High-throughput protein structure determination using grid computing

Jason W. Schmidberger[1], Blair Bethwaite[4], Colin Enticott[4], Mark A. Bate[1], Steve G. Androulakis[1], Noel Faux[5], Cyril F. Reboul[1,2], Jennifer M. N. Phan[1], James C. Whisstock[1,2], Wojtek J. Goscinski[3] , Slavisa Garic[4], David Abramson[3,4], and Ashley M. Buckle[1,2]

[1]Department of Biochemistry and Molecular Biology,
[2]ARC Centre of Excellence in Structural and Functional Microbial Genomics,
[3]Monash eResearch Centre,
[4]Clayton School of Information Technology,
Monash University, Victoria 3800, Australia.
[1]NICTA Victoria Research Laboratory at The University of Melbourne, Australia.

Abstract

Determining the X-ray crystallographic structures of proteins using the technique of molecular replacement (MR) can be a time and labor-intensive trial-and-error process, involving evaluating tens to hundreds of possible solutions to this complex 3D jigsaw puzzle. For challenging cases indicators of success often do not appear until the later stages of structure refinement, meaning that weeks or even months could be wasted evaluating MR solutions that resist refinement and do not lead to a final structure. In order to improve the chances of success as well as decrease this timeframe, we have developed a novel grid computing approach that performs many MR calculations in parallel, speeding up the process of structure determination from weeks to hours. This high-throughput approach also allows parameter sweeps to be performed in parallel, improving the chances of MR success.

1. Introduction

Proteins perform the functions necessary for life in all organisms. Protein function is to a large extent dictated by the 3-dimensional structure, and thus knowledge of the atomic structure of a protein is a prerequisite to understanding its function. The understanding of protein structure now has a firm role in the molecular basis of all diseases, and as such is a vital underpinning for the future promise of de novo drug design. X-ray crystallography is the most common technique for the structure elucidation of proteins. Briefly, this method involves first the production of large amounts of (usually recombinant) pure protein, followed by crystallization and X-ray diffraction analysis. The atomic structure is then calculated from the diffraction pattern using one of several methods. Over the last 5 years adoption of automation technologies has eased the bottlenecks at the cloning, protein production and crystallization stages. Availability of synchrotron radiation has increased the rate at which high-quality diffraction data can be collected. Although the development of computational methods of structure elucidation has also undergone significant improvement, the high-throughput nature of the pipeline places an increasing emphasis on the computational resources available for structure calculation.

Structure determination can take days to months, and is frequently complicated by the heterogeneity of hardware, software and data

978-1-4244-3750-4/09/$25.00 ©2009 IEEE

formats encountered. Despite efforts to improve the user-friendliness of software the learning curve for novice structural biologists can be steep, particularly for researchers with a biological sciences background. Furthermore, crystallographers have been slow to harness the power of high-performance distributed computing. In this paper we describe the development of a novel approach to performing common crystallographic calculations in a high throughput fashion, using grid computing.

2. Background

The most common method of crystallographic protein structure determination is molecular replacement (MR). This technique involves using the structure of a protein that shares significant sequence similarity with the protein of unknown structure as a starting point in the structure determination (otherwise known as solving the *phase problem*). The process generally involves four steps: (1) Using sequence-based searching methods such as PSI-BLAST [1] to identify suitable structures that can be used for MR; (2) modification of structures to yield *search models*; (3) Finding the orientation and position of the search model in the unit cell of the target crystal; (4) Refinement of the model.

Molecular replacement has been used to determine the structure of approximately half of the 50,000 structures deposited in the Protein Data Bank (PDB; 67% of 2006 releases were solved by MR [2]). It is anticipated that the proportion determined by MR will grow for three reasons: First, the probability that the unknown target structure belongs to readily identifiable fold is steadily increasing, due to the rapid growth of the PDB. Second, the emergence of more sophisticated sequence searching algorithms, such as profile-profile matching [3], improve the probability of finding a suitable search model, even in cases of very low similarity (<20% identity). Third, MR algorithms consistently improve.

3. Parallel Molecular Replacement

Where the sequence similarity between the unknown target and the search model is high (sequence identity >40%) the success rate of MR is very good, even without optimization of the search model. However, in cases where sequence similarity is low (identity <30%) MR, and subsequent structure refinement becomes non-trivial, and emphasis must be placed on the optimization of the search model. A key breakthrough in successfully applying the MR approach to situations where sequence identity is low was the development of the PHASER maximum likelihood approach [4].

Even in cases where a MR solution with low overall sequence identity can be obtained these solutions are commonly challenging to refine (the so called “model bias” trap). This situation occurs where errors in regions of the starting model cannot be adequately identified and corrected due to model bias. There are several criteria that affect the outcome of the MR calculation; 1) structural similarity between search model and target structure (measured by root mean square deviation (RMSD)); 2) percentage of residues missing from the search model (*coverage*); 3) the amount of conserved side chains that are expected to remain structurally conserved (for example in the protein interior). These factors, and thus the outcome of the MR calculation, can be influenced by improvement of the search model, using several methods. The simplest approach is to remove regions of the structure that are predicted to be different in the search model and target, typically loops. However, this process is a subjective one and relies on sequence alignments, which are often incorrect, particularly at low sequence identity. Thus it is often unclear which loops should be removed and how much of the loop should be removed, and each model must be tested. We have developed a robust solution to this method, called *sieving*, which produces search models with structurally divergent regions removed in an objective fashion (Schmidberger *et al.,* unpublished data). The ideal starting model (*e.g.* one with least model bias) is difficult to obtain *a priori,* however it is possible to test multiple sieved models and asses the refinement process using statistically robust validation, providing a generally applicable method for model bias reduction. A further method of search model improvement leverages protein flexibility. Proteins are typically not rigid but undergo conformational changes, resulting in a population of many distinct conformational states. This can be modelled using the technique of Normal Mode Analysis (NMA [5]), which produces many alternative molecular conformations that can all be tested as search models in the MR calculation. In some cases the true symmetry of the crystal is unknown and several alternatives must be tested in the MR calculation. Finally, the estimated RMSD between the search model and unknown structure can affect the outcome of the MR calculation, leading in the worst case to probable solutions being missed.

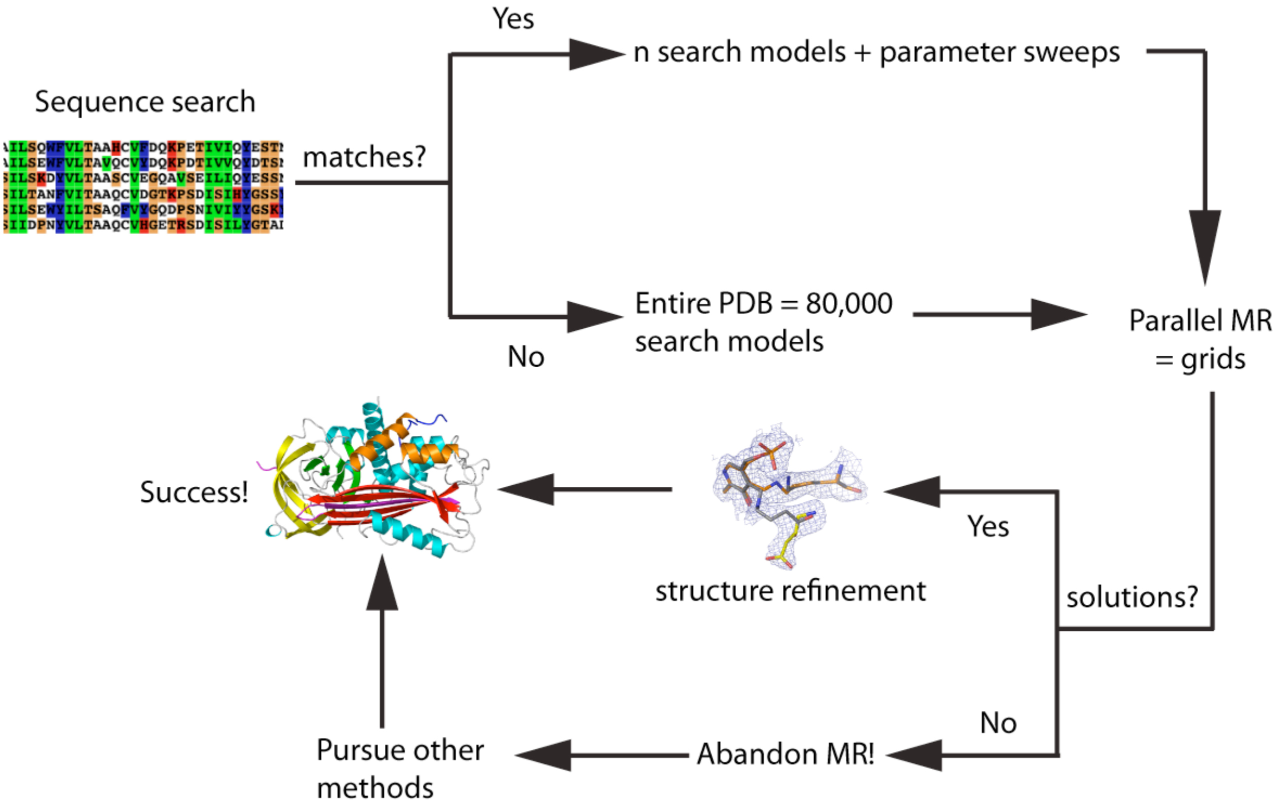

Figure 1 – Parallel MR Approach

Therefore, for challenging cases the combination of multiple search models produced by sieving and NMA, symmetry and RMSD values makes MR potentially time and labor intensive, and puts an emphasis on the availability and power of computational resources.

Given the complexities of the MR technique outlined above we have developed a novel grid computing approach that is able to perform independent MR calculations using hundreds to thousands of candidate search models in parallel, whilst also performing near-exhaustive parameter-sweeps in order to increase chances of success. Furthermore, we have extended this approach so that the entire protein fold universe (currently ~80,000 folds) can be tested. This massively parallel approach may be useful in cases where the similarity between a protein of unknown structure and a "known fold" cannot be detected by sequence matching methods, despite them sharing the same fold. For such proteins, an MR-based approach may be achievable, but up until recently, the computational resources required for such an approach would be prohibitive. However, the exponential growth of computing power and recent advances in harnessing this power in a massively parallel fashion, using grid computing, means this approach is now feasible. Figure 1 summarizes the rationale behind this approach.

4. Implementation

We have implemented hierarchical grid-based approach that leverages a range of distributed computational resources. Generally, the approach performs multiple PHASER-associated MR calculations across a grid of networked computers, permitting high-throughput MR. This approach is summarized in Figure 2.

There is significant heterogeneity in the resources available to us accessed by a range of different middleware solutions. For example, a collection of Apple Macintosh's in the department of biochemistry at Monash University use Apple's proprietary Xgrid technology. We have exploited the rapid and easy implementation of Xgrid for developing the parallel MR methodology and for proof-of principle testing. Once convinced of its merit and for larger scale computations, we used a combination of a large Condor Pool, that aggregates many of the student laboratory machines, and a range of clusters distributed globally. These latter resources were accessed via Globus and the Nimrod/G middleware. These are discussed in more detail below.

4.1 Xgrid

We have built a web based application written in Java/JSP and Ruby, and taking advantage of Apple Xgrid technology (see: http://www.apple.com/server/macosx/technology/xgrid.html), which we call *MR Grid*. Designed to interface with a user defined Xgrid resource the package manages the distribution of multiple MR runs to the available nodes on the grid and reports all returned results. Utilizing the maximum likelihood based molecular replacement program PHASER [4], MR Grid enables the user to retrieve and manage the results of hundreds of MR calculations via a single web interface, as well as broadening the range of strategies that can be attempted, increasing the likelihood of success.

MR Grid is distributed as a self-contained software package, and downloaded and executed across a local grid resource. Once set up MR Grid is accessed through a web portal. Apple Xgrid software is preinstalled on Apple operating systems OS X 10.4 and 10.5, allowing machines to be

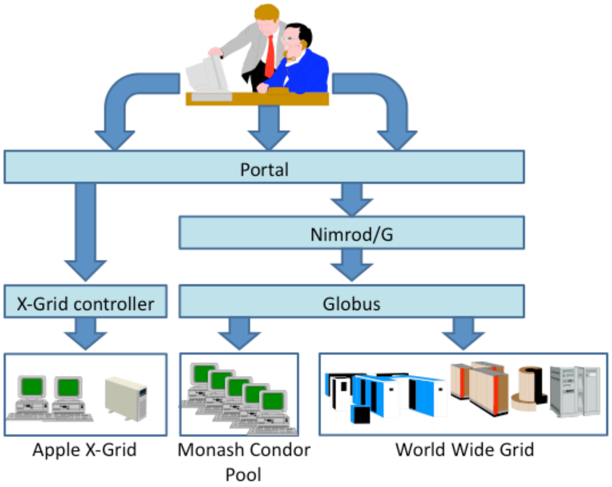

Figure 2 – Parallel MR grid architecture

configured as Xgrid clients by simply activating a setting in system preferences. By default MR Grid processes on the client are given low priority, such that the client remains fully responsive. The remaining requirement is a networked machine acting as the Xgrid controller. This can be running either the server or client version of Mac OS 10.4/10.5.

MR Grid takes as input the X-ray diffraction data and a compressed file of search models. MR Grid then parses the input and distributes jobs to available nodes on an available grid resource. Each job runs to completion independent of all other jobs, and a URL where the results of the submission can be accessed is returned to the user. Job distribution and queuing is performed entirely by the Xgrid controller and requires no programming. MR Grid source can be found at http://code.google.com/p/mrgrid/.

4.2. Condor and World Wide Grid

MR Grid is well suited to performing parallel MR on a modest laboratory-based network of Apple computers, and performs useful validation of our approach. In order to attack more challenging problems, we have leveraged two different classes of resource that were available to us, namely a ~1000 CPU Condor pool built from otherwise idle desktop machines in Monash teaching laboratories; and a World Wide Grid of machines leveraging computers (mostly clusters or Condor pools) in an Australian University Enterprise Grid, the Pacific Rim and Grid Middleware Assembly (PRAGMA) testbed, and the US based Open Science Grid (OSG). This provides thousands of processors and has allowed us to perform high throughput MR based calculations.

We have used Nimrod/G [6] to allow the distribution and execution of a large number of jobs to both resources. Nimrod/G interfaces with the Globus middleware that provides a common access layer for all resources. Thus, we are able to access the Monash Condor Pool in exactly the same way as the Grid resources that are available through our Global collaborations. Combining all of these machines, has allowed us to perform an extremely large MR experiment consisting of some 80,000 PHASER executions.

Table 1. List of test case proteins, extracted from the Protein DataBank (PDB). Details about respective datasets are also listed

|PDB id	Protein Name/Type	Space Group	Resolution Limit (Å)	Molecular Mass (Da)	Ave. %ID	# Search Models	# SGs (or RMSDs) tested	# Jobs (RMSD)	Grid Run time (hh:mm:ss)	Linear Run time (hh:mm:ss)
2GPZ	transthyretin-like protein	$P6$	2.5	12700	66.4	4	6 (5)	24 (20)	0:13:24 (0:15:05)	1:16:53 (0:57:33)
2NO4	Haloacid Dehalogenase	$P3_121$	1.9	24000	37.7	5	3	15	7:16:48	22:18:36
2CWQ	Hypothetical protein TTHA0727	$P3_121$	1.9	12581	<30	8	3	24	5:09:03	23:54:00
2ENX	Mn-dependant inorganic pyrophosphatase	$H32$	2.8	33597	57.5	7	1	7	0:04:07	0:17:39
2RH5	Adenylate kinase	$C222_1$	2.48	23231	43.2	8	2	16	0:20:36	0:40:38
1S3G	Adenylate kinase	$P3_121$	2.25	23888	41.3	8	3 (4)	24 (32)	0:25:12 (0:33:34)	1:38:19 (5:08:41)
2JCB	5-Formyl-tetrahydrofolate cycloligase	$P1$	1.6	23385	31	4	1 (4)	4 (16)	1:06:58 (1:21:00)	7:18:00 (6:55:35)
2H74	Thioredoxin	$P6_1$	2.4	11807	49.5	9	6	54	0:22:12	4:32:28
1FB0	Thioredoxin	$P3_121$	2.26	11782	45.1	9	3 (5)	27 (45)	0:29:40 (0:12:28)	3:24:34 (1:44:54)
2MM1	Myoglobin	$P3_221$	2.8	17184	53.9	12	3	36	0:10:42	1:20:41

5. Experiments and Results

In all cases MR calculations using the program PHASER produce possible solutions having a maximum likelihood Z-score. Z-scores greater than 7 were chosen as probable solutions worthy of structure refinement. Specific methodologies and test data were chosen based upon the class of grid resource available, and are described in detail below.

5.1 Xgrid

A set of 10 proteins were used as test cases, representing 8 different SCOP [7] families (Table 1), and allowing for the parallel execution of 4 to 54 jobs at any one time. PDB entries were selected on the basis of having 3 or more homologous structures in the PDB, with datasets from a range of crystal symmetries. MR search models were generally chosen on the basis of a >30% sequence identity across >75% of the monomer of interest (*i.e.* no partial matches).

Experimental data taken from PDB for the 10 proteins listed in Table 1 were each used in test case experiments in order to demonstrate the utility of the system under typical situations. For each protein example, data were screened against each homologue search model (including self), searching all possible space group symmetries. In this way, the number of jobs submitted to our local grid varied between 4 and 54, and the corresponding speed up factors showed a clear linear relationship. Featuring an average speed up value of 5.7 across all the tests, it is clear that our approach has the capacity to significantly reduce the time taken to achieve a MR result when screening numerous parameters, thus demonstrating the validity of the approach.

5.2 Condor

Having demonstrated the benefits of a parallel approach to MR, we sought to apply it to a challenging, real-world case. We chose a specific X-ray dataset in our laboratory that had proved recalcitrant to several methods of structure determination, including MR (unpublished). The target protein was of molecular weight of 44500 Da, with one molecule in the crystal asymmetric unit (~53% solvent). Using sequence searching we found two potential search models in the PDB, having 24 and 16% sequence identity, respectively. X-ray diffraction data extended to 1.6Å resolution. For such a challenging case we chose to perform extensive parameter sweeps coupled with testing alternative models from a Normal Mode Analysis calculation (using the *El Nemo* server: http://www.igs.cnrs-mrs.fr/elnemo/) as well as a sieving approach (Schmidberger *et al,* unpublished). NMA analysis generated 11 models for each of the 5 lowest modes, giving 55 models in total. Each mode was then sieved separately, giving 495 models in total [11 x 9 (sieve levels) x 5 (NMA modes)]. When combined with RMSD parameter sweeps (5 alternatives) this produced a total of 2475 MR runs.

Calculations executed on up to 469 Condor nodes at once, and completed within 15 hours. After sorting the results according to the maximum likelihood Z-score in PHASER, we obtained a unique solution having a Z score of 7.5 (previous manual MR calculations failed to produce a Z-score greater than 5.5). After rigid-body refinement using the program REFMAC [8] the model was subjected to automatic model building and structure refinement using the program ARP/wARP [9]. This resulted in a near-complete structure (329 of 390 residues (85%) built), having crystallographic R(work) and R(free) values of 0.20 and 0.22, respectively (R-values are the typical measure of model correctness; typically correct, fully refined models have R(free) < 0.30).

5.3 World Wide Grid

In order to test the validity of the ambitious approach of using the entire set of known protein structures in a MR calculation, we chose two test cases of contrasting complexity. The first case, *Thioredoxin* represents a typical example of a small-sized protein: (Molecular weight 11.6 kDa; PDB ID 2E0Q, one molecule per crystal asymmetric unit, data resolution 1.5 Å). We generated approx 70,988 search models from the PDB using the domain classification according to the SCOP database [7]. The side chains of all residues were truncated to alanine, and all loops were removed prior to the MR calculation. We performed 70988 independent MR calculations using the Thioredoxin test dataset across a total of 300 nodes worldwide. The entire run took 94 hours to complete, producing high Z-scores, as expected, for search models having clear structural homology with thioredoxin.

For the second test we selected a more challenging example of a larger protein; *Thiamin Phosphate Synthase* (Molecular weight 25 kDa; PDB ID 2TPS, two molecules per crystal asymmetric unit, data resolution 1.25 Å). This MR calculation produced 70,988 independent MR runs, searching for one molecule in the crystal asymmetric unit, and a total of 1050 nodes worldwide, and took 52 hours to complete. Nodes utilized in both runs included resources at VPAC and Monash University in Victoria, Australia and PRAGMA grid resources in Japan, Switzerland, Thailand and the United States. In total this experiment utilized approximately half a million CPU hours.

6. Conclusions and Future Plans

Our MR calculations over small laboratory-based networks of Apple computers show clearly how tens to hundreds of search models can be performed in parallel, along with parameter sweeps, increasing the chances of success in a relatively short timeframe. Testing this approach using a challenging, unsolved X-ray dataset required a larger grid, and we were able to show that for our real-world case, parallel MR calculations could be performed on a medium-sized university campus grid in less than 15 hours (e.g., overnight). Importantly, this experiment allowed us to perform near-exhaustive parameter sweeps, and resulted in a successful structure determination of a protein structure that had previously resisted structure determination by conventional (e.g., serial) MR calculations. This result shows clearly the promise of performing MR in a parallel fashion. By extending this concept to cases where no suitable search models could be obtained, we demonstrated that using worldwide grid resources available to academic scientists, MR could be performed using the entire known protein structure universe as independent search models, in a timeframe of two days in the case of a medium sized protein.

We expect that the current trend towards multi-core architecture will strengthen our parallel MR approach. It is important to note that apart from the clear advantage of rapid structure determination, even negative results (e.g., failure to find MR solutions) will prove useful: an unsuccessful exhaustively-parallel MR calculation allows the protein crystallographer to make an objective decision on alternative methods to be pursued, at the early stages of the project, offering potential labor and cost savings.

7. Acknowledgements

We thank Ruby Law, Michelle Dunstone, and Randy Read for helpful discussions. We also thank our colleagues in PRAGMA, the Enterprise Grid, The Swiss National Grid and the Open Science Grid for their assistance. We acknowledge the resources provided by the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science; the resources provided by the Pacific Rim Applications and Grid Middleware Assembly and its member organizations, PRAGMA member institutes are funded by funding agencies of their respective governments. PRAGMA is supported by NSF Grant No. INT-0216895, INT-0314015 and OCI-0627026. We acknowledge support from the DEST (DIISR) funded DART and ARCHER projects, support from the Australian Research Council (ARC), the National Health and Medical Research Council (NHMRC), Victorian Partnership for Advanced Computing, the Victorian Bioinformatics Consortium, the state government of Victoria (Australia), the Monash library and the Monash e-Research Centre. A. M. B is an NHMRC Senior Research Fellow. J. C. W. is an ARC Federation Fellow, a Monash University Logan Fellow and an honorary NHMRC Principal Research Fellow. D. A. is an ARC Professorial Research Fellow.

8. References

[1] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25 (1997) 3389-402.

[2] F. Long, A.A. Vagin, P. Young, and G.N. Murshudov, BALBES: a molecular-replacement pipeline. Acta Crystallogr D Biol Crystallogr 64 (2008) 125-32.

[3] L. Jaroszewski, L. Rychlewski, Z. Li, W. Li, and A. Godzik, FFAS03: a server for profile--profile sequence alignments. Nucleic Acids Res 33 (2005) W284-8.

[4] A.J. McCoy, R.W. Grosse-Kunstleve, P.D. Adams, M.D. Winn, L.C. Storoni, and R.J. Read, Phaser crystallographic software. Journal of Applied Crystallography 40 (2007) 658-674.

[5] K. Suhre, and Y.H. Sanejouand, ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res 32 (2004) W610-4.

[6] D. Abramson, Giddy, J. and Kotler, L., High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid?, International Parallel and Distributed Processing Symposium (IPDPS), Cancun, Mexico, 2000, pp. 520- 528.

[7] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247 (1995) 536-40.

[8] G.N. Murshudov, A.A. Vagin, and E.J. Dodson, Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53 (1997) 240-55.

[9] S.X. Cohen, R.J. Morris, F.J. Fernandez, M. Ben Jelloul, M. Kakaris, V. Parthasarathy, V.S. Lamzin, G.J. Kleywegt, and A. Perrakis, Towards complete validated models in the next generation of ARP/wARP. Acta Crystallogr D Biol Crystallogr 60 (2004) 2222-9.