Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Molecular replacement using *ab initio* polyalanine models generated with *ROSETTA*

Daniel J. Rigden,[a]* Ronan M. Keegan[b] and Martyn D. Winn[b]

[a]School of Biological Sciences, University of Liverpool, Crown Street, Liverpool L69 7ZB, England, and [b]STFC Daresbury Laboratory, Daresbury, Warrington WA4 4AD, England

Correspondence e-mail: drigden@liv.ac.uk

Received 11 July 2008
Accepted 13 October 2008

The success of the molecular-replacement method for solving protein structures from experimental diffraction data depends on the availability of a suitable search model. Typically, this is derived from a previously solved structure, sometimes by homology modelling. Very recently, Baker, Read and coworkers have demonstrated a successful molecular-replacement case based on an *ab initio* model generated by *ROSETTA* [Qian *et al.* (2007), *Nature (London)*, **450**, 259–264]. In this contribution, a number of additional test cases in which *ab initio* models generated using modest computational resources give correct molecular-replacement solutions are reported. Unsuccessful cases are also reported for comparison and the factors influencing the success of this route to structure solution are discussed.

1. Introduction

Molecular replacement (MR) is one of the key methods available for supplementing experimental diffraction data with phasing information and thereby solving the 'phase problem'. Continual improvement of MR methodology has allowed the effective use of increasingly inaccurate and/or incomplete search models (Evans & McCoy, 2008). Historically, the models successfully used for MR have been either existing experimentally determined structures or predicted structures produced by conventional homology modelling.

The past few years have seen the progressive maturation of a fundamentally different methodology for protein structure determination: *ab initio* modelling. Such methods produce large numbers of models (or 'decoys') through combining fragments of known structures thought to be compatible with local sequences. The models are clustered and the conformations represented by large clusters, particularly the largest, are supposed to be the most likely to accurately represent the true structure. *Ab initio* modelling through more extensive search protocols has recently produced highly accurate structures (Bradley *et al.*, 2005; Qian *et al.*, 2007).

An examination of the ability of *ab initio* modelling to produce models that are sufficiently accurate for the solution of structures by MR in cases where there is no related experimental structure is timely. A recent paper (Qian *et al.*, 2007) provided a valuable proof of principle but used supercomputers as well as the distributed computing of *Rosetta@home* (http://boinc.bakerlab.org/rosetta/). Here, we consider whether models generated with *ROSETTA* running on more typical hardware can be used successfully for MR and assess the future potential of this route to structure solution.

2. Materials and methods

2.1. Preparation of models

A set of test cases was collected by searching the Protein Data Bank (PDB; Berman *et al.*, 2007) for proteins of fewer than 100 residues that were determined to better than 2.2 Å resolution and for which experimental data were available. A maximum of 30% sequence identity to previously determined structures was allowed

Table 1
Results of trials for which a correct MR solution was found.

The contents of the asymmetric unit (ASU) are given as the number of copies of the target protein multiplied by the number of residues in the most complete chain deposited. The presumed biological unit is given in parentheses. The size of the largest *ROSETTA* cluster (cluster 0) is given as a percentage of the total number of models, followed by the location of the best model. N_{match} is the match of the best *ROSETTA* model to the deposited structure expressed as number of C^{α} atoms matched; RMS_{match} is the r.m.s. deviation of matched C^{α} atoms (Å). RMS_{sol} is the r.m.s. deviation on C^{α} atoms (Å) for each copy of the target protein in the best MR solution (r.m.s. deviations for ensembles refer arbitrarily to the first member in each case).

PDB code	Structural class	Resolution (Å)	Contents of ASU (residues)	No. of *ROSETTA* models	Size of cluster 0 (%)	Cluster of top model	N_{match}; RMS_{match}	MR attempts (singly and in superposition; complete or polyalanine models)	MR successes	RMS_{sol}
2pmr	α	1.32	1 × 76 (dimer)	7000	75.9	0	70; 1.324	Nine models from clusters 0 and 2 or one model from cluster 0, truncated at both termini	Ensembles of 5 or 9 all-atom models or single all-atom model	0.95 for single all-atom model
2nn4	α	2.10	3 × 62 (monomer)	3000	48.3	1	59; 1.245	Seven models from cluster 1	Ensemble of 7 all-atom models OR single polyalanine model	1.98, 2.03, 2.08 for ensemble of 7 all-atom models
2fzt	α	2.05	2 × 78 (dimer)	5000	47.3	0	58; 1.279	Nine models from cluster 0, truncated or untruncated at the C-terminus	Ensemble of 9 polyalanine models (truncated or untruncated) OR single polyalanine model (truncated)	1.69, 1.62 for ensemble of 9 polyalanine models (truncated)
2o3l	α	2.05	2 × 81 (monomer)	3000	38.8	0	70; 1.385	Six models from cluster 0	Ensemble of 6 all-atom models	2.45, 3.47 for ensemble of 6 all-atom models
2duy	α	1.75	1 × 65 (dimer)	3000	11.1	0	57; 1.559	Seven models from clusters 0, 1 and 6, truncated at the N-terminus	Ensembles of 4, 6 or 7 polyalanine models	1.78 for ensemble of 4 polyalanine models

and most of our cases were in fact novel folds. We thus confined ourselves to cases for which homology modelling would not have been readily applicable. The selected structures were deposited no earlier than 2006, so that the fragment libraries employed in model construction contained no pieces of closely related proteins, forcing *ROSETTA* to operate in a 'pure' *ab initio* mode.

For each test case, at least 3000 models were produced and clustered by *ROSETTA* v.2.1.2 (Simons *et al.*, 1997, 1999; Shortle *et al.*, 1998) using default protocols, with secondary-structure predictions provided by *PSIPRED* (Jones, 1999). *ROSETTA* was run on Dell Precision workstations with Xeon processors and typically took 13–25 processor hours to produce a set of 3000 polyalanine models using a single processor. Side chains were added to the polyalanine models produced by *ROSETTA* using *SCWRL* (Dunbrack & Cohen, 1997) and limited energy minimization was carried out with *MODELLER* (Sali & Blundell, 1993). In some cases, limited manual editing of the models was applied, in particular the truncation of divergent regions at the termini. Uniform *B* factors of 20 $Å^2$ were applied to the models.

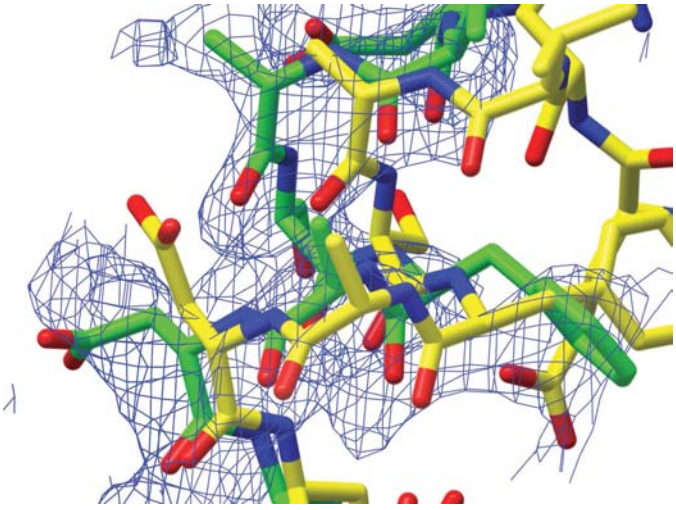

Figure 1
Illustration of the solution of 2o3l using an ensemble of six all-atom models. For clarity, only one member of the ensemble is shown (backbone in yellow). For comparison, the deposited structure is also shown (backbone in green). The electron-density map is calculated from the weighted map coefficients output by *DM* (Cowtan, 1994) following phase improvement of the *Phaser* solution of the ensemble. This figure was prepared with *CCP4MG* (Potterton *et al.*, 2004).

Models were compared individually with the known crystal structure in order to decide whether or not to proceed with MR attempts and, if so, with which models. The Match Index measurement, which includes consideration of match length, match r.m.s. and overall length of matched proteins, was used to rank the models with *LSQMAN* (Kleywegt, 1999).

2.2. Molecular replacement

Phaser (McCoy *et al.*, 2007) was used for MR attempts and *REFORIGIN* (Collaborative Computational Project, Number 4, 1994) was used to compare *Phaser* solutions with the deposited structure. MR was attempted with individual models and with ensembles and with or without added side chains from *SCWRL*. The predicted r.m.s. error parameter required by *Phaser* was typically set to around 2.5 Å, but was varied on a case-by-case basis. *MrBUMP* (Keegan & Winn, 2007) was used to automatically process sets of models.

Our initial criterion for success was placement of the model to less than 2.2–2.8 Å r.m.s. deviation from the deposited structure (the more relaxed criterion being applied to untruncated models with frequently divergent termini). A more stringent criterion for success is the ability to proceed to a refined structure from the MR solution. To this end, where MR solutions were successfully obtained, automatic model rebuilding was attempted with *ARP/wARP* (Cohen *et al.*, 2008). Where this was unsuccessful, electron-density maps were inspected for evidence that manual rebuilding could succeed.

3. Results

Of the 16 test cases considered, ten produced models that were promising enough to be tried in MR. In the other six cases, there were no models close to the known crystal structure, as could have been predicted from the poor modelling results. Of the ten promising cases, *ab initio* models gave correct MR solutions in five cases as judged by comparison with the deposited structures (Table 1). In the cases of

Table 2
Results for cases in which MR was unsuccessful.

Columns are as in Table 1.

PDB code	Structural class	Resolution (Å)	Contents of ASU (residues)	No. of *ROSETTA* models	Size of cluster 0 (%)	Cluster of top model	N_{match}; RMS_{match}	MR attempts (singly and in superposition; complete or polyalanine models)
2dsy	$\alpha+\beta$	1.90	4 × 80 (tetramer)	5000	5.1	14	57; 1.541	Single best model, from cluster 14, truncated at both termini
2gf4	α	2.07	2 × 89 (tetramer)	5000	18.7	5	55; 1.774	Single best model, from cluster 5, truncated at C-terminus
2i5u	α	1.50	1 × 77 (monomer)	3000	14.1	31	60; 1.468	Six models from cluster 3, plus one model from cluster 31 truncated at N-terminus
2nzc	α/β	1.95	4 × 80 (tetramer)	3000	14.4	5	50; 1.704	Five models from cluster 5, truncated at both termini
2o6k	α	2.10	2 × 72 (monomer)	3000	2.4	21	31; 0.846	Single best model, from cluster 21, truncated at N-terminus

2pmr and 2nn4, the MR solution could be automatically rebuilt in *ARP*/*wARP* (74 of 76 and 174 of 186 residues built and docked, respectively) and refined to results that were near-indistinguishable from the crystal structures. Thus, for at least two cases *ab initio* models were sufficient to obtain a complete structure solution.

For the other three cases, although we know from comparison with the deposited solution that the *ab initio* model has been correctly placed, the automated methods used in this short study were not sufficient to complete the structure solution. For 2fzt, an ensemble of nine polyalanine models finds both copies of the target protein to within about 1.6 Å r.m.s. deviation from the deposited model. *ARP*/*wARP* is able to rebuild 69 out of 156 residues of the main chain, but is unable to dock any significant portion of the sequence.

For 2o3l, the first copy is located easily with an ensemble of six all-atom models. However, location of the second copy requires a more exhaustive MR search using a non-default protocol in *Phaser*. Although the MR solution is essentially correct, the accuracy, particularly of the second copy, is not good and automatic rebuilding fails. Nevertheless, inspection of the electron density suggests that it may be possible to complete structure solution. In the lower half of Fig. 1, we see that the MR solution is essentially correct, although the density suggests a change to the side-chain conformation of the aspartate residue. The upper half shows the beginning of a loop which is displaced from the model by an intermolecular contact. Importantly, the electron density here clearly reflects the deposited model rather than the MR solution. Finally, an ensemble of four polyalanine models for 2duy gives a correct solution in MR. However, the *ab initio* models are the least accurate of those in Table 1 and automatic model rebuilding fails.

The models yielding MR solutions varied in type. All-atom models including side chains from *SCWRL* worked for 2pmr and 2o3l and polyalanine models worked for 2fzt and 2duy, while 2nn4 could be solved with either. Ensemble models were required for 2o3l and 2duy, whereas the others could also be solved with single models. Truncation of flexible termini regions which were inaccurately modelled was necessary for success with 2pmr and 2duy. In the case of 2fzt, correct solutions were obtained for a polyalanine ensemble whether truncated or not (Table 1), although the truncated ensemble gave a slightly more accurate solution.

It must be emphasized that the models chosen for testing were not blindly chosen, as representatives of top clusters for example, but were instead those known *a priori* to most closely resemble the true structure. Nevertheless, an analysis of the successful cases and the common factors in the editing required strongly suggests that protocols to automatically select and process models could be devised that would be capable of providing sets of models with which MR attempts would be justified. For example, trimming of excessively divergent termini from cluster-derived ensembles could be readily automated.

We analysed several factors that might be expected to influence the success of MR with *ab initio* models. We find that greater success in the modelling, as defined by the existence of top clusters including a significant proportion of the overall model set, leads to success in MR. Thus, for successful MR cases the top cluster contained on average 44.3% of all models, while this figure was just 10.9% for unsuccessful MR cases (see Table 2) and 4.6% for the six examples where MR was not attempted. Evidently, the six cases where we chose not to test MR could have been selected *a priori* through the clear failure of their modelling. The greater success of the *ab initio* modelling in the 'MR successful' category is also evident in the fact that the most accurate model overall within this group was found in the top cluster of models or, in a single case, the second largest cluster. It is probably also no coincidence that all of the proteins in the 'MR successful' category belong to the all-α structural class for which *ab initio* modelling is particularly successful, in part owing to better secondary-structure predictions (see, for example, Karypis, 2006). Conversely, the six cases not selected for MR trials comprise three $\alpha+\beta$ folds, the single all-β structure in our original set of proteins and two further all-α folds.

The oligomeric state of the protein would be expected to influence modelling success since protein–protein interfaces are different from solvent-exposed protein surfaces. Consistent with this idea, the present cases show success with monomers and dimers, but not with higher oligomers. Interestingly, however, in the present cases the resolution of the available diffraction data seems not to influence the chance of success: both the 'MR successful' and the 'MR unsuccessful' categories contain structures determined to mean resolutions of between 1.85 and 1.90 Å. Finally, larger numbers of protein subunits per asymmetric unit would be expected to make the MR solution more difficult. Indeed, although successful examples contain up to three subunits, the mean number of subunits per ASU is slightly lower in the 'MR successful' category than in the 'MR unsuccessful' group (1.8 *versus* 2.6).

Illustrative figures of the solutions obtained have been deposited as supplementary material[1].

4. Discussion

These data are preliminary but offer strong encouragement to further explore the potential of *ab initio* models for structure solution by MR. Importantly, the production of suitable models for MR is strongly linked to the overall success of the *ab initio* modelling. This in turn is reflected by the size of the top cluster; thus, it can be predicted whether efforts at MR have a chance of success from the results of *ab*

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: FW5187). Services for accessing this material are described at the back of the journal.

initio modelling. In addition, we note two advantageous correspondences between *ab initio* modelling and MR. Firstly, *ab initio* modelling involves the production of ensembles of clustered models: such ensembles are often more effective than single structures in MR attempts (Leahy *et al.*, 1992; Table 1). Secondly, MR may be successful for rather incomplete models: we note that *ab initio* models are often very similar in core regions (sometimes even between clusters) but differ at termini or in loops. Such structural divergences could in future be automatically identified and removed prior to MR. We also note some obvious possibilities for improving performance, such as defining *B* factors in models according to structural variation. This could be within the ensemble of structures used for MR or even more broadly among the set of *ab initio* models. The *ab initio* modelling step could also be applied to the smallest identifiable homologue of the target protein rather than necessarily to the target itself. Finally, rather than attempt complete side-chain placement on to a potentially inaccurate backbone, only more reliable side chains, such as those with few well populated rotamers, could be modelled.

It is clear from our negative results that even for small target proteins producing an *ab initio* model on typical hardware that is sufficiently accurate for MR remains challenging. Nevertheless, the successful case studies presented here show that it is now possible in principle and may be a more convenient approach than experimental phasing in certain cases. Qian *et al.* (2007), working with a more recent computer-intensive all-atom algorithm of *ROSETTA*, were able to find MR solutions for the CASP7 target T0283 using an *ab initio* model truncated at both termini. T0283 is an all-α protein, as are the proteins in our 'MR successful' category. At 112 residues, T0283 is somewhat larger than the proteins considered here and its successful prediction may well be related to the more advanced algorithm employed. Unfortunately, it may be some time before typically available computing hardware is sufficiently powerful for routine use of the all-atom algorithm of *ROSETTA*. Dodson (2007) commented that the use of *ab initio* models in MR should be 'assessed further using known structures' and we have shown that *ab initio* models may indeed work even with relatively modest computing resources.

Presently, success with *ROSETTA* is limited to around 100–120 residues. This limits its use at present, although we note that this is the most common size for a protein domain (Wheelan *et al.*, 2000). However, with recent improvements in speed (Wu *et al.*, 2007) and size capability (Taylor *et al.*, 2008) providing further evidence of the upward trajectory of *ab initio* methodology development and with the ever-increasing availability of computing power, we are encouraged to embark on a larger more detailed analysis of the performance of *ab initio* models in MR experiments. Ultimately, *ab initio* modelling may be seamlessly incorporated into automatic MR pipelines.

References

Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. (2007). *Nucleic Acids Res.* **35**, D301–D303.

Bradley, P., Misura, K. M. & Baker, D. (2005). *Science*, **309**, 1868–1871.

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.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cowtan, K. (1994). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.

Dodson, E. J. (2007). *Nature (London)*, **450**, 176–177.

Dunbrack, R. L. Jr & Cohen, F. E. (1997). *Protein Sci.* **6**, 1661–1681.

Evans, P. & McCoy, A. (2008). *Acta Cryst.* D**64**, 1–10.

Jones, D. T. (1999). *J. Mol. Biol.* **292**, 195–202.

Karypis, G. (2006). *Proteins*, **64**, 575–586.

Keegan, R. M. & Winn, M. D. (2007). *Acta Cryst.* D**63**, 447–457.

Kleywegt, G. J. (1999). *Acta Cryst.* D**55**, 1878–1884.

Leahy, D. J., Axel, R. & Hendrickson, W. A. (1992). *Cell*, **68**, 1145–1162.

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.

Potterton, L., McNicholas, S., Krissinel, E., Gruber, J., Cowtan, K., Emsley, P., Murshudov, G. N., Cohen, S., Perrakis, A. & Noble, M. (2004). *Acta Cryst.* D**60**, 2288–2294.

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J. & Baker, D. (2007). *Nature (London)*, **450**, 259–264.

Sali, A. & Blundell, T. L. (1993). *J. Mol. Biol.* **234**, 779–815.

Shortle, D., Simons, K. T. & Baker, D. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 11158–11162.

Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). *J. Mol. Biol.* **268**, 209–225.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999). *Proteins*, **34**, 82–95.

Taylor, W. R., Bartlett, G. J., Chelliah, V., Klose, D., Lin, K., Sheldon, T. & Jonassen, I. (2008). *Proteins*, **70**, 1610–1619.

Wheelan, S. J., Marchler-Bauer, A. & Bryant, S. H. (2000). *Bioinformatics*, **16**, 613–618.

Wu, S., Skolnick, J. & Zhang, Y. (2007). *BMC Biol.* **5**, 17.