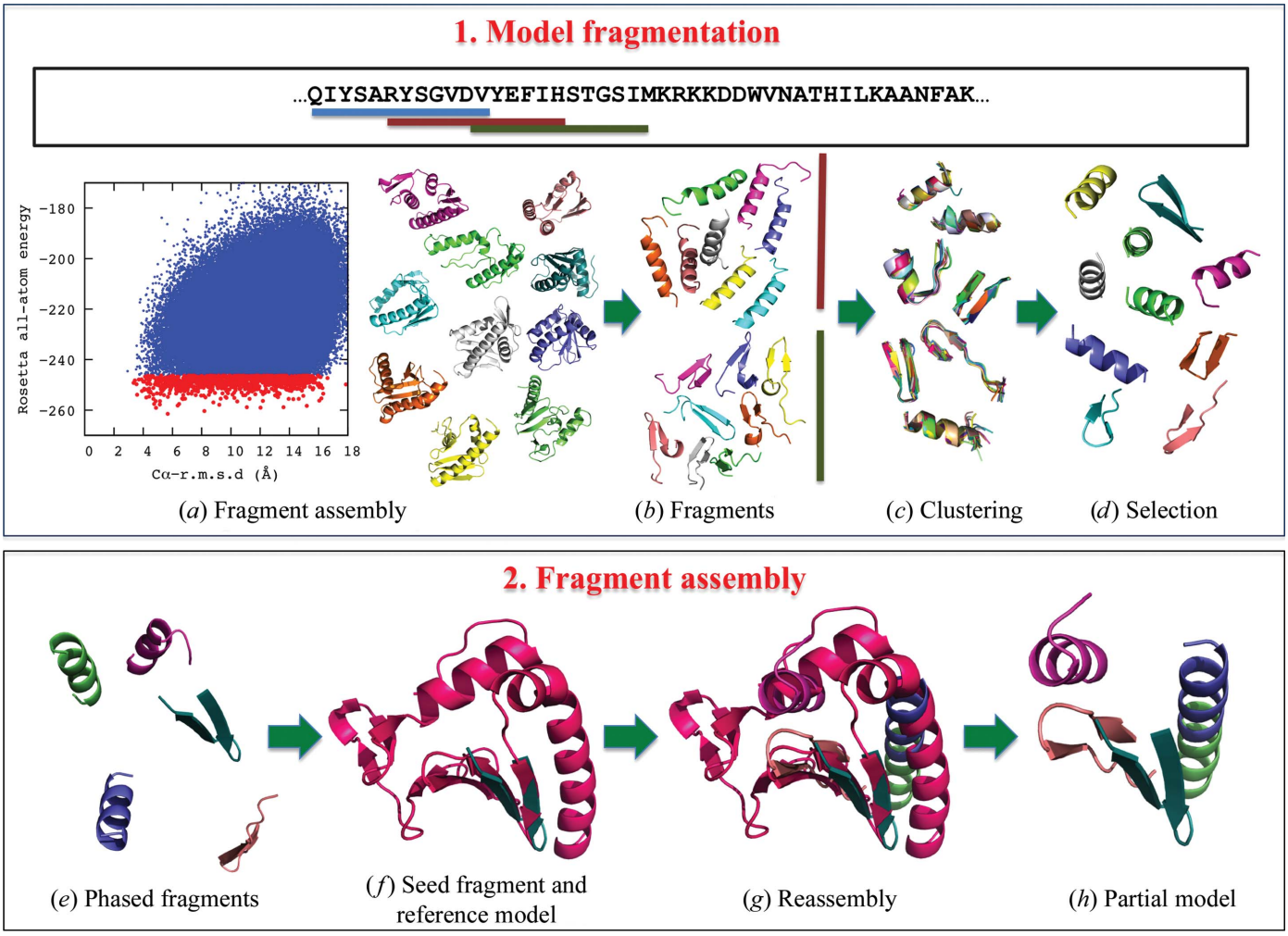

Figure 1
Schematic diagram of the *FRAP* method. The procedure is illustrated as two major sections and each step is labelled and described in the text.

r.m.s.d. from the target structure in general or within 3.0 Å under favourable circumstances, although there is significant variation depending on the protein target and diffraction data (Abergel, 2013; Chen *et al.*, 2000; Shrestha *et al.*, 2012). The development of *de novo* modelling has enabled the generation of templates for successful MR starting from low-homology structures, NMR models or computationally predicted models (Qian *et al.*, 2007). Moreover, it has been shown that combining protein structure modelling with density-guided and energy-guided optimization together with model auto-building can be a powerful approach to solve difficult MR targets (DiMaio *et al.*, 2011). Despite all of this progress, structural solution by MR using *de novo* modelling is far from routine. There are still a significant number of structures that cannot be solved by this approach. While improving the accuracy of a template is a commonly used strategy to achieve success in MR, we have sought to develop methods that could solve structures by MR using low-accuracy models.

Here, we describe a new method for protein structure determination by MR with templates from *de novo* models of low accuracy. These models are the best that could be generated by *de novo* modelling and yet their accuracy is not high enough for use as templates for successful MR. Our method starts with an ensemble of these low-accuracy *de novo* models, breaks them into overlapping fragments and groups them into clusters, tries to place representative fragments from these clusters in the crystallographic unit cell by MR, and then assembles the successfully placed fragments into a complete model. We have tested this method on ten targets that could not be solved by MR using the best *de novo* models and have shown that our method could solve eight out of these ten targets.

2. Method

Our method combines fragment assembly, disassembly and reassembly in an iterative procedure for the *ab initio* phasing of protein crystallographic diffraction data (Fig. 1). It starts from a pool of low-energy *de novo* models predicted by fragment-assembly methods (Fig. 1*a*). These *de novo* models are disassembled into overlapping fragments of various sizes (Fig. 1*b*), and the corresponding fragments are clustered (Fig. 1*c*). A representative set of fragments is selected for each fragment window (Fig. 1*d*). Each fragment in this set is placed separately into the unit cell by MR (Fig. 1*e*). These independently placed fragments refer to different permissible origins and are reassembled together with the aid of a seed fragment and a reference model (Figs. 1*f* and 1*g*). The reassembled model (Fig. 1*h*) is significantly closer to the native structure than any of the starting *de novo* models and can provide sufficient phase information to enable the determination of the complete structure by automated model building. Our method is implemented in C++ and has been given the acronym *FRAP*, which stands for *FRagment Assembly Phasing*. The *FRAP* software is available from http://www.riken.jp/zhangiru/software.html.

2.1. Data-set and initial model generation

We selected ten proteins of different topologies (all-α, all-β and $\alpha + \beta$) that are challenging targets for phasing using full *de novo* models. The criterion for the selection of each target was that the full all-atom models could not be used as templates for successful MR using *Phaser*. In order to select these targets, we generated 120 000 all-atom models for each protein using *Rosetta*3.2 with default parameters. The 9-mer and 3-mer fragment libraries were obtained from the *Robetta* server (Chivian *et al.*, 2003) with homologous proteins excluded. From these models, 1000 lowest energy models were selected for phasing by MR. The solution for each model was verified to find whether MR was able to place the model correctly in the unit cell.

2.2. Fragment generation and phasing

The 1000 lowest energy models were divided into multiple overlapping fragments. We have used three different fragment sizes that contain 13, 17 or 21 residues. Fragments for each residue position were clustered using *Durandal* (Berenger *et al.*, 2011) with a clustering radius of 1.0 Å. 200 fragments were picked from the top ten clusters based on the number of fragments in each cluster and the cluster ranking. When the number of clusters was less than ten, all clusters were used for selection. Conversely, when fragments were sparsely distributed in the clusters, more than ten clusters were allowed for fragment selection. These fragments were independently used as templates in *Phaser* (McCoy *et al.*, 2007) using default parameters and the estimated coordinate error was set to 0.5 Å. Placed fragments were ranked based on *Phaser Z*-score and log-likelihood gain (LLG).

2.3. Reassembly of placed fragments

The independently placed fragments by MR refer to different permissible origins in the unit cell. These fragments have to be put together by referring to the same permissible origin in order to obtain useful phase information. We have taken a real-space approach to reassemble these independently placed fragments, although a reciprocal-space approach is also possible. A seed fragment was first selected from the pool of placed fragments based on their *Phaser Z*-scores, LLG scores and secondary-structure contents. A full model was randomly chosen from the pool of 1000 lowest energy *de novo* models. This selected full model was superposed onto the seed fragment using the Kabsch algorithm (Kabsch, 1976), imposing the matched region covering the sequence of the seed fragment. All the other placed fragments were reassembled using this superposed full model as a reference. Each placed fragment was transformed by a combination of permissible origins, crystallographic symmetry operators and unit-cell translations. The Euclidean distance between the geometric centres of each transformed fragment and that of the corresponding region in the reference model was calculated. The transformed fragment that is the closest to the corresponding region in the reference model was saved. Clashes between the main-chain atoms in the transformed

fragments were measured. The transformed fragment that contained clashes with non-matching residues was removed. If the clashes were between matching residues in the sequence, the clashed region was trimmed off. After removing fragments with clashes and trimming off overlapping residues, the remaining fragments constituted the reassembled model that was considered as the MR solution for the target structure. This reassembled model was often a partial structure but might provide estimated phases of sufficient quality so that model autobuilding methods, such as *phenix.autobuild*, could complete the entire structure using default parameters without manual intervention. If the model autobuilding was unsuccessful, another full model was selected as the reference model and the fragment-reassembly procedure was repeated. This was iterated five times at most in our experiment.

Table 1
Summary of protein targets with the C^{α} r.m.s.d. of the best predicted models.

No.	Target	Resolution (Å)	Space group	No. of molecules in asymmetric unit	Sequence length	SCOP classification	C^{α} r.m.s.d./ MR-C^{α} r.m.s.d.†
1	1opd	1.50	$P1$	1	85	$\alpha + \beta$	2.78/19.42
2	1cm3	1.60	$P2_1$	1	85	$\alpha + \beta$	2.72/14.79
3	1ew4	1.40	$P3_221$	1	106	$\alpha + \beta$	4.98/8.57
4	2eff	1.80	$P3_221$	1	106	$\alpha + \beta$	4.99/21.44
5	3o55	1.90	$C222_1$	1	119	α	6.18/19.17
6	3nzl	1.20	$P2_12_12_1$	1	73	α	3.48/20.91
7	1ctf	1.70	$P4_32_12$	1	68	$\alpha + \beta$	3.16/10.44
8	1mb1	2.10	$P4_12_12$	1	98	$\alpha + \beta$	2.43/18.31
9	4esp	1.10	$P4_12_12$	1	130	$\alpha + \beta$	5.58/16.59
10	3mx7	1.76	$P3_121$	1	90	β	3.40/18.48

† MR-C^{α} r.m.s.d. is the C^{α} r.m.s.d. between the model after molecular replacement and the native structure. To obtain the score, the model after molecular replacement is rotated and translated according to the symmetry operators, permissible origin and unit-cell translation. If the molecular-replacement solution is correct, the MR-C^{α} r.m.s.d. should be very close to the C^{α} r.m.s.d. The unit of r.m.s.d. shown is Å.

Table 2
The fragmentation and reassembly results at different fragment sizes.

		13-residue fragments			17-residue fragments			21-residue fragments		
No.	Target	C^{α} r.m.s.d. 1†	$R/R_{\rm free}$	C^{α} r.m.s.d. 2†	C^{α} r.m.s.d. 1	$R/R_{\rm free}$	C^{α} r.m.s.d. 2	C^{α} r.m.s.d. 1	$R/R_{\rm free}$	C^{α} r.m.s.d. 2
1	1opd	1.03	0.26/0.31	0.68	0.99	0.26/0.27	0.76	1.06	0.28/0.33	0.83
2	1cm3	1.03	0.30/0.36	1.93	1.45	0.29/0.34	1.42	—	—	—
3	1ew4	0.68	0.34/0.37	0.88	1.98	0.31/0.34	1.20	1.73	0.29/0.33	1.34
4	2eff	2.31	0.38/0.42	2.41	1.40	0.34/0.37	1.15	1.34	0.32/0.36	1.21
5	3o55	—	—	—	—	—	—	—	—	—
6	3nzl	1.45	0.31/0.35	1.15	1.76	0.35/0.39	1.35	2.51	0.36/0.39	1.85
7	1ctf	1.44	0.31/0.34	0.09	1.66	0.29/0.34	1.46	1.86	0.27/0.31	1.41
8	1mb1	—	—	—	—	—	—	—	—	—
9	4esp	1.38	0.35/0.34	0.28	1.77	0.34/0.34	1.22	1.92	0.33/0.34	0.49
10	3mx7	—	—	—	1.55	0.35/0.39	1.43	—	—	—

† C^{α} r.m.s.d. 1 is for the partial models after reassembly and C^{α} r.m.s.d. 2 is for the autobuilt models using the phases obtained from reassembled models. The unit of r.m.s.d. shown is Å.

The fragments independently placed by MR can be related by an arbitrary translation along the polar axes in a polar space group. Under these circumstances, the translation vectors between each fragment and the seed fragment along the polar axes were determined by the fast cross-translation Patterson function implemented in *Phaser* with the seed fragment and each successive fragment entered as known partial structures.

3. Results and discussion

3.1. *Ab initio* phasing by fragment assembly, disassembly and reassembly

Our method was tested on the ten targets listed in Table 1 and the results with different fragment sizes are shown in Table 2. The fragmentation with 17 residues achieved the highest success rate, with eight structures solved out of a total of ten. Fragment sizes of 13 and 21 residues were successful in solving seven and six targets, respectively. Since none of these ten targets could be solved by MR using templates from the 1000 lowest energy models (Table 1), these success rates represent a significant achievement of our method. The C^{α} r.m.s.d.s of the *de novo* model closest to the target structure range from 2.48 to 6.18 Å, which far exceed the limit of the accuracy of templates for successful MR by conventional approaches. The fact that our *FRAP* method can succeed in a majority of these cases demonstrates the power of our fragmentation and reassembly approach. Our target structures contain not only all-α-helical proteins, which are considered to be relatively easier targets for this type of approach (Rodríguez *et al.*, 2009), but also all-β-sheet proteins, as well as mixed $\alpha + \beta$ proteins.

It has been observed that even a low-accuracy template might contain some partial structures that could be used for successful phasing by MR. Trimming off loop regions in a homologous structure is a commonly used strategy to solve structures by MR (Bunkóczi & Read, 2011; Stein, 2008). To identify a conserved core from an ensemble of predicted models and use it for MR is another powerful approach (Bibby *et al.*, 2012; Mao *et al.*, 2011). Our method is a new approach that systematically examines structural elements and automatically identifies good ones for successful phasing. Instead of identifying a conserved core domain, accurately predicted fragments were identified. This approach is especially suitable when the large overall structural difference between the template and target structures is owing to the misalignment of otherwise correct structural elements. This often occurs in *de novo* models predicted by fragment-assembly approaches.

Two examples where the secondary-structure elements were correctly predicted but some of those elements were assembled in wrong orientations are given for 1ctf and 1ew4 (Fig. 2). When the best predicted models were compared with their native structures, the C^{α} r.m.s.d.s were 3.16 Å for 1ctf and 4.98 Å for 1ew4. These errors in the reassembled partial models were reduced to 1.86 Å for 1ctf and 1.73 Å for 1ew4.

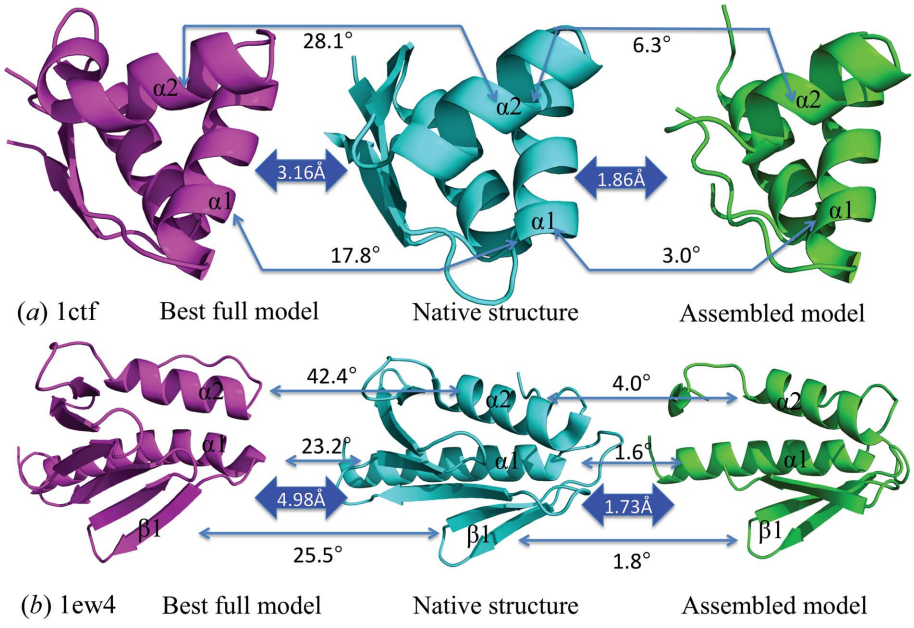

Figure 2
The effect of fragmentation and reassembly in *ab initio* phasing. The best full models (pink), native structures (cyan) and assembled partial models (green) for (*a*) 1ctf and (*b*) 1ew4 are shown. The structures for each target are shown in the same orientation. The overall r.m.s.d.s between each pair of structures are shown. The rotation angles between each pair of labelled secondary-structure elements are also indicated.

Figure 3
The translation-function *Z*-score (TFZ) and log-likelihood gain of seed fragments. The number following the PDB code is the fragment size.

When individual secondary-structure elements were examined, it was found that the α1 helix (Ala63–Gly77) in 1ctf differs by a 17.8° rotation between the best full model and the native structure (Fig. 2*a*). This was reduced to 3.0° in the reassembled partial model. Similarly, the orientation of the α2 helix (Gly79–Glu88) in 1ctf was reduced from 28.1° in the best full model to 6.3° in the reassembled partial model when compared with the native structure (Fig. 2*a*). For 1ew4, the α1 helix (Met1–Trp24) in the best full model differs from that in the native structure by 23.2° and this is reduced to 1.6° in the reassembled partial model (Fig. 2*b*). Similarly, the misalignment angle for the α2 helix (Thr86–Gly100) was reduced from 42.4 to 4.0° when comparing the best full model with the reassembled partial model (Fig. 2*b*). For the β-hairpin (β1; Ile30–Phe43) in 1ew4, the misalignment angle was reduced from 25.5° in the best full model to 1.8° in the reassembled partial model (Fig. 2*b*). These large misalignments of regular secondary-structure elements cannot be handled by trimming off loops. However, the structure can be effectively solved using our *FRAP* method. One natural extension of this strategy is to use fragments identified directly from the Protein Data Bank based on sequence similarity. This is similar to the method implemented in the program *BORGES* (Sammito *et al.*, 2013). Our method has the advantage of using fragments improved by the torsion angle and all-atom refinement protocols in *Rosetta*. Moreover, our method seeks to reassemble multiple fragments found by MR instead of using partial structure expansion of a single correctly placed fragment as in *BORGES* (Sammito *et al.*, 2013).

It should be pointed out that the protein targets in our test set are relatively small. This is a general limitation of *de novo* modelling (Rohl *et al.*, 2004). The average computing time used for the generation of 120 000 all-atom models for each protein target is about 4×10^3 core hours. The computing time would increase tremendously and the model accuracy would decrease significantly for large proteins. These small proteins were chosen in order to contain the computational cost. Multiple molecules in the asymmetric unit increase the challenge to MR, especially with

Table 3
The distribution of the assembled fragments in the *de novo* models.

No. of fragments represents the total number of fragments in the assembled model. No. of models represents the total number of unique *de novo* models from which these assembled fragments came.

No.	Targets	No. of fragments	No. of models
1	1opd	16	9
2	1cm3	10	8
3	1ew4	25	24
4	2eff	11	11
5	3nzl	10	10
6	1ctf	25	22
7	4esp	11	11
8	3mx7	8	5

small fragments. Therefore, our test targets were selected to contain one molecule in the asymmetric unit. The diffraction data for our test set are at high resolution (better than 2.0 Å). This has made it easier for the automated model building since it is used to assess whether the phases from the assembled partial model can lead to the native structure. This is not necessarily a limitation of our method. A more time-consuming manual model-building process can be used at medium resolution or when autobuilding fails.

3.2. Seed fragment and reference model

Our method starts with an initial seed fragment, which determines the orientation and location of a reference model upon which the rest of the placed fragments are reassembled. Therefore, the seed fragment and reference model are important for the success of our method. The selection of the seed fragment is challenging. In our study, the TFZ and LLG scores were used to select the seed fragment. The LLG and TFZ scores of seed fragments are shown as a histogram in Fig. 3. Seed fragments often showed TFZ scores of more than 7.0. Although the choices of seed fragment and reference model directly impact the success of our method, this is mitigated by our iterative multi-solution strategy of selecting a different seed fragment and a new reference model when the previous choice did not lead to successful solution.

Our experiment showed that the seed fragments are often either α-helices or antiparallel β-strands. For 17-residue fragments, the seed fragments were α-helices for five proteins (1ctf, 1cm3, 1opd, 3nzl and 4esp) and antiparallel β-strands for three proteins (1ew4, 3eff and 3mx7).

3.3. Fragment assembly

The individually placed fragments were reassembled using the reference model aligned with a seed fragment. *FRAP* placed more than 60% of residues in the correct orientation and position on average for the successfully solved proteins under three different fragmentation settings (Table 2). The assembled partial models reached a high level of accuracy, with an average C^{α} r.m.s.d. from their respective native structures of about 1.75 Å (Table 2).

FRAP placed 65.1% of residues for the eight successfully phased proteins on average using 17-residue fragments. The

Table 4
Mean phase errors of the assembled and autobuilt models.

The mean phase error of the assembled model represents the average difference between the phases from the assembled model and the phases from the native structure. The mean phase error of the autobuilt model represents the average difference between the phases from the model that was autobuilt based on the assembled model and the phases from the native structure.

		Mean phase error (°)	
No.	Target	Assembled model	Autobuilt model
1	1opd	69.8	31.8
2	1cm3	69.7	40.3
3	1ew4	65.7	30.2
4	2eff	64.9	38.1
5	3nzl	69.1	37.9
6	1ctf	60.0	31.0
7	4esp	71.0	35.0
8	3mx7	72.5	40.6

average C^{α} r.m.s.d. of the assembled structures from the native structures is 1.57 Å. *Phaser* identified the correct placement of the secondary-structure elements in the unit cell including α-helices and antiparallel β-strands connected by small loops. The example of protein 1ew4, which is an $\alpha + \beta$ protein containing two α-helices and six antiparallel β-strands with long loops, is discussed below. *FRAP* started with the seed fragment of an antiparallel β-strand (TFZ = 7.6, LLG = 54; residues 31–47). One of the low-energy models, which deviates by 9.33 Å in C^{α} r.m.s.d. from the native structure, was superimposed on the seed fragment (Kabsch, 1976). *FRAP* searched for the correct position and orientation of other non-overlapping and overlapping placed fragments. *FRAP* selected fragments that are the nearest to the reference model using crystallographic symmetry operators, permissible origin shifts and unit-cell translations. *FRAP* assembled 73.6% of the residues, which belonged to two α-helices, three antiparallel β-strands and a few loops. This partial model provided sufficient phase information to enable *phenix.autobuild* to complete the structure, yielding R and $R_{\rm free}$ factors of 0.31 and 0.34, respectively.

FRAP assembled placed fragments in polar space groups (1opd in $P1$ and 1cm3 in $P2_1$) differently because their permissible origins are infinite along the polar axes. In order to solve the origin-shift problem, we ran the fast translation function in *Phaser* on selected fragments phased from the initial run with the seed fragment as the known partial structure. Subsequently, *FRAP* identified the crystallographic operator and unit-cell translation vector that bring the selected fragment closest to the reference model. This process was repeated until all of the placed fragments had been assembled. *FRAP* succeeded in assembling the placed fragments for 1opd using all three different fragment lengths. For 1cm3, it succeeded for 13-residue and 17-residue fragments.

The distribution of the assembled fragments in the *de novo* models was analyzed. For the eight targets successfully solved with the assembled partial models from 17-residue fragments, the assembled fragments came from a diverse set of *de novo* models (Table 3). This suggests that it is important to use an

ensemble of *de novo* models instead of trying to identify one best model.

3.4. Model building using partial structures

We further assessed the quality of the phases obtained from assembled partial structures by building the complete structures using the automated model-building method *phenix.autobuild*. We monitored the R and $R_{\rm free}$ factors to evaluate the autobuilt models (Table 2). The R factors ranged from 26 to 38% for successfully phased proteins with all three different fragment sizes. Similarly, the $R_{\rm free}$ factors ranged from 27 to 42%. Protein 1opd achieved the best R and $R_{\rm free}$ factors (26 and 27%) with 17-residue fragments.

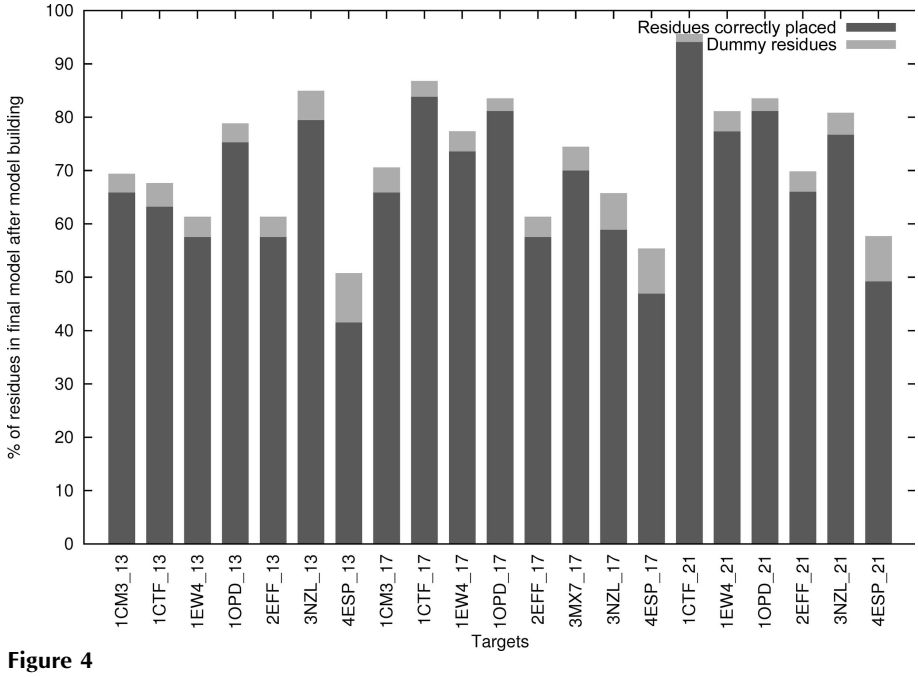

Figure 4
The proportion of interpreted and dummy residues computed from dummy atoms in the electron-density map. The percentage of interpreted models is shown in dark grey. The percentage of dummy residues is shown in light grey and is added on top of the dark grey bars.

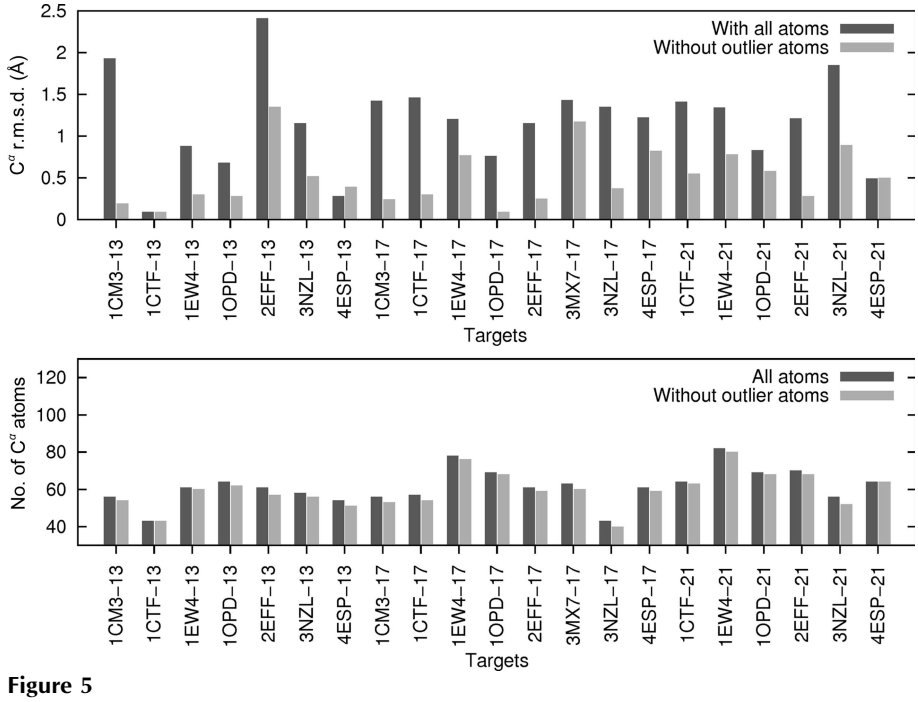

Figure 5
The overall quality of the autobuilt models. A comparison of the accuracy of the final model with and without outlier atoms is shown in the top panel. The numbers of C^{α} atoms included to calculate the model accuracy with and without the outlier atoms are shown in the bottom panel.

The completeness of the autobuilt model needs to be considered in order to properly evaluate the R and $R_{\rm free}$ factors. We have assessed the fraction of missing atoms in the autobuilt model compared with the native structure. We then evaluated the fraction of dummy atoms that occupy the positions of residues in the protein structure. This gives us an estimate of the fraction of uninterpreted electron densities that correspond to protein residues. The result is shown in Fig. 4. The partial structures of 4esp and 1ctf contained the maximum and minimum numbers of dummy atoms, and these numbers were equivalent to nine and one residues, respectively. These correspond to a small fraction of the total number of residues. This shows that these low R/$R_{\rm free}$ factors are mainly owing to the interpreted portion of the structure, with small contributions from the dummy atoms. It was noticed that the R and $R_{\rm free}$ factors for these successfully solved targets are higher than expected for well refined structures, although they are in the range for correct MR solutions. These higher than expected R and $R_{\rm free}$ factors might be owing to the incompleteness of the autobuilt structures and their coordinate errors.

The quality of the phases derived from the assembled partial models can be measured by the mean phase error (MPE). The MPEs for the eight targets successfully solved by the assembled partial models using 17-residue fragments were calculated and compared with that from the corresponding autobuilt models (Table 4). The MPEs for the assembled partial models range from 60.0 to 72.5°. These relatively large phase errors are probably owing to the coordinate errors in the model and its incompleteness. Moreover, the assembled partial model was given a uniform B factor estimated from the Wilson plot and no individual B factors were assigned. However, these phases are of

sufficient quality to enable the autobuilding to bring these models much closer to the native structure, with the resulting MPEs ranging from 31.8 to 48.4°.

The coordinate errors of the autobuilt structures were measured to further evaluate the quality of these structures. The C^{α} r.m.s.d.s of these autobuilt structures from their native structures range from 0.09 Å (1ctf) to 2.41 Å (2eff). The larger than expected coordinate error for some autobuilt structures were owing to a few outlier residues observed at the termini of the assembled fragments. When these outliers were removed using *SUPERPOSE* (Krissinel & Henrick, 2004), the average C^{α} r.m.s.d. was significantly reduced for all three cases. The average C^{α} r.m.s.d.s are 0.45, 0.50 and 0.60 Å for 13-residue, 17-residue and 21-residue fragments, respectively. The C^{α} r.m.s.d.s with and without outlier atoms were compared and the numbers of C^{α} atoms included in the comparison are shown in Fig. 5. Two examples, 1ew4 and 1opd, are shown in Fig. 6, where their assembled partial models and autobuilt models are superposed on their respective native structures and the r.m.s.d. distributions for C^{α} atoms are also plotted. There is significant improvement in the coordinate error for each residue after autobuilding with the exception of a

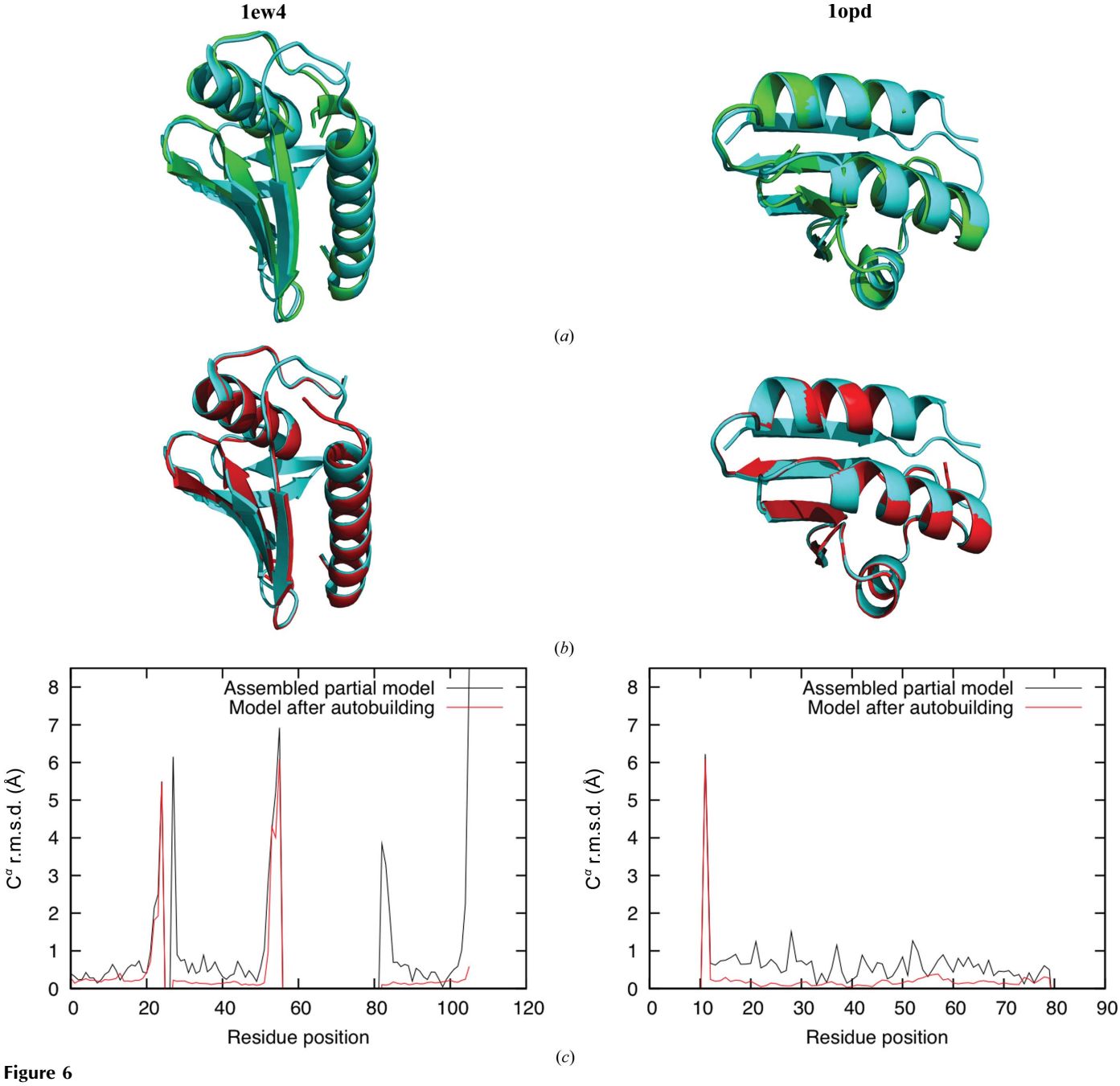

Figure 6
Comparison of the assembled partial models and the autobuilt models with the native structures for two representative targets (1ew4 and 1opd). (*a*) Superposition of the assembled partial model (green) on the native structure (cyan). (*b*) Superposition of the autobuilt model (red) on the native structure (cyan). (*c*) The C^{α} r.m.s.d. distributions of each residue in the sequence for the assembled partial model and the autobuilt model.

few outliers, which also explains the large reduction in the MPEs.

4. Conclusion

We have developed a fragment-assembly method for the *ab initio* phasing of protein crystallographic diffraction data with low-accuracy *de novo* models. Tests with ten targets have shown that while each *de novo* model predicted for these targets used as a whole cannot serve as a template for successful MR, our method can solve up to eight of these ten structures. Our method is a useful addition to the current tool set developed for *ab initio* phasing with *de novo* models. Although one unique aspect of our method is to generate improved and target specific fragments for phasing with MR, our approach may be applicable to the assembly of fragments retrieved directly from the Protein Data Bank (Pröpper *et al.*, 2014). It is also conceivable to apply our approach to distant homologous structural templates (Sammito *et al.*, 2014) or to an ensemble of NMR models for crystallographic structure determination (Mao *et al.*, 2011; Bibby *et al.*, 2013).

We would like to thank the Advanced Center for Computing and Communication, RIKEN, Japan for the computing resources of the RIKEN Integrated Cluster of Clusters (RICC) system. We are grateful to Dr Kevin Cowtan for his advice on the Clipper library. RS is supported by the International Program Associate (IPA) program of RIKEN.

References

Abergel, C. (2013). *Acta Cryst.* D**69**, 2167–2173.
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.
Baker, D. & Sali, A. (2001). *Science*, **294**, 93–96.
Berenger, F., Zhou, Y., Shrestha, R. & Zhang, K. Y. J. (2011). *Bioinformatics*, **27**, 939–945.
Berman, H. M. *et al.* (2002). *Acta Cryst.* D**58**, 899–907.
Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2012). *Acta Cryst.* D**68**, 1622–1631.
Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2013). *Acta Cryst.* D**69**, 2194–2201.
Bradley, P., Misura, K. M. S. & Baker, D. (2005). *Science*, **309**, 1868–1871.
Bunkóczi, G. & Read, R. J. (2011). *Acta Cryst.* D**67**, 303–312.
Chen, Y. W., Dodson, E. J. & Kleywegt, G. J. (2000). *Structure*, **8**, R213–R220.
Chivian, D., Kim, D. E., Malmstrom, L., Bradley, P., Robertson, T., Murphy, P., Strauss, C. E. M., Bonneau, R., Rohl, C. A. & Baker, D. (2003). *Proteins*, **53**, 524–533.
Das, R. & Baker, D. (2009). *Acta Cryst.* D**65**, 169–175.
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.
Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.
Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., Jaskolski, M. & Baker, D. (2011). *Nature Struct. Mol. Biol.* **18**, 1175–1177.
Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* D**60**, 2256–2268.
Mao, B., Guan, R. & Montelione, G. T. (2011). *Structure*, **19**, 757–766.
Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F. & Šali, A. (2000). *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.
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., Nicholls, R. A. & Schneider, T. R. (2013). *Acta Cryst.* D**69**, 2216–2225.
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.
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.
Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. (2004). *Methods Enzymol.* **383**, 66–93.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.
Roy, A., Kucukural, A. & Zhang, Y. (2010). *Nature Protoc.* **5**, 725–738.
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., 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.
Shrestha, R., Berenger, F. & Zhang, K. Y. J. (2011). *Acta Cryst.* D**67**, 804–812.
Shrestha, R., Simoncini, D. & Zhang, K. Y. J. (2012). *Acta Cryst.* D**68**, 1522–1534.
Shrestha, R. & Zhang, K. Y. J. (2014). *Proteins*, **82**, 2240–2252.
Simoncini, D., Berenger, F., Shrestha, R. & Zhang, K. Y. J. (2012). *PLoS One*, **7**, e38799.
Stein, N. (2008). *J. Appl. Cryst.* **41**, 641–643.
Xu, D. & Zhang, Y. (2012). *Proteins*, **80**, 1715–1735.