Acta Cryst D STRUCTURAL BIOLOGY
ISSN 2059-7983

Initiating heavy-atom-based phasing by multi-dimensional molecular replacement

Bjørn Panyella Pedersen,* Pontus Gourdon,‡ Xiangyu Liu,§ Jesper Lykkegaard Karlsen and Poul Nissen

Centre for Membrane Pumps in Cells and Disease, Danish National Research Foundation, Department of Molecular Biology and Genetics, Aarhus University, Gustav Wieds Vej 10C, DK-8000 Aarhus, Denmark. *Correspondence e-mail: bpp@mbg.au.dk

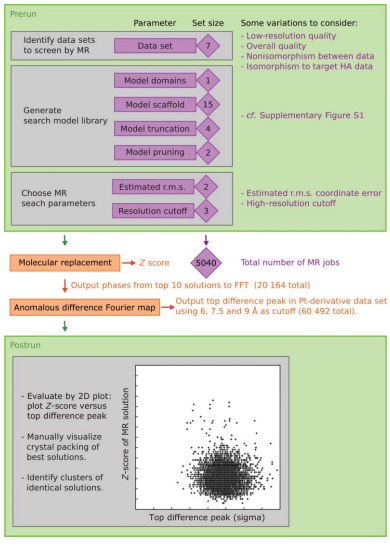
Received 12 January 2015
Accepted 24 November 2015

‡ Current address: Department of Biomedical Sciences, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark and Department of Experimental Medical Science, Lund University, Sölvegatan 19, SE-221 84 Lund, Sweden.
§ Current address: School of Medicine, Tsinghua University, Beijing 100084, China.

Keywords: experimental phasing; molecular replacement; heavy-atom substructure.

Supporting information: this article has supporting information at journals.iucr.org/d

To obtain an electron-density map from a macromolecular crystal the phase problem needs to be solved, which often involves the use of heavy-atom derivative crystals and concomitant heavy-atom substructure determination. This is typically performed by dual-space methods, direct methods or Patterson-based approaches, which however may fail when only poorly diffracting derivative crystals are available. This is often the case for, for example, membrane proteins. Here, an approach for heavy-atom site identification based on a molecular-replacement parameter matrix (MRPM) is presented. It involves an *n*-dimensional search to test a wide spectrum of molecular-replacement parameters, such as different data sets and search models with different conformations. Results are scored by the ability to identify heavy-atom positions from anomalous difference Fourier maps. The strategy was successfully applied in the determination of a membrane-protein structure, the copper-transporting P-type ATPase CopA, when other methods had failed to determine the heavy-atom substructure. MRPM is well suited to proteins undergoing large conformational changes where multiple search models should be considered, and it enables the identification of weak but correct molecular-replacement solutions with maximum contrast to prime experimental phasing efforts.



OPEN ACCESS

1. Introduction

To determine the structure of a macromolecular crystal, the phase problem must be solved. For isomorphous replacement and anomalous scattering methods (referred to as experimental phasing in this paper), phasing can be considered a two-step procedure in which the heavy-atom (HA) substructure is initially derived, after which the substructure is used to calculate phases for the entire macromolecular structure (Hendrickson, 1991; Dauter *et al.*, 2002). Knowing the substructure, reasonable experimental maps can often be generated from surprisingly weak data thanks to improvements in statistical phase probability calculations and density-modification procedures (Terwilliger, 2000, 2001; McCoy, 2002; Jenni *et al.*, 2006; Keller *et al.*, 2006; Maier *et al.*, 2006; McCoy & Read, 2010; Abrescia *et al.*, 2011; Li & Li, 2011; Liu *et al.*, 2011)

Typically, the heavy-atom substructure is found using Patterson-based methods, direct methods or frequently dual-space methods, which can combine Patterson-based seeding with direct methods and real-space steps (Hendrickson & Ogata, 1997; Weeks & Miller, 1999; Burla *et al.*, 2003; Grosse-Kunstleve & Adams, 2003; Sheldrick, 2008). Such heavy-atom site identification is nontrivial when only weak diffraction data

of poor quality are available and is often complicated by crystal and data pathologies such as radiation damage and severe anisotropy (Skubák & Pannu, 2013; Bunkóczi *et al.*, 2015).

Molecular replacement (MR) is an alternative method for obtaining phase estimates. However, if the experimental data are of low resolution and low quality, the end result will be highly biased by the model (Read, 1986; DeLaBarre & Brunger, 2006), obscuring rebuilding and refinement of the proper target structure.

Nonetheless, MR is still useful in such difficult cases. By using molecular replacement at low resolution, an initial starting model, despite very low sequence identity, can generate phases which allow the identification of HA peaks through anomalous difference Fourier maps (de La Fortelle & Bricogne, 1997; McCoy & Read, 2010). After positioning of the heavy atom(s), the model-biased MR phases can in principle be discarded and phase calculation and improvement can be conducted using traditional methods. For examples, see Pedersen *et al.* (2007) and Mourão *et al.* (2014).

Here, we present a systematic expansion of this approach that we developed for the structure determination of the copper-transporting P-type ATPase CopA (Gourdon, Liu *et al.*, 2011). The identification of heavy-atom sites in CopA HA-derivative data turned out to be highly challenging. While an extensive effort was put into the generation of improved derivative and native crystals, a strategy to systematically screen MR parameters was developed that we have dubbed molecular-replacement parameter matrix (MRPM) search, since traditional approaches failed to facilitate structure determination and refinement.

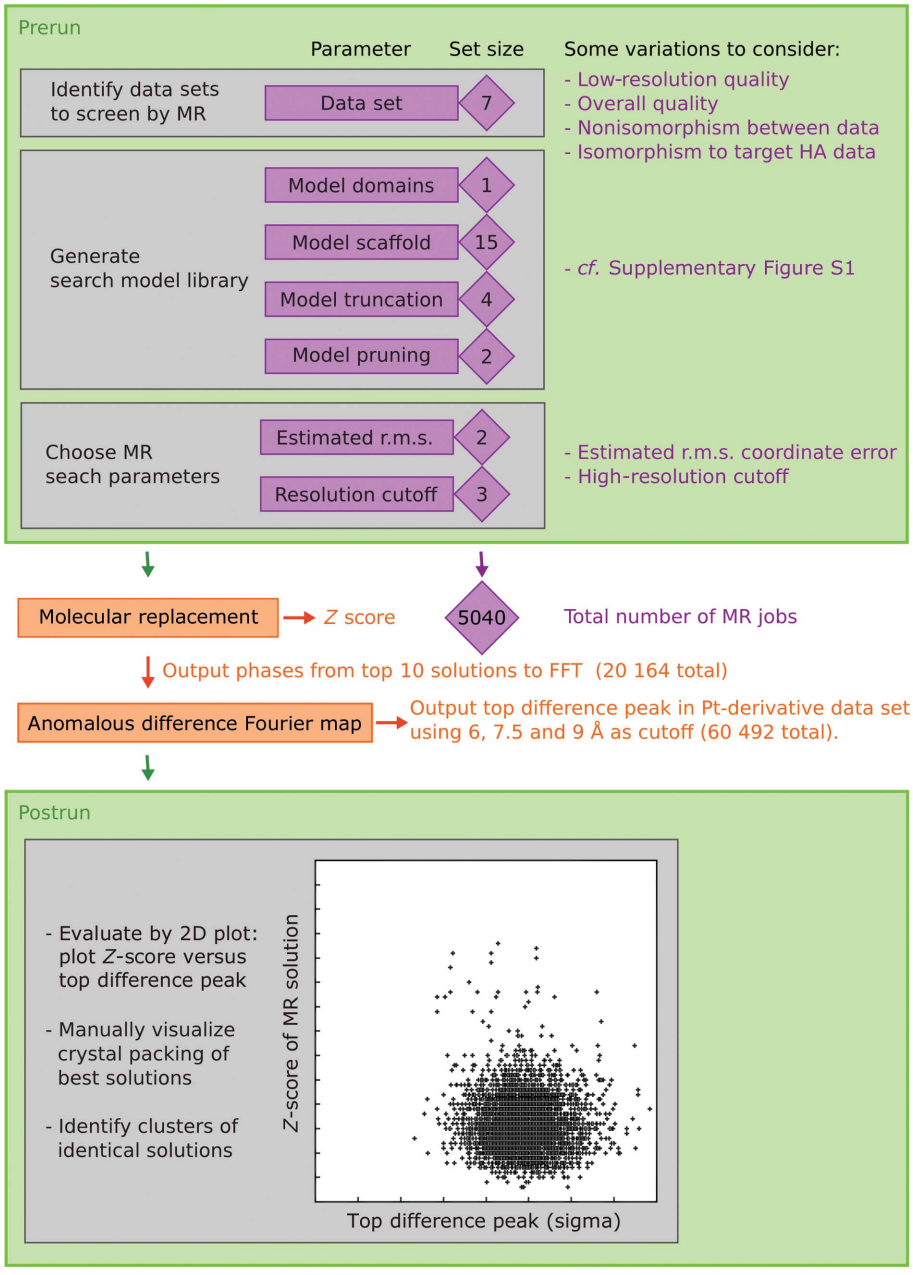


Figure 1
Overview of the MRPM search strategy. Prerun considerations (top green box) have to be made to identify parameters (dimensions) and sets of values to test for each parameter. The parameters and set size for each parameter shown here are specific for the CopA case. After each MR and FFT calculation, the result is plotted on a two-dimensional plot to identify clusters of MR solutions that both have a high *Z*-score and generate large difference peaks in the Pt-derivative data set.

2. Materials and methods

2.1. Sample description

CopA is a copper-exporting membrane protein that belongs to the well studied family of primary transporters known as P-type ATPases (Møller *et al.*, 1996; Axelsen & Palmgren, 1998; Palmgren & Nissen, 2011). This family has a transmembrane (M) domain with a common core of six transmembrane (TM) helices, and three soluble domains, known as the A, N and P domains (Morth *et al.*, 2011). Crystallization of a CopA family member from *Legionella pneumophila* (LpCopA) resulted in crystals that diffracted to about 3 Å resolution in the best case but suffered from severe non-isomorphism between most data sets (Supplementary Table S1; Gourdon, Andersen *et al.*, 2011; Gourdon, Liu *et al.*, 2011).

2.2. Method description

The identification of a correct MR solution is not trivial when the search model and/or experimental data are of poor quality. The use of various

high-resolution data cutoffs and estimated root-mean-square coordinate errors (r.m.s.) of the search model should be explored, and search-model completeness *versus* correctness should be ensured (Pedersen *et al.*, 2010; Oeffner *et al.*, 2013). If conformational flexibility of the target is possible, different conformational states should be tested as well.

Here, we include a number of model conformations and search parameters in a systematic expansion to explore a large MR parameter space. Since the end goal is to identify consistent HA peaks in a substructure determination, the numerous MR solutions are scored using this criterion and simultaneously the corresponding *Z*-score to help to distinguish correct solutions from noise.

2.3. Hardware and software

The computer used was a regular Linux desktop computer [4× Intel Xeon CPU W3540 (2.93 GHz), 24 GB RAM]. A total of 397 CPU hours were used for this analysis. In real time, the calculations took 4 d 3 h 20 min.

All scripts were made using the Bourne shell (sh). Example scripts sufficient to perform a similar analysis are provided as Supporting Information. The programs used were *Phaser* (McCoy *et al.*, 2007), *PEAKMAX* (Winn *et al.*, 2011), *SCALEIT* (Howell & Smith, 1992), *FFT* (Ten Eyck, 1973), *SUPERPOSE* (Krissinel & Henrick, 2004), *PyMOL* (http://www.pymol.org) and *gnuplot* (http://www.gnuplot.info).

3. Results and discussion

A schematic representation of the MRPM strategy is shown in Fig. 1. Manually analyzing the heavy-atom derivative data sets collected, a K_2PtCl_6-derivative data set was identified to be our superior HA data set, *i.e.* that with the most significant anomalous difference signal, in this case extending to 5.5 Å resolution (Supplementary Table S2). A strategy was therefore designed to evaluate whether MR phases could identify significant anomalous difference peaks in this Pt-derivative data set.

3.1. Generation of the search-model library

Several full-length P-type ATPase structures (predominantly of the rabbit sarcoplasmic reticulum Ca^{2+}-ATPase 1a) are available in the Protein Data Bank (PDB), representing a library of conformational states that are characteristic of this protein family. We regard a search model as composed of a number of domains placed according to different scaffold structures representing conformational states. To further increase sampling, the domains are subjected to different truncations of loop regions or whole domains and pruning of the side-chain atoms, leading to sublibraries of related search models.

For scaffolds, 33 P-type ATPase structures were downloaded and an r.m.s. deviation matrix of the C^{α} atoms was calculated (Supplementary Table S3). Redundant scaffold structures were identified, resulting in 15 unique scaffolds with greater than 1 Å r.m.s. deviation from each other (Supplementary Table S4).

Structures of isolated A, N and P domains with high sequence identity to our LpCopA target were identified by *BLAST*. For the M domain, the six core TM helices of each of the 15 scaffolds were used. These four domains together cover ∼71% of the CopA sequence (Supplementary Table S5). Missing parts of CopA included the heavy-metal binding domain and the two N-terminal TM helices; both are specific features of heavy-metal pumps and had unknown positions relative to the scaffolds.

The four domains were placed by superposition into the 15 scaffolds, resulting in 15 starting models representing the conformational variability observed in the database of P-type ATPase structures (Supplementary Fig. S1, steps 1 and 2; Supplementary Fig. S2).

To compensate for potential domain flexibility or domain-structure errors, we included three truncated versions for each starting model (A, N and M domain removed, respectively; Supplementary Fig. S1, step 3), and these four versions of each starting model were generated in two forms: either with all atoms included or pruned to polyalanines only (Supplementary Fig. S1, step 4). In total, the final library contained 120 different search models (Supplementary Tables S6–S9).

3.2. Setting up the MR parameter-matrix search

Six native data sets were selected, based on criteria such as good quality of the low-resolution data, highest obtained resolution and best scaling overall to the Pt-derivative data set (Supplementary Table S2). Assuming one monomer per asymmetric unit, the solvent content was estimated to be about 62%, which is typical of membrane-protein crystals.

Based on previous experience with MR and low-quality data (Pedersen *et al.*, 2010), we tested different values for the expected r.m.s. coordinate error (2 or 3 Å) and high-resolution limits of the data (4, 6 and 8 Å), while leaving other parameters constant.

The final parameter matrix systematically combined these six search-parameter setups with seven data sets and 120 search models, parsing a total of 5040 MR searches for analysis (Fig. 1). As the correct solution was expected to be weak, the ten best final solutions from each run were saved and evaluated. Postrun analysis shows that a total of 20 164 suggested MR solutions were output from the 5040 MR searches.

3.3. Evaluation

An anomalous difference Fourier map of the Pt-derivative data set was calculated for each of the 20 164 MR solutions. Peaks are expected to be weak in such maps and very sensitive to the resolution cutoff. To adress this, three cutoff values (6, 7.5 and 9 Å) were used. The highest anomalous difference peak for each of the 60 492 maps was identified and plotted (peak height in σ units) as a function of the *Z*-score of the input MR solution.

The majority of MR solutions had low *Z*-scores (<5.5) and did not give rise to significant difference peaks (<5σ),

indicating failed MR searches. However, a number of favourably scored MR solutions were apparent and through evaluation according to the various screened parameters a tantalizing pattern emerged (Fig. 2).

A broad selection of top-scoring solutions were manually analyzed and we found that 30 of these were virtually identical and all identified the same difference peak (highlighted in Fig. 2). All of these required the exclusion of high-resolution data, a scaffold with an 'outward-facing occluded' conformation and a polyalanine model excluding the M domain. Depending on the MR data set used, the parameters would either give a notable *Z*-score or a notable difference peak.

The phases from MR using these parameters allowed the determination of two initial positions of Pt atoms leading to experimental phases and allowing structure determination to proceed (Gourdon, Liu *et al.*, 2011).

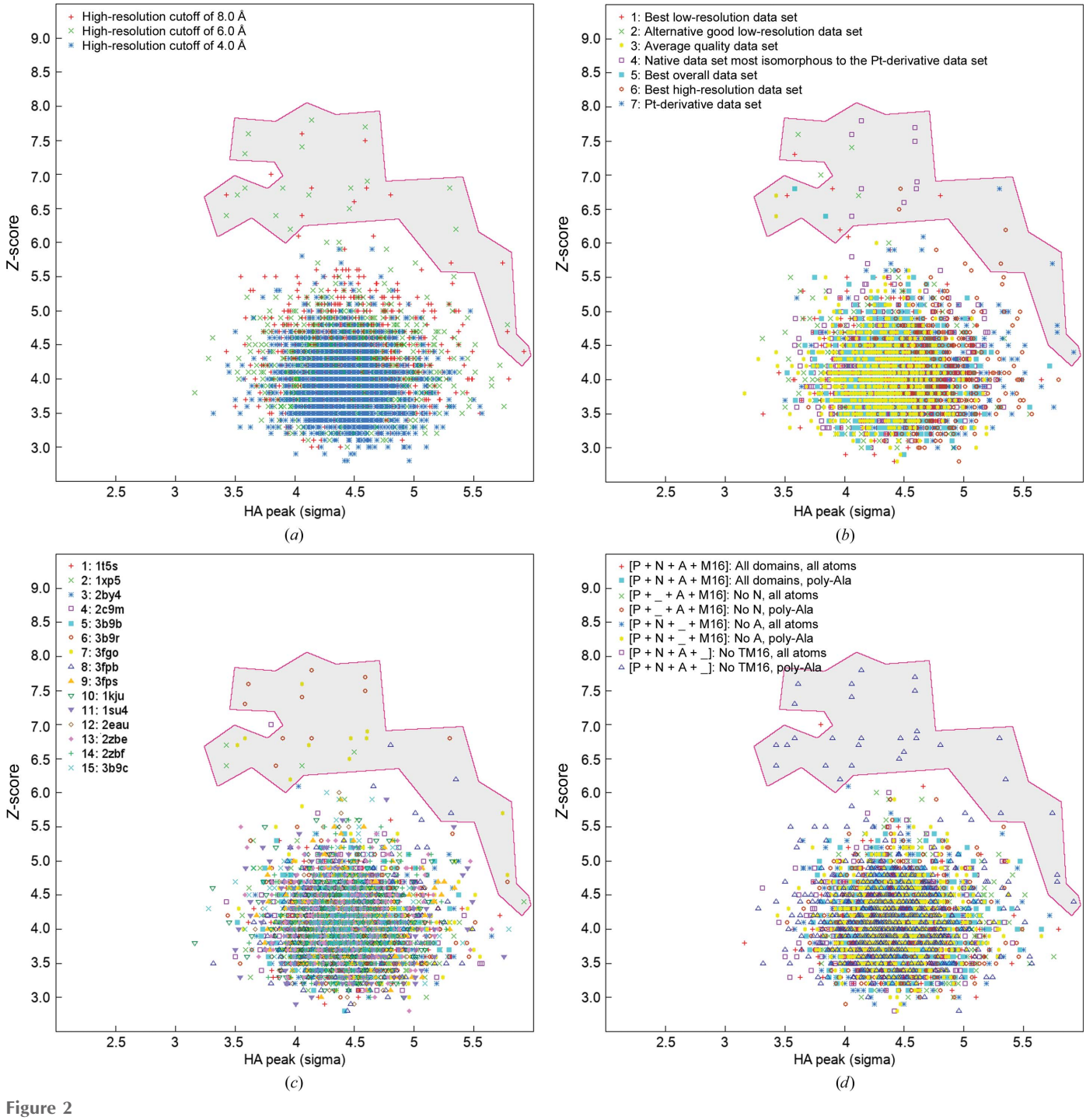


Figure 2
Two-dimensional plot of the result of the MR parameter search. All solutions are plotted as a function of *Z*-score and corresponding highest difference peak in the Pt-derivative data set. The grey area highlights the MR solutions that turned out to be identical and correct. (*a*) High-resolution cutoff. (*b*) Data set used. (*c*) Scaffold used. The PDB code is noted. (*d*) Truncation and pruning used.

The best MR solution as evaluated by *Z*-score alone (*Z*-score 7.8) was a correct solution, but the Pt peak calculated using the phases from this particular solution was insignificant (4.14σ), likely owing to non-isomorphism to the Pt-derivative data set. We must emphasize that even if by serendipity the best possible selection of parameters tested here had been used in a single MR run, the result would still not be sufficiently clear in its own right to indicate a correct solution. Only by comparing a large number of solutions did a consistent picture emerge, which lent confidence to the subsequent analysis. One solution, for example, had an MR *Z*-score of 7.0, and another produced an anomalous difference peak at 5.79σ, which both appeared to be promising indications of a successful solution but which both also turned out to be wrong (Fig. 2).

4. Concluding remarks

For CopA, the molecular-replacement parameter matrix search presented here was our workaround to initiate phasing for a structure determination that was plagued by weak diffraction properties and poor crystal-to-crystal isomorphism. We believe that the MRPM search strategy is of general interest for numerous projects with analogous challenges as well as in more standard applications. It can easily be extended to use more or different dimensions than those presented here. Employing an array of different domains (for example, domains solved from different organisms) is one example. Testing more data sets and using alternative methods of search-model pruning, as well as full mutagenesis to the target sequence or the creation of mixed models, are other obvious possibilities [using, for example, *CHAINSAW* (Stein, 2008) and *Sculptor* (Bunkóczi & Read, 2011)]. Furthermore, multiple derivative data sets could easily be employed to identify consistent sets of different HA peaks.

An ever-increasing number of programs target the phase problem in different ways, and our choice of programs is not necessarily the best one for any given case. Instead, we wish to emphasize the general value of systematic sampling for difficult cases, and this may also include different programs or approaches. For instance, one could try using log-likelihood-gradient completion in *Phaser* to find the heavy-atom sites (Read & McCoy, 2011) instead of calculating anomalous difference maps, or for relatively good-resolution data use *SHELXE* to reduce model bias and obtain an indication of whether MR solutions are correct without using any derivative data and experimental phasing (Thorn & Sheldrick, 2013). Keeping in mind the advent of improved protein-folding algorithms (Qian *et al.*, 2007; Rigden *et al.*, 2008; DiMaio *et al.*, 2011), generic search models (Strop *et al.*, 2007) and automated procedures (Keegan & Winn, 2007; Stokes-Rees & Sliz, 2010), as well as pipelines using large numbers of input search models (Bibby *et al.*, 2012; Sammito *et al.*, 2013), the importance of testing different conformational states is accentuated by the work presented here, and it emphasizes an aspect of modelling that is not currently addressed by *in silico* modelling.

Traditionally, crystallographic structure determination has proceeded through either experimental phasing or molecular replacement. MRPM is a hybrid approach in which heavy-atom derivative-based scoring is used to distinguish proper MR solutions that conversely determine the heavy-atom substructure to initiate experimental phasing. As another example of this, *phenix.mr_rosetta* takes a set of potential MR solutions and rebuilds each of these using *Rosetta* force fields to obtain the correct solution (Terwilliger *et al.*, 2012).

In general, systematic MR searches should be strongly preferred over single MR runs, using for instance an MRPM strategy as described here in conjunction with powerful combinatory approaches such as *MrBUMP* and *Wide Search Molecular Replacement* (Keegan & Winn, 2007; Stokes-Rees & Sliz, 2010). Even if derivative data sets are not available, a systematic search is more likely to help to identify a correct solution and distinguish it from false positives when only data of limited quality are available.

Acknowledgements

We are grateful to Gregers R. Andersen for discussions about molecular replacement strategies. We thank Robert M. Stroud and Janet Finer-Moore for reading and reviewing the final manuscript. BPP was supported by a postdoctoral fellowship from the Carlsberg Foundation and a fellowship from the Danish Cancer Society and later by a Sapere Aude Grant from the Danish Council for Independent Research and an AIAS fellowship; PG was supported by the Lundbeck Foundation and the Swedish Research Council and XL by the China Scholarship Council. This project has received funding from the European Research Council (grant agreement No. 637372 and grant agreement No. 250322).

References

Abrescia, N. G. A., Grimes, J. M., Oksanen, H. M., Bamford, J. K. H., Bamford, D. H. & Stuart, D. I. (2011). *Acta Cryst.* D**67**, 228–232.
Axelsen, K. B. & Palmgren, M. G. (1998). *J. Mol. Evol.* **46**, 84–101.
Bibby, J., Keegan, R. M., Mayans, O., Winn, M. D. & Rigden, D. J. (2012). *Acta Cryst.* D**68**, 1622–1631.
Bunkóczi, G., McCoy, A. J., Echols, N., Grosse-Kunstleve, R. W., Adams, P. D., Holton, J. M., Read, R. J. & Terwilliger, T. C. (2015). *Nature Methods*, **12**, 127–130.
Bunkóczi, G. & Read, R. J. (2011). *Acta Cryst.* D**67**, 303–312.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2003). *Acta Cryst.* D**59**, 662–669.
Dauter, Z., Dauter, M. & Dodson, E. J. (2002). *Acta Cryst.* D**58**, 494–506.
DeLaBarre, B. & Brunger, A. T. (2006). *Acta Cryst.* D**62**, 923–932.
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.
Gourdon, P., Andersen, J. L., Hein, K. L., Bublitz, M., Pedersen, B. P., Liu, X.-Y., Yatime, L., Nyblom, M., Nielsen, T. T., Olesen, C., Møller, J. V., Nissen, P. & Morth, J. P. (2011). *Cryst. Growth Des.* **11**, 2098–2106.
Gourdon, P., Liu, X.-Y., Skjørringe, T., Morth, J. P., Møller, L. B., Pedersen, B. P. & Nissen, P. (2011). *Nature (London)*, **475**, 59–64.
Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Acta Cryst.* D**59**, 1966–1973.

Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Hendrickson, W. A. & Ogata, C. M. (1997). *Methods Enzymol.* **276**, 494–523.
Howell, P. L. & Smith, G. D. (1992). *J. Appl. Cryst.* **25**, 81–86.
Jenni, S., Leibundgut, M., Maier, T. & Ban, N. (2006). *Science*, **311**, 1263–1267.
Keegan, R. M. & Winn, M. D. (2007). *Acta Cryst.* D**63**, 447–457.
Keller, S., Pojer, F., Heide, L. & Lawson, D. M. (2006). *Acta Cryst.* D**62**, 1564–1570.
Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* D**60**, 2256–2268.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Li, W. & Li, F. (2011). *Structure*, **19**, 155–161.
Liu, Q., Zhang, Z. & Hendrickson, W. A. (2011). *Acta Cryst.* D**67**, 45–59.
Maier, T., Jenni, S. & Ban, N. (2006). *Science*, **311**, 1258–1262.
McCoy, A. J. (2002). *Curr. Opin. Struct. Biol.* **12**, 670–673.
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. & Read, R. J. (2010). *Acta Cryst.* D**66**, 458–469.
Møller, J. V., Juul, B. & le Maire, M. (1996). *Biochim. Biophys. Acta*, **1286**, 1–51.
Morth, J. P., Pedersen, B. P., Buch-Pedersen, M. J., Andersen, J. P., Vilsen, B., Palmgren, M. G. & Nissen, P. (2011). *Nature Rev. Mol. Cell Biol.* **12**, 60–70.
Mourão, A., Nager, A. R., Nachury, M. V. & Lorentzen, E. (2014). *Nature Struct. Mol. Biol.* **21**, 1035–1041.
Oeffner, R. D., Bunkóczi, G., McCoy, A. J. & Read, R. J. (2013). *Acta Cryst.* D**69**, 2209–2215.
Palmgren, M. G. & Nissen, P. (2011). *Annu. Rev. Biophys.* **40**, 243–266.
Pedersen, B. P., Buch-Pedersen, M. J., Morth, J. P., Palmgren, M. G. & Nissen, P. (2007). *Nature (London)*, **450**, 1111–1114.
Pedersen, B. P., Morth, J. P. & Nissen, P. (2010). *Acta Cryst.* D**66**, 309–313.
Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J. & Baker, D. (2007). *Nature (London)*, **450**, 259–264.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Read, R. J. & McCoy, A. J. (2011). *Acta Cryst.* D**67**, 338–344.
Rigden, D. J., Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* D**64**, 1288–1291.
Sammito, M., Millán, C., Rodríguez, D. D., de Ilarduya, I. M., Meindl, K., De Marino, I., Petrillo, G., Buey, R. M., de Pereda, J. M., Zeth, K., Sheldrick, G. M. & Usón, I. (2013). *Nature Methods*, **10**, 1099–1101.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.
Skubák, P. & Pannu, N. S. (2013). *Nature Commun.* **4**, 2777.
Stein, N. (2008). *J. Appl. Cryst.* **41**, 641–643.
Stokes-Rees, I. & Sliz, P. (2010). *Proc. Natl Acad. Sci. USA*, **107**, 21476–21481.
Strop, P., Brzustowicz, M. R. & Brunger, A. T. (2007). *Acta Cryst.* D**63**, 188–196.
Ten Eyck, L. F. (1973). *Acta Cryst.* A**29**, 183–191.
Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.
Terwilliger, T. C. (2001). *Acta Cryst.* D**57**, 1763–1775.
Terwilliger, T. C., DiMaio, F., Read, R. J., Baker, D., Bunkóczi, G., Adams, P. D., Grosse-Kunstleve, R. W., Afonine, P. V. & Echols, N. (2012). *J. Struct. Funct. Genomics*, **13**, 81–90.
Thorn, A. & Sheldrick, G. M. (2013). *Acta Cryst.* D**69**, 2251–2256.
Weeks, C. M. & Miller, R. (1999). *Acta Cryst.* D**55**, 492–500.
Winn, M. D. (2011). *Acta Cryst.* D**67**, 235–242.