Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Accelerating *ab initio* phasing with *de novo* models

**Rojan Shrestha, Francois Berenger and Kam Y. J. Zhang***

Zhang Initiative Research Unit, Advanced Science Institute, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

Correspondence e-mail: kamzhang@riken.jp

Received 7 April 2011
Accepted 11 July 2011

*Ab initio* phasing is one of the remaining challenges in protein crystallography. Recent progress in computational structure prediction has enabled the generation of *de novo* models with high enough accuracy to solve the phase problem *ab initio*. This '*ab initio* phasing with *de novo* models' method first generates a huge number of *de novo* models and then selects some lowest energy models to solve the phase problem using molecular replacement. The amount of CPU time required is huge even for small proteins and this has limited the utility of this method. Here, an approach is described that significantly reduces the computing time required to perform *ab initio* phasing with *de novo* models. Instead of performing molecular replacement after the completion of all models, molecular replacement is initiated during the course of each simulation. The approach principally focuses on avoiding the refinement of the best and the worst models and terminating the entire simulation early once suitable models for phasing have been obtained. In a benchmark data set of 20 proteins, this method is over two orders of magnitude faster than the conventional approach. It was observed that in most cases molecular-replacement solutions were determined soon after the coarse-grained models were turned into full-atom representations. It was also found that all-atom refinement was hardly able to change the models sufficiently to enable successful molecular replacement if the coarse-grained models were not very close to the native structure. Therefore, it remains critical to generate good-quality coarse-grained models to enable subsequent all-atom refinement for successful *ab initio* phasing by molecular replacement.

## 1. Introduction

The structures of proteins can reveal crucial information about their biological functions and molecular interactions. X-ray crystallography is the predominant method of obtaining high-resolution three-dimensional protein structures. Protein crystallographic phasing is a major bottleneck in solving crystal structures of proteins when diffraction data have been collected. Traditionally, protein crystallographic phasing is achieved through experimental methods such as multiple isomorphous replacement (MIR; Perutz *et al.*, 1960) or multiple anomalous dispersion (MAD; Hendrickson, 1991) or computational methods such as molecular replacement (MR; Blow & Rossmann, 1961).

Recently, *de novo* protein structure prediction using amino-acid sequences has reached a high level of accuracy and thus created new possibilities for *ab initio* phasing using MR (Qian *et al.*, 2007). The *Rosetta* structure-prediction algorithm (Rohl *et al.*, 2004) has been able to predict the structures of small proteins with high accuracy (Bradley *et al.*, 2005; Das *et al.*,

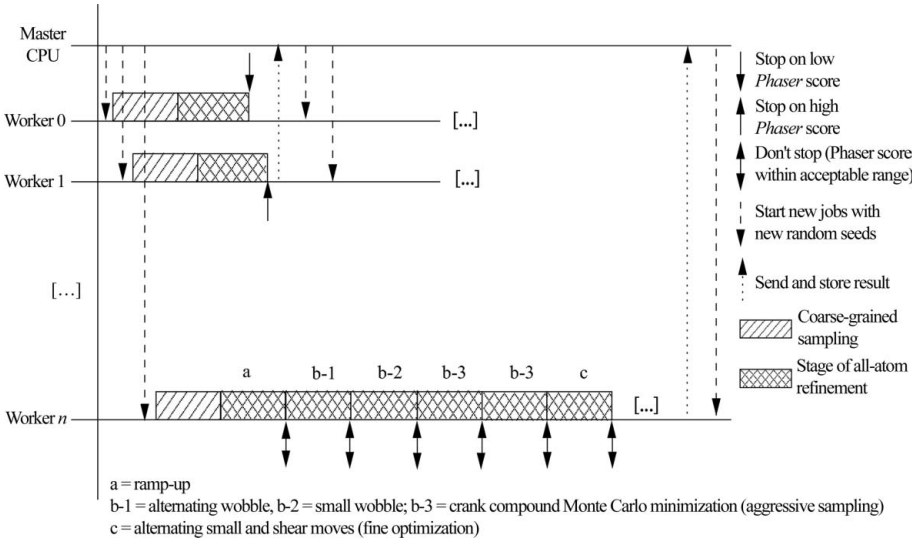


**Figure 1**
Runtime behaviour of MR during all-atom refinement of *Rosetta*. The all-atom refinement method has three major stages (Qian *et al.*, 2007).

2007). *Rosetta* starts with the assembly of fragments, which is followed by random perturbation of torsion angles, combinatorial optimization of side-chain conformations and energy minimization (Rohl *et al.*, 2004; Qian *et al.*, 2007). Atomic level accuracy *de novo* models produced using the *Rosetta* all-atom prediction methodology (Das *et al.*, 2007; Qian *et al.*, 2007) were able to pass the stringent test of successful MR in the absence of suitable templates from the Protein Data Bank (PDB) and without providing any experimental phase information. For example, Baker and coworkers successfully solved the crystal structure of a target in the Critical Assessment of Structure Prediction (CASP) using a *Rosetta*-predicted model with *Phaser* (Qian *et al.*, 2007). In another study, they showed that the atomic level accuracy of *de novo* models predicted using the *Rosetta* method was adequate to achieve successful MR solutions (Das & Baker, 2009). These two studies demonstrated that crystallographic phasing can be solved *ab initio* with *de novo* models for small proteins using currently available MR tools.

However, this '*ab initio* phasing with *de novo* models' method has limitations. Firstly, it can only handle relatively small proteins. The *ab initio* phasing method with *de novo* models needs highly accurate models. The *de novo* structure-prediction algorithm is limited to predicting the structures of small-size proteins because the large number of degrees of freedom associated with a polypeptide chain creates a very large conformational space to be sampled and evaluated. Secondly, the computation time is long. Typically, the method for *ab initio* phasing first produces a large number of *de novo* models and MR is then run on a few selected models. These models are chosen either using all-atom energy (Das & Baker, 2009; Qian *et al.*, 2007) or sometimes using a clustering method from a pool of *de novo* models. Lowest energy *de novo* models are selected using an all-atom energy function and are often very near to the native structure (Das *et al.*, 2007; Das & Baker, 2009), but these models are identified by comparing the all-atom energy after generating all models. This requires the completion of the simulation in order to identify the lowest energy models and it takes a large amount of time to generate all of these models. For example, even for proteins of about 100 residues at least 100 CPU days are needed to generate a few hundred thousand *de novo* models.

In this work, we propose an approach to accelerate *ab initio* phasing with *de novo* models. The approach primarily focuses on filtering the potentially good and bad all-atom models during folding. Whether MR can be used to determine the necessity for further all-atom refinement[1] soon after the coarse-grained models are changed to all-atom models was investigated. Subsequently, whether MR during folding can reduce the number of conformations to be sampled was checked. It was found that by initiating MR during *de novo* modelling the speed of *ab initio* phasing can be accelerated by more than two orders of magnitude. This would make the *ab initio* phasing with *de novo* models method more accessible to researchers with moderate computing resources.

## 2. Methods

Our method achieves acceleration of *ab initio* phasing with *de novo* models by filtering out unproductive models early and stopping the entire folding simulation once a few good models have been obtained. This is achieved by embedding the MR program *Phaser* inside the *Rosetta* code and executing *Phaser* at the beginning and each subsequent major stage of all-atom refinement in *Rosetta*. A folding simulation that generates a model with a *Phaser* score below the lower bound is terminated since there is little chance that this model could be refined to a quality sufficient to succeed in phasing. A folding simulation that generates a model with a *Phaser* score above the upper bound is also terminated without further all-atom refinement since successful phasing has been achieved. Only those models with a *Phaser* score that falls between the lower and upper bounds are permitted to undergo further all-atom refinement until either the upper bound is reached or the simulation has ended. When a few simulation runs have generated models with a *Phaser* score above the upper bound the entire simulation is terminated. This is achieved by communication between multiple runs spawned by the MPI process and through a kill signal sent to the individual MPI runs. This protocol is illustrated in Fig. 1.

[1] The word 'refinement' here refers to the optimization of model conformations against the empirical *Rosetta* all-atom energy function during structure prediction. It should not be confused with the refinement of coordinates against diffraction data during crystallographic structure determination.

### 2.1. *Rosetta* and *Phaser* modification

*Phaser* (McCoy *et al.*, 2007) is a program for phasing macromolecular crystal structures by molecular replacement using maximum likelihood. A standalone *Phaser* v.2.1.4, distributed in *PHENIX* v.1.3, was converted into an object-oriented version that is callable *via* a library (*Phaser* library). Although a Python interface to *Phaser* is available, the *Phaser* library was developed and called in *Rosetta* because of computational performance. *Rosetta* (Rohl *et al.*, 2004) is designed to predict the three-dimensional structure of a protein given its amino-acid sequence using a fragment-assembly approach. *Rosetta* v.3.0 was modified to incorporate *Phaser* in the all-atom refinement stage. The modified *Rosetta* is referred to as *RosettaX* in this paper. The *Phaser* translation-function *Z* score or rotation-function *Z* score (*Phaser* score) was employed to determine the fate of a folding-simulation run.

### 2.2. *Phaser* score threshold determination

To determine the optimal cutoff *Phaser* score, the success rate of MR trials on *de novo* models was estimated. In order to calculate the probability of success, a few thousand models with different random seeds were generated using the *Rosetta* fragment-assembly algorithm followed by all-atom refinement (Rohl *et al.*, 2004; Bradley *et al.*, 2005). *Rosetta* v.3.0 was used to produce *de novo* models. Fragments from homologues were included in the three-residue and nine-residue fragment libraries in order to generate good structures for this experiment. 11 structure-factor data sets (PDB entries 1ab6, 1be7, 1ctf, 1ig5, 1m6t, 1opd, 2fka, 2hsh, 2igd, 3chy and 6chy) were selected with different space groups. To check the consistency and reliability of the *Phaser* score, a group of 200 randomly selected models were used. *Phaser* v.1.4.2 (McCoy *et al.*, 2007) was subsequently executed using default parameters and with an estimated C$\alpha$ root-mean-square deviation (CA-RMSD) of 1.5 Å to perform the MR trials. Models after *Phaser* runs were verified using the difference in the CA-RMSD calculated using the rigid-body transformation and origin-permutation methods. This verification method is described in detail in §2.4. Small CA-RMSD differences indicates that models are likely to have succeeded in MR trials.

### 2.3. Data-set and model generation

In this study, a subset of structure factors and model sequences from Das & Baker (2009) were used. 20 structure-factor data sets with the number of molecules in the unit cell varying from one to four were selected with corresponding sequences in the range 50–130 residues. The other ten data sets could not be included owing to runtime errors while executing the parallel runs using the message-passing interface (MPI) on our machine. A fragment library (Simons *et al.*, 1997) for each of these sequences was generated using *RobettaServer* (Chivian *et al.*, 2003). It was further checked to confirm that structures of targets and of proteins homologous in sequence or structure were excluded from the fragment libraries. *RosettaX* was run to produce all the models on the RIKEN Integrated Cluster of Clusters (RICC), which is a massively parallel integrated cluster with 8000 cores. Since *Phaser* was incorporated into *RosettaX*, it was run using customized parameters with the top five orientations selected for translation search. *Phaser* was executed after each stage during all-atom conformation sampling, as shown in Fig. 1. Therefore, *Phaser* was run at least once and at most six times during each trajectory. The score from *Phaser* at each stage determined whether to continue or terminate the trajectory. The entire simulation was stopped when some good models according to *Phaser* score were produced. During production time, not only were decoys generated but intermediate lowest energy decoys were also tested against the diffraction data.

### 2.4. Molecular-replacement verification

Molecular-replacement verification is an important step to ensure that the MR solution is correct and unambiguous. Initially, the model after MR was moved to all permissible origins in the space group and the closest moved and symmetry-expanded model compared with the native structure was selected. The CA-RMSD to the native structure was computed after this origin permutation. Two publicly available programs, *origins.com* (http://bl831.als.lbl.gov/~jamesh/pickup/origins.com) and *match.py* (http://boscoh.com/protein/matchpy), were used to calculate CA-RMSD. Next, the CA-RMSD of the input model to the native structure was calculated using a rigid-body transformation (Kabsch, 1976), in which an optimal translation vector and rotation matrix were found that minimized the sum of the squared distances between corresponding atoms in the two coordinate sets. For successful MR solution, the CA-RMSD from the origin permutation should match that from the rigid-body transformation. Therefore, the difference between the CA-RMSDs of input models to the native structure computed after origin permutation and rigid-body transformation was used to assess the success of MR solution. It was observed that a small

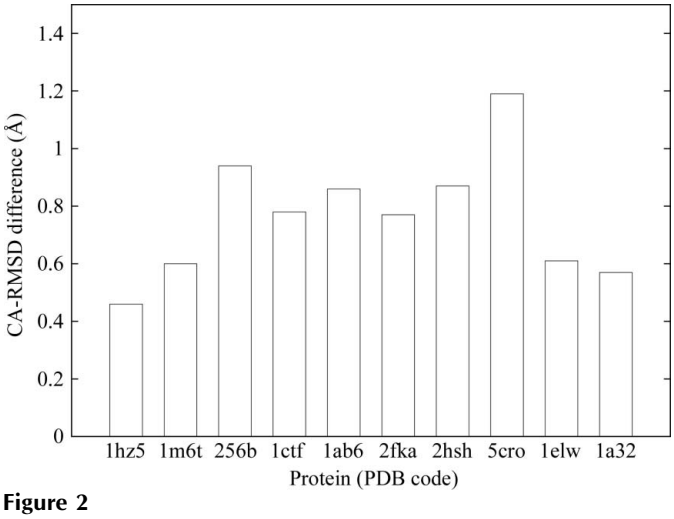


**Figure 2**
Determination of the threshold for MR verification. The *x* axis lists proteins in different space groups. The *y* axis is the difference in the CA-RMSD between the origin permutation and the rigid-body transformation.

**Table 1**
Protein targets successful in molecular replacement.

Here, r.m.s.d. indicates all-atom root-mean-square deviation. NA, not available.

| Structure factors | Sequence | Space group | No. of copies in ASU | Sequence length | Solvent content (%) | Resolution (Å) | $R_{merge}$ (%) | $d_{min}$ (Å) | No. of models targeted | Molecular-replacement solution | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Model No. | *Phaser* score | CA-RMSD, r.m.s.d. (Å) | $R$ factor, $R_{free}$ |
| 2igd | 1pgx | $P2_12_12_1$ | 1 | 55 | 45.0 | 1.10 | 0.037 | 10.10 | $3.00 \times 10^5$ | 88652 | 7.6 | 0.83, 1.70 | 0.20, 0.22 |
| 5cro | 5cro | $H32$ | 4 | 55 | 67.0 | 2.30 | 0.096 | 20.00 | $1.80 \times 10^5$ | 12431 | 9.1 | 0.80, 1.98 | 0.28, 0.34 |
| 1hz5 | 1hz6 | $P3_221$ | 2 | 61 | 71.2 | 1.80 | 0.066 | 30.00 | $1.60 \times 10^5$ | 123500 | 7.3 | 2.14, 3.86 | 0.26, 0.30 |
| 1hz6 | 1hz6 | $P2_12_12_1$ | 3 | 61 | 54.9 | 1.70 | 0.074 | 25.00 | $1.40 \times 10^5$ | 91579 | 8.5 | 2.14, 3.57 | 0.22, 0.28 |
| 1a32 | 1a32 | $P2_12_12_1$ | 1 | 70 | 38.0 | 2.10 | 0.075 | 100.00 | $2.80 \times 10^5$ | 7101 | 6.9 | 0.85, 1.79 | 0.30, 0.39 |
| 1ig5 | 1ig5 | $P4_32_12$ | 1 | 75 | 43.0 | 1.50 | 0.075 | 22.30 | $5.00 \times 10^5$ | 117040 | 7.0 | 2.36, 3.13 | † |
| 2hsh | 1aiu | $C2$ | 1 | 105 | 35.0 | 1.35 | 0.028 | 22.40 | $2.50 \times 10^5$ | 192086 | 6.9‡ | 2.12, 2.79 | 0.21, 0.25 |
| 1m6t | 256b | $C222_1$ | 1 | 106 | 42.8 | 1.81 | 0.300§ | 18.90 | $1.25 \times 10^5$ | 176 | 8.2 | 2.34, 2.91 | 0.24, 0.29 |
| 256b | 256b | $P1$ | 2 | 106 | 44.2 | 1.40 | NA | 32.29 | $1.50 \times 10^5$ | 1444 | 6.5 | 2.60, 2.90 | 0.22, 0.26 |
| 1elw | 1elw | $P4_1$ | 2 | 117 | 42.4 | 1.60 | 0.041 | 20.00 | $1.50 \times 10^5$ | 246 | 13.2 | 1.46, 1.97 | 0.22, 0.27 |

† *ARP*/*wARP* was not able to build the final models. ‡ TFZ and RFZ scores were used to assess the models and the higher value is reported. § $R_{merge}$ from PDB entry 1m6t.

difference in CA-RMSDs infers an unambiguous molecular-replacement solution. The cutoff was determined from the maximum difference found for an unambiguous solution of selected proteins as shown in Fig. 2. A structural difference in CA-RMSD of less than 1.20 Å was used as the cutoff in this study. This verification method cannot be employed in the absence of a crystal structure of the target protein, but this procedure was only used to determine threshold values of the *Phaser* score.

## 3. Results and discussion

### 3.1. *Phaser* score threshold

The *Phaser* score was used to evaluate the model quality. Two threshold values, upper and lower bounds, were estimated by calculating the probability of success in different intervals of the *Phaser* score. The translation-function $Z$ score (TFZ) value of the models was closely monitored in all space groups except $P1$. The rotation-function $Z$ score (RFZ) was employed to evaluate the model with a single molecule in space group $P1$, whereas the TFZ score of the last placed molecule was utilized when the unit cell contained multiple copies. In our experiments, the TFZ and RFZ scores showed better correlation than the TFZ score alone for polar space groups because the translation search is only performed over a plane in these space groups. Therefore, both the TFZ and the RFZ scores were considered in the case of polar space groups.

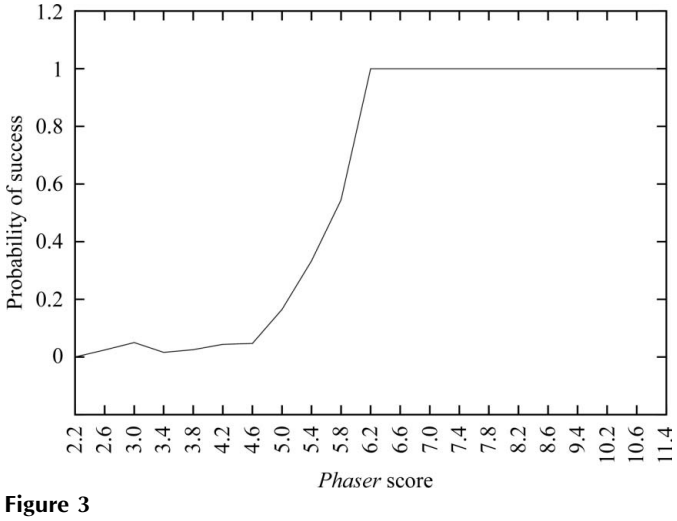


**Figure 3**
Probability of success of *de novo* models in MR using *Phaser*. The $x$ axis is the *Phaser* score with an average bin size of 0.4. The $y$ axis is the probability of success.

In the current work, the lower and upper thresholds of the *Phaser* score were determined by an extensive analysis of MR solutions from different space groups. *De novo* models were randomly chosen from different space groups. As shown in Fig. 3, the lower bound could be arbitrarily chosen as a *Phaser* score of between 4.6 and 5.8. However, the computation time and the probability that the *Phaser* score could reach a level where the success rate becomes 100% are crucial factors when selecting a threshold. It was noticed that it was very difficult to reach a *Phaser* score of 6.2 or more when the initial score was not closer. In our experiment, the lower bound *Phaser* score is set to 5.8 because it has a higher probability of success. The lower bound implies that a model with a score less than this threshold has little chance of improving the model quality for successful MR during all-atom refinement. For the upper bound, Fig. 3 showed that the probability of success reached 100% when the *Phaser* score was 6.2 or higher. A *Phaser* score of 7.2 was decided on as a conservative upper bound in this study. A higher threshold value was set in order to make sure that the model generated had succeeded in phasing by MR since subsequent simulation will be terminated. These two cutoffs can be customized by the end user. Models with a *Phaser* score between the lower and upper limits were allowed to undergo further *Rosetta* all-atom refinement because their structures could be changed to improve the accuracy.

### 3.2. Molecular-replacement solution and reduced conformational sampling

Near-native models can be generated during coarse-grained modelling and all-atom conformational sampling can improve the accuracy of these models, but completion of the simulation is required in order to identify these models. *A posteriori*

execution of MR after structure prediction needs a complete run to generate all decoys. Initiating MR during all-atom refinement could detect accurate models for phase estimation; thus, it is not necessary to wait until the completion of the structure-prediction simulation.

Out of the 20 structure-factor sets, ten diffraction sets succeeded in MR trials. Those molecules that produced successful MR solutions are shown in Table 1. Of the ten successful cases using *RosettaX* shown in Table 1, *Rosetta* was unable to find MR solutions for five molecules (2igd–1pgx, 5cro–5cro, 1ig5–1ig5, 2hsh–1aiu and 256b–256b) using 100 CPU days computation time; however, four of these molecules (with the exception being 1ig5–1ig5) were subsequently successfully phased in the 'large-scale' experiment (Das & Baker, 2009). A summary of cases of success and failure in phasing by *Phaser* using models generated by *RosettaX* compared with *Rosetta* from both 100 CPU day and large-scale experiments is shown in Table 2. The first few solutions for each target were monitored closely. The solutions for diffraction data and sequence combination of five models, 1m6t–256b, 256b–256b, 1elw–1elw, 1a32–1a32 and 5cro–5cro, were found very early. They required less than 15 000 models to be generated to obtain the first solution. For 1m6t–256b, which contains one molecule in the unit cell, only 176 conformers were needed to obtain a solution. In this case, instead of generating all of the requested models ($1.5 \times 10^5$), only 176 *de novo* models were sufficient for phasing. This reduced the computation time tremendously. Likewise, for 5cro–5cro, instead of generating all of the requested conformers ($1.8 \times 10^5$ models), the solution was obtained in the first 12 500 models. This is around 15 times fewer even though the unit cell contains four copies. For 2igd–1pgx, 1hz5–1hz6, 1hz6–1hz6,

**Table 2**
Summary of success and failure in phasing by MR using models generated by *RosettaX* and *Rosetta* from both 100 CPU day and large-scale experiments.

'0' represents failure and '1' represents success in phasing by *Phaser* using models generated by methods with protocols shown in the header of each column; '—' indicates that these structure–sequence pairs were not available with *RosettaX* owing to runtime errors while executing the parallel runs using the message-passing interface (MPI) on our machine.

| | | | | | | *Rosetta* | | |
|---|---|---|---|---|---|---|---|---|
| Structure factors | Sequence | Space group | No. of copies in ASU | Sequence length | Solvent content (%) | 100 CPU days | Large-scale | *RosettaX* |
| 2igd | 1pgx | $P2_12_12_1$ | 1 | 55 | 45.0 | 0 | 1 | 1 |
| 5cro | 5cro | $H32$ | 4 | 55 | 67.0 | 0 | 1 | 1 |
| 1hz5 | 1hz6 | $P3_221$ | 2 | 61 | 71.2 | 1 | 1 | 1 |
| 1hz6 | 1hz6 | $P2_12_12_1$ | 3 | 61 | 54.9 | 1 | 1 | 1 |
| 1a32 | 1a32 | $P2_12_12_1$ | 1 | 70 | 38.0 | 1 | 1 | 1 |
| 1ig5 | 1ig5 | $P4_32_12$ | 1 | 75 | 43.0 | 0 | 0 | 1 |
| 2hsh | 1aiu | $C2$ | 1 | 105 | 35.0 | 0 | 1 | 1 |
| 1m6t | 256b | $C222_1$ | 1 | 106 | 42.8 | 1 | 1 | 1 |
| 256b | 256b | $P1$ | 2 | 106 | 44.2 | 0 | 1 | 1 |
| 1elw | 1elw | $P4_1$ | 2 | 117 | 42.4 | 1 | 1 | 1 |
| 1be7 | 1bq9 | $H3$ | 1 | 51 | 43.0 | 0 | 0 | 0 |
| 1bq9 | 1bq9 | $P2_12_12_1$ | 1 | 51 | 41.5 | 0 | 0 | 0 |
| 1ctf | 1ctf | $P4_32_12$ | 1 | 68 | 41.9 | 0 | 0 | 0 |
| 1cm3 | 1opd | $P2_1$ | 1 | 85 | 27.9 | 0 | 0 | 0 |
| 1opd | 1opd | $P1$ | 1 | 85 | 32.6 | 0 | 0 | 0 |
| 2bc5 | 256b | $P2_12_12_1$ | 4 | 106 | 41.3 | 1 | 0 | 0 |
| 2fka | 2chf | $F432$ | 1 | 128 | 78.7 | 1 | 1 | 0 |
| 3chy | 2chf | $P2_12_12_1$ | 1 | 128 | 41.0 | 0 | 0 | 0 |
| 6chy | 2chf | $P2_12_12$ | 2 | 128 | 42.5 | 0 | 1 | 0 |
| 1ab6 | 2chf | $P3_1$ | 2 | 128 | 61.0 | 0 | 1 | 0 |
| 1aar | 1ubi | $P1$ | 2 | 71 | 35.0 | 0 | 0 | — |
| 1f9j | 1ubi | $I4_122$ | 2 | 71 | 60.0 | 0 | 0 | — |
| 1ubq | 1ubi | $P2_12_12_1$ | 1 | 71 | 33.0 | 0 | 1 | — |
| 2fcq | 1ubi | $P4_332$ | 2 | 71 | 58.0 | 0 | 0 | — |
| 2ojr | 1ubi | $P3_221$ | 1 | 71 | 73.0 | 0 | 0 | — |
| 1dt4 | 1dtj | $P4_22_12$ | 1 | 74 | 54.0 | 1 | 1 | — |
| 1dtj | 1dtj | $C2$ | 4 | 74 | 60.0 | 1 | 1 | — |
| 1a19 | 1a19 | $I4_1$ | 2 | 89 | 49.0 | 0 | 0 | — |
| 2hxx | 1a19 | $C2$ | 2 | 89 | 46.0 | 0 | 0 | — |
| 1mb1 | 1bm8 | $P4_12_12$ | 1 | 99 | 51.0 | 0 | 0 | — |

**Table 3**
Protein targets that were unsuccessful in molecular replacement.

NA, not available.

| Structure factors | Sequence | Space group | No. of copies in ASU | Sequence length | Solvent content (%) | Resolution (Å) | $R_{merge}$ (%) | $d_{min}$ (Å) | No. of models targeted | CA-RMSD, r.m.s.d.† (Å) | CA-RMSD, r.m.s.d.‡ (Å) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1be7 | 1bq9 | $H3$ | 1 | 51 | 43.0 | 1.67 | 0.068 | 30.00 | $3.50 \times 10^5$ | 2.27, 2.42 | 1.75, 2.07 |
| 1bq9 | 1bq9 | $P2_12_12_1$ | 1 | 51 | 41.5 | 1.20 | 0.060 | 50.00 | $3.60 \times 10^5$ | 2.14, 2.86 | 1.56, 1.68 |
| 1ctf | 1ctf | $P4_32_12$ | 1 | 68 | 41.9 | 1.70 | NA | 33.67 | $6.30 \times 10^5$ | 2.29, 3.21 | 1.26, 1.76 |
| 1cm3 | 1opd | $P2_1$ | 1 | 85 | 27.9 | 1.60 | NA | 10.00 | $1.80 \times 10^5$ | 3.45, 4.07 | 2.50, 3.20 |
| 1opd | 1opd | $P1$ | 1 | 85 | 32.6 | 1.50 | 0.029 | 19.89 | $3.80 \times 10^5$ | 3.19, 4.07 | 1.41, 2.09 |
| 2bc5 | 256b | $P2_12_12_1$ | 4 | 106 | 41.3 | 2.25 | 0.079 | 24.40 | $1.25 \times 10^5$ | 1.16, 1.85 | 1.10, 2.09 |
| 2fka | 2chf | $F432$ | 1 | 128 | 78.7 | 2.00 | 0.058 | 20.00 | $1.30 \times 10^5$ | 3.05, 3.46 | 1.77, 2.30 |
| 3chy | 2chf | $P2_12_12_1$ | 1 | 128 | 41.0 | 1.66 | NA | 35.48 | $2.00 \times 10^5$ | 2.70, 3.25 | 2.31, 2.82 |
| 6chy | 2chf | $P2_12_12$ | 2 | 128 | 42.5 | 2.33 | 0.051 | 30.86 | $1.50 \times 10^5$ | 2.93, 4.32 | 1.91, 3.19 |
| 1ab6 | 2chf | $P3_1$ | 2 | 128 | 61.0 | 2.20 | 0.048 | 15.00 | $1.30 \times 10^5$ | 2.74, 3.30 | 1.77, 2.16 |

† CA-RMSD and r.m.s.d. of the best model generated during our experiment without giving any biased constraints. ‡ Minimum model accuracy required for successful MR solution. Biased constraints from native models were used to generate models.

1ig5–1ig5 and 2hsh–1aiu, a larger conformational space was explored to find near-native models. Despite the larger search space, solutions were found earlier before the completion of simulation. The method selected all-atom *de novo* models and then also reduced conformation sampling.

### 3.3. Difficult targets for molecular replacement

In our experiment, the phasing of ten proteins was not successful, as shown in Table 3. Overall, *Rosetta* (Das & Baker, 2009) with 100 CPU days computing time successfully predicted *de novo* models for four targets (2bc5 and 2fka) which became harder targets for *RosettaX*. However, both methods were unsuccessful in generating models for successful phasing for the remaining eight targets. This is also summarized in Table 2. The potential causes for the failure of MR solution could be complex. The primary cause of failure might be the absence of accurate models during simulation. However, it is not clear what the minimum quality of the model required in order to obtain a successful MR solution is. One of the best models of the native structure for 2bc5 (space group $P2_12_12_1$, four molecules in the asymmetric unit) had a CA-RMSD of 1.25 Å and was unsuccessful in the MR trial. Meanwhile, a *de novo* model generated using the same sequence with a greater structural divergence from the native structure (∼3.0 Å CA-RMSD) gave a successful solution for 256b (space group *P*1, two molecules in the asymmetric unit). One possible reason for failure in this case could be the presence of four copies of the molecule in the asymmetric unit. As the number of copies in the asymmetric unit increases, finding an unambiguous MR solution becomes more difficult. However, *Phaser* was able to determine solutions for other multicopy targets. Thus, the reason for the unsuccessful MR solution for 2bc5 could be more than just the model accuracy or the presence of multiple copies in the asymmetric unit.

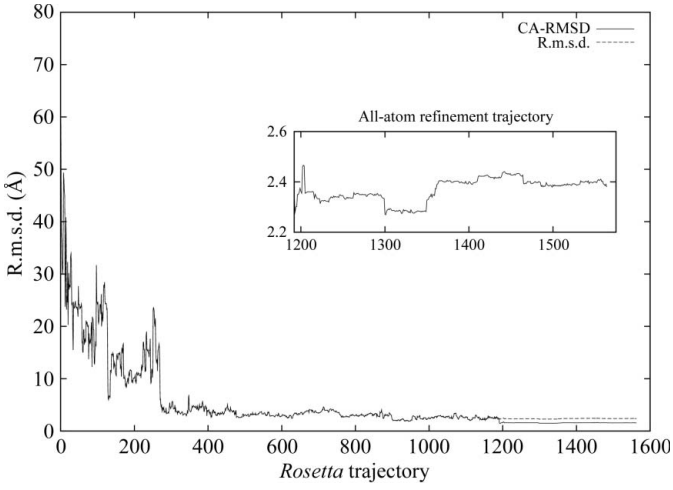


**Figure 4**
*Rosetta* trajectory during *de novo* structure prediction for 256b. The *x* axis is the model number generated during a *Rosetta* folding trajectory. The *y* axis is the r.m.s.d. of the generated model *versus* the native structure. In the inset, the course of all-atom refinement is shown. For this particular trajectory, the structure is not significantly changed during all-atom refinement. All-atom refinement started with a ramp-up stage from 1192 and finished at 1200. *Rosetta* performed aggressive sampling and produced conformations 1201–1491 (1201–1270 alternating wobble, 1271–1342 small wobble and 1343–1491 crank compound Monte Carlo minimization). Finally, fine optimization was carried out and generated conformations 1492–1564.

There could be numerous factors that affect the success of MR with a given model (Rigden *et al.*, 2008). One factor investigated was the minimum model quality required for successful MR. A group of 100 models were randomly selected from the pool of models generated using native constraints. The MR solution was verified after its execution and the maximum r.m.s.d. of the model to the native structure that gave an unambiguous solution was monitored. A boundary was obtained as shown in Table 3 in terms of the r.m.s.d. which was necessary to pass the MR trial. The quality of the predicted structure should be improved to at least the minimum accuracy necessary to enable successful MR. The minimum structural similarity to the native structure required to achieve successful MR cannot be generalized and varies case by case, as shown in Table 3. For example, *de novo* models of 6chy can be phased up to a CA-RMSD of 1.91 Å away from the native structure, but this was limited to 1.26 Å for 1ctf. 2bc5 was an exceptional case; a continuous trend in success of phasing in our constrained data set could not be observed, although the best model had a 0.99 Å CA-RMSD from the native structure. In addition, for a particular target, 2bc5, an MR solution was found when 100 CPU days were spent generating models;

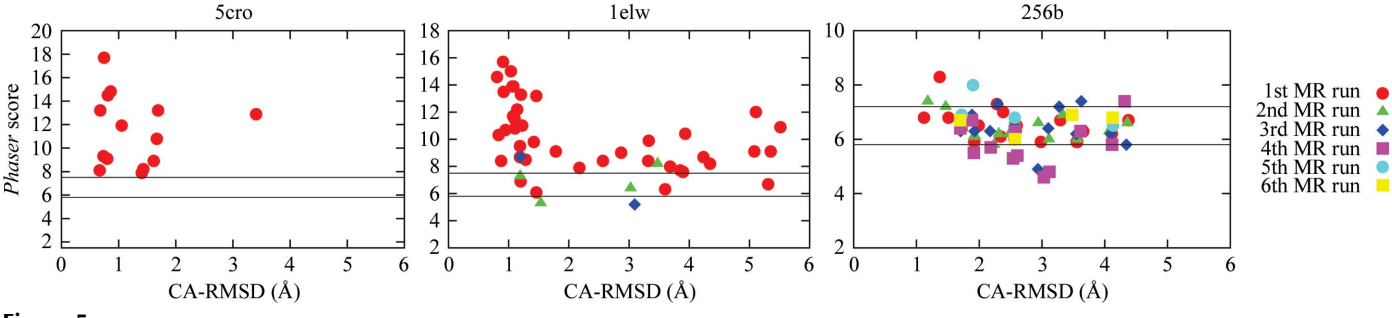


**Figure 5**
Distribution of CA-RMSD and *Phaser* score in MR trials during the *Rosetta* all-atom refinement procedure. The *x* axis is the CA-RMSD to the native structure and the *y* axis is the *Phaser* score. The two boundary lines are the thresholds determined using the *Phaser* score. The trajectories are terminated when the *Phaser* score is outside the accepted range. The red spheres indicate the first *Phaser* run; subsequent *Phaser* runs are represented with colours and shapes as indicated.

however, in large-scale conformation sampling it was unsuccessful in MR trials (Das & Baker, 2009). One promising method could be the trimming of possibly wrong regions to find the MR solution in practice (Rigden *et al.*, 2008).

### 3.4. *Rosetta* all-atom conformational sampling and molecular replacement

The global topology of models is assembled using randomly selected short fragments of known structures using the Monte Carlo strategy (Rohl *et al.*, 2004). Coarse-grained conformation sampling is thus responsible for generating near-native folds. One of the trajectories from *Rosetta* simulation for sequence 256b is shown in Fig. 4. It can be seen that all-atom refinement could not significantly alter the protein structure. However, the accuracy of the overall structure could be improved when all-atom refinement of low-resolution protein structure models was mainly focused on the regions that were most likely to contain errors (Qian *et al.*, 2007).

The MR method *Phaser* was set up to execute in the course of all-atom refinement from the beginning in our approach. Both *Rosetta* all-atom refinement and *Phaser* are computationally intensive. To enable effective computing, *Phaser* has been set to run with customized parameters (select top five orientations) to reduce the elapsed time. *Phaser* was also executed with default parameters on these input models to confirm the effect of using customized options. There were no changes in the final result. Since *Phaser* was executed several times during a *Rosetta* trajectory, it always had a greater chance of success with new and potentially more accurate models.

As shown in Fig. 5, all *de novo* models of 5cro and most models of 1elw passed the upper threshold of a predefined *Phaser* score in the first MR run. This implies that no further refinement is necessary for those models. Improvements in accuracy and *Phaser* score (256b in Fig. 5) were not significant when *Phaser* was run multiple times during folding. This trend was also observed in other molecules. Hence, phasing could be easier and achieved earlier when coarse-grained models were predicted close to the native fold.

When a huge amount of CPU time was spent in searching all-atom conformational space (Das & Baker, 2009; Qian *et al.*, 2007) phasing of difficult targets could be successful. However, it could be more efficient if MR were used to select models during structure prediction. From this perspective, stopping the refinement of bad models and using this time for the generation of further new conformations would improve the current *ab initio* phasing method with *de novo* models.

As an alternative means of assessing the success of MR solution and confirming the utility of *de novo* models for

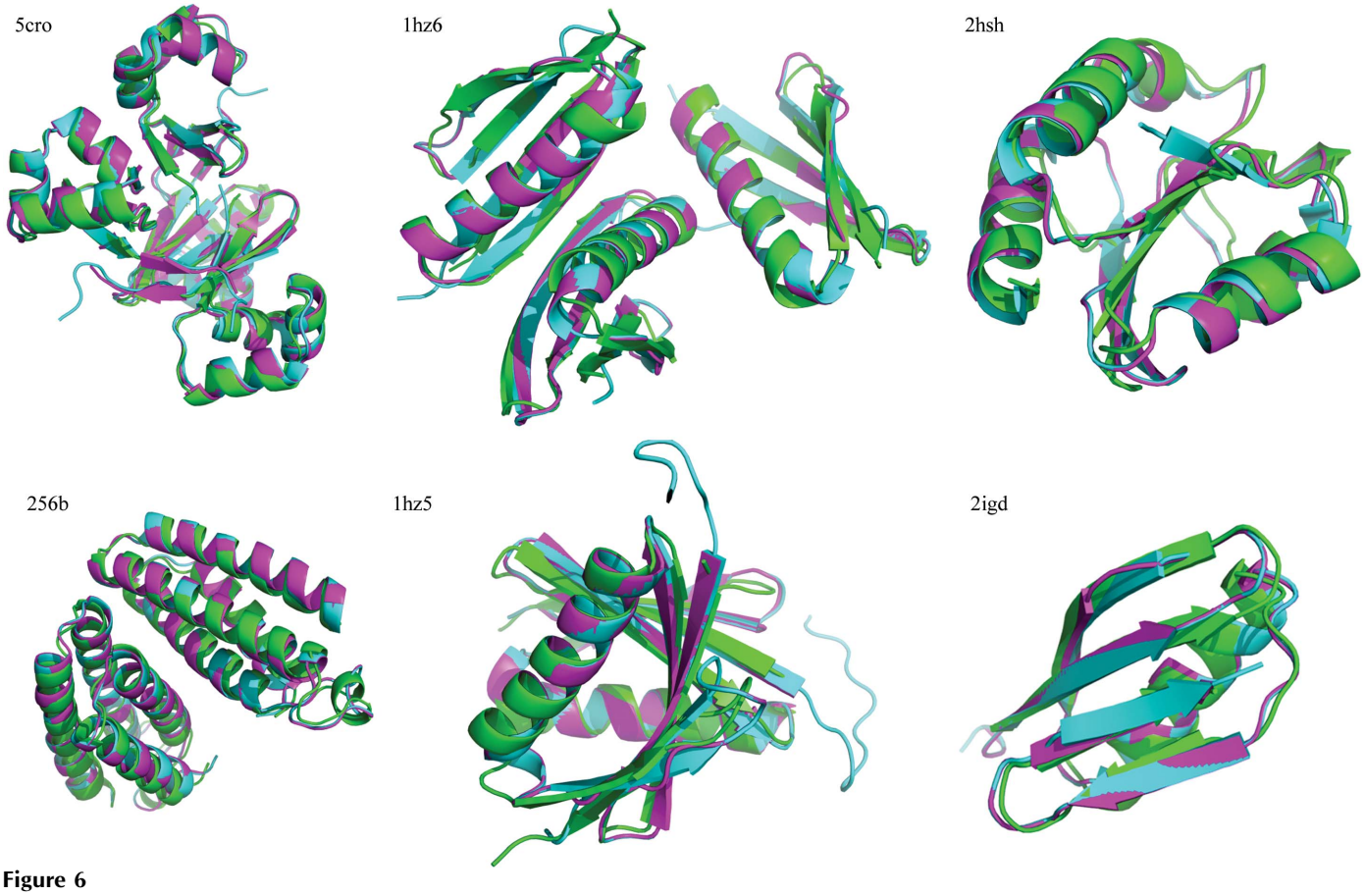


**Figure 6**
Superposed structures of successful MR models and models after *ARP*/*wARP* with native structures. Native structures are shown in cyan. Successful MR models are shown in green. Models built by *ARP*/*wARP* after MR are shown in magenta.

phasing, *ARP/wARP* v.7.1 (Perrakis *et al.*, 1999) was used to rebuild and complete the models obtained by MR because it can generate a complete and accurate model if the MR solution is correct (Cohen *et al.*, 2008). *ARP/wARP* was executed using default parameters on successful MR models to confirm the quality of the electron-density map. This widely used tool produced highly accurate models close to the crystal structures from the PDB, as shown in Fig. 6. *R*-factor and $R_{\rm free}$ values for each successful target are shown in Table 1. *ARP/wARP* was not able to build the final model of one molecule (1ig5) using default parameters. However, the model after molecular replacement was very close to the native structure when the model was manually inspected.

### 3.5. Elapsed time after *Phaser* in *Rosetta* all-atom refinement

Computation time is the primary concern in this study because the *Rosetta* all-atom refinement algorithm requires a huge amount of computing power. Although diffraction data can be used to drive the all-atom conformational sampling, computation time will be greatly increased when an MR method such as *Phaser* runs at every step of the protocol. The number of copies in the unit cell can also increase the CPU time. The quality of *de novo* models was found to determine the running time of *Phaser*; when *de novo* models are very near to the native structure *Phaser* gives the solution very early, while inaccurate models take a very long time without any solution.

In our protocol, the elapsed time was reduced in three ways. Firstly, *de novo* models that crossed the upper bound of the *Phaser* score were prevented from undergoing computationally expensive all-atom refinement. Secondly, the trajectories of models with *Phaser* scores below the lower bound were also terminated. Finally, when a few models had *Phaser* scores higher than the upper bound the entire simulation was terminated; this reduced the computation time tremendously.

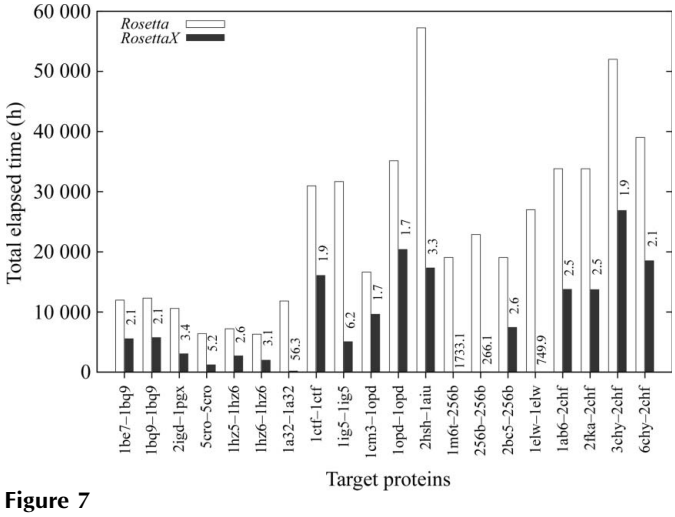


**Figure 7**
Total elapsed time of protein-folding simulation for *Rosetta* and *RosettaX*. Structure-factor and model-sequence pairs are shown on the *x* axis and the *y* axis shows the total elapsed time in hours. The labels above the black bars (*RosettaX*) indicate the ratio of elapsed time spent by *Rosetta* and *RosettaX*.

Alternatively, the elapsed time could be significantly reduced if MR were run after each trajectory because the entire simulation could be stopped after a few solutions were found. However, this approach would not be as efficient as our approach. Firstly, this procedure would miss the elapsed time saved by avoiding all-atom refinement of good-quality and bad-quality models. Secondly, in a massively parallel run our method can detect suitable models for phasing early in a trajectory and can send the message to the master node earlier to convey whether further conformational sampling is necessary. Thirdly, small structural changes in protein structure can have a large impact on MR (Giorgetti *et al.*, 2005); running the MR program multiple times (as in our protocol) is more likely to identify successful MR models.

Both *Rosetta* and *RosettaX* were independently run in the same cluster to compute the total elapsed time. The number of models to be generated by *Rosetta* for each target was taken from Das & Baker (2009) for 100 CPU day simulation and the same protocol was used. The same number of models was set to be generated by *RosettaX* at the start of the run; however, the actual number of models generated was significantly smaller owing to the termination criteria used. The actual elapsed time was monitored for both programs. As shown in Fig. 7, *RosettaX* appears to be an efficient method in terms of computation time even for unsuccessful proteins without compromising the efficiency. Among the 20 selected targets, *Rosetta* succeeded in finding solutions for 12 targets when large-scale CPU time was spent in generating models; the number of successful cases was subsequently reduced to seven with 100 CPU days computation time (Das & Baker, 2009). In our method, ten cases out of 20 targets were successful in phasing. This result suggests that *RosettaX* has achieved acceleration without reducing its effectiveness compared with *Rosetta*.

Our method saved a huge amount of computational power when MR solutions were found earlier (1a32–1a32, 256b–256b, 1m6t–256b and 1elw–1elw). Even when the unit cell contained more than one copy of the molecule (5cro–5cro, 1hz6–1hz5 and others), the computation time used by *RosettaX* was less than half of that used by *Rosetta* to find a solution. Overall, *RosettaX* is 142 times faster than Rosetta on our benchmark data set when an average of the ratio of elapsed times for each molecule is computed. During the production run, thousands of protein models were generated and only a few of them were able to determine unambiguous MR solutions. Our method efficiently evaluated all the models and selected accurate models quickly.

## 4. Conclusions

In this study, we explored the efficient use of the MR method in *Rosetta* all-atom conformational sampling instead of running it as an *a posteriori* process on selected models for phasing after the completion of a *Rosetta* run. The important aspect of our approach to *ab initio* phasing with *de novo* models is to determine the solution as early as possible. Firstly, MR solutions are found very early in all-atom refinement soon after

the coarse-grained models are changed to all-atom models. Secondly, the number of conformations to be generated to achieve successful MR is reduced. The total elapsed time of simulation is reduced by more than two orders of magnitude. Our method will expand the utility of *ab initio* phasing with *de novo* models owing to the significantly reduced computational time required. These results suggest that our approach is an efficient way to phase small proteins with novel folds using *de novo* models.

The effectiveness of all-atom refinement of *de novo* models is determined by the accuracy of the backbone conformation in coarse-grained models. All-atom refinement was not able to adequately improve the overall accuracy of the structure for MR when backbone atoms were not accurately predicted. *Ab initio* phasing with *de novo* models can be easier if the model quality is closer to the native structure; however, the minimum accuracy of protein models for phasing varies from case to case.

We wish to thank the Advanced Centre for Computing and Communication, RIKEN, Japan for the computing resources of the RIKEN Integrated Cluster of Clusters (RICC) system. We are grateful to Dr David Baker and Dr Randy Read for making the *Rosetta* and *Phaser* source codes available. We thank Dr James Holton and Dr Bosco Ho for the source codes of *origins.com* and *match.py*. We thank Dr Rhiju Das for helpful e-mails. We thank Dr Yong Zhou for stimulating discussions. We acknowledge the Initiative Research Unit program from RIKEN, Japan for funding.

## References

Blow, D. M. & Rossmann, M. G. (1961). *Acta Cryst.* **14**, 1195–1202.
Bradley, P., Misura, K. M. & Baker, D. (2005). *Science*, **309**, 1868–1871.
Chivian, D., Kim, D. E., Malmström, L., Bradley, P., Robertson, T., Murphy, P., Strauss, C. E., Bonneau, R., Rohl, C. A. & Baker, D. (2003). *Proteins*, **53**, 524–533.
Cohen, S. X., Ben Jelloul, M., Long, F., Vagin, A., Knipscheer, P., Lebbink, J., Sixma, T. K., Lamzin, V. S., Murshudov, G. N. & Perrakis, A. (2008). *Acta Cryst.* D**64**, 49–60.
Das, R. *et al.* (2007). *Proteins*, **69**, 118–128.
Das, R. & Baker, D. (2009). *Acta Cryst.* D**65**, 169–175.
Giorgetti, A., Raimondo, D., Miele, A. E. & Tramontano, A. (2005). *Bioinformatics*, **21**, ii72–ii76.
Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.
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.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G. & North, A. C. (1960). *Nature (London)*, **185**, 416–422.
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.
Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. (2004). *Methods Enzymol.* **383**, 66–93.
Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). *J. Mol. Biol.* **268**, 209–225.