pxGrid - High-Throughput Protein Structure Determination using Computational Grids

Noel Faux[1,2], Mark Bate[1], Colin Enticott[3], Khalid Mahmood[1,2], David Abramson[3], Ashley M. Buckle[1]

[1]The Department of Biochemistry and Molecular Biology
[2]The ARC Centre of Excellence in Structural and Functional Microbial Genomics
Faculty of Medicine, [3]CSIT, Monash University, Clayton, Victoria 3800, Australia

Email: Ashley.Buckle@med.monash.edu.au

Abstract

Structural biology research places significant demands upon high-performance computing. The elucidation of protein structures at atomic resolution is computationally demanding and requires user-friendly interfaces to high-performance computing resources. Fortunately, critical calculations are embarrassingly parallel and thus ideally suited to distributed computing. Here we discuss how we are using Grid computing to determine the three dimensional structures of proteins in a massively parallel fashion, in a timeframe of hours to days - orders of magnitude faster than is currently possible.

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 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.

2. Protein Crystal Structure Determination by Molecular Replacement (MR)

X-ray crystallography is the most powerful technique for determining the 3-dimensional structures of proteins. The most common method in structure determination is Molecular Replacement (MR). This 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 three steps: (1) Using sequence-searching methods such as PSI-BLAST [1] to identify suitable structures that can be used for molecular replacement; (2) modification of probe structures (e.g.,

removal of flexible loop regions and non-identical side chains), 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 using iterative model-building and maximum likelihood atomic refinement. Although there are other methods of structure determination, molecular replacement is predicted to become an increasingly common technique, for two reasons. Firstly, the probability that the unknown target structure belongs to a known fold is steadily increasing, due to the exponential growth of the Protein Database (PDB) [2]. Secondly, 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).

In this paper we describe work that addresses two key problems in computational protein crystallography. Firstly, we describe the development and deployment of a computational Grid using Apple XGrid technology, designed for medium scale molecular replacement calculations. Secondly, we describe how we are using large scale distributed computing to perform intelligently guided brute force calculations to identify candidate models for structure determination in the event that no obvious search model (based on sequence similarity) is available.

2.1. Highly Parallel Molecular Replacement Using XGrid

In a typical MR calculation, several structural homologues of the target protein can be identified using sequence searching methods. These search models are then tested against the experimental data, in a serial fashion. The true symmetry of the data is often ambiguous until the latter stages of the MR calculation, so an MR calculation must be repeated in every possible symmetry system (space group). For example, if the diffraction data has orthorhombic symmetry, and we wish to trial 10 different search models, there are 8 x 10 combinations of space group and search model to test. This is currently achieved in a manual, serial approach. We are therefore developing methods of submitting each of the combinations to a node on a computational grid, in order to achieve linear speedup times. We have developed a web front-end to the MR program PHASER [4] that allows the user to upload diffraction data, and an unlimited amount of search

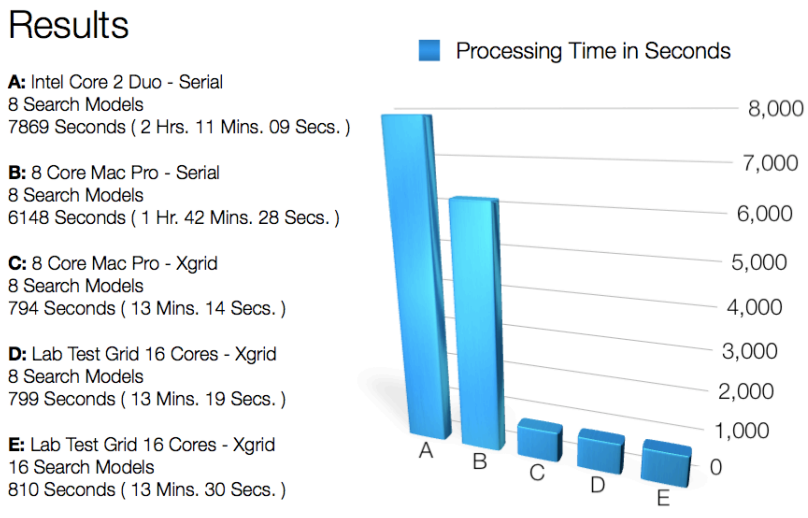

Figure 1. Execution times for serial and parallel, XGrid enabled PHASER MR calculations.

models. With minimal configuration the application then submits the necessary jobs to available Apple Mac OS X computers on the local network, using Apple XGrid technology to manage the batch queue system. We currently have >50 departmental computing nodes available, which are all available for use as nodes "out of the box" due to the standard incorporation of XGrid. Preliminary testing is producing near linear speedup factors, allowing us to test 10-100 search models in under an hour, compared to days or weeks in traditional serial computations (Figure 1).

2.2. Intelligently Guided Brute Force Molecular Replacement Using Large Grids

A key problem in bioinformatics is that structural similarity can be retained long after detectable sequence similarity is lost [5]. Thus it is common for similarity between a protein of unknown structure and a "known fold" to become apparent only after structure determination. For such proteins, an MR-based approach may have been achievable, however, the inability to detect the fold or family by sequence matching methods restricts its application. One approach in this scenario is to attempt brute force molecular replacement experiments using every single structure in the PDB (>3000 families). 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. To date, the structures of approximately 1000 different 3D folds have been described [6], from ~3000 families. Further, structural genomics programs have launched targeted attempts in order to provide the biological community with representatives of all folds [7; 8], estimated at ~1700 [8]. We have developed a "brute force" molecular replacement approach using all known folds, which does not rely on sequence similarity. Using the SCOP database [9], we have developed a library consisting of ~3000 MR search models derived from the representative highest resolution structure of each SCOP family. In initial experiments, we have developed a resource where each family representative is used as a search model in a PHASER MR calculation. In order to perform >3000 PHASER calculations in a timeframe of days rather than years we have developed a highly parallel approach using computational facilities at the Victorian Partnership for Advanced Computing (VPAC; *Brecca* – 97 dual

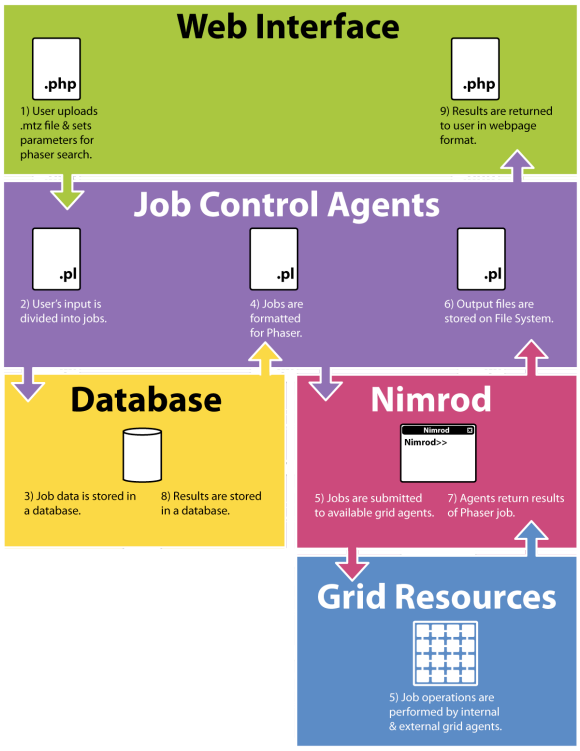

Figure 2. Architecture of Brute-Force, Grid Molecular Replacement

Xeon 2.8 GHz CPUs, 160 GB (2 GB per node) total memory; *Edda* – 185 Power5 CPUs, 552 GB (8-16 GB per node) total memory) and Monash University ITS Sun Grid (54 dual 2.3 GHz CPUs, 208.7 GB (3.8 GB per node) total memory). The PostgreSQL database (www.postgresql.org) system is used to store and manage the MR jobs and results. PERL scripts are used to farm out MR jobs to free CPU's, launch the MR programs and collect the results. The web front end is written using PHP software (www.php.net) and served using Apache server software (www.apache.org). This is represented schematically in Figure 2.

In order to expand this approach to all protein domains in the PDB (~80, 000) we require an order of magnitude increase in computing nodes. This is being achieved using the software tool *Nimrod/G [10],* which distributes individual jobs over the Pacific Rim Application and Grid Middleware Assembly (PRAGMA) testbed. As such, the availability of ~1000 nodes makes the scale of this task practical in a timeframe of days and at most, weeks. Specifically, each individual PHASER job consists of the csh script (describing the PHASER job), reflection file, search PDB, and any PDB file that will be fixed during the run. These files are copied over to the resource and the csh script is then run / submitted on the allocated resource. Upon completion of the job, the results files are copied back to the submission machine and the initial copied files are removed.

3. Conclusions

In summary, we are developing major new tools to solve the three-dimensional structures of proteins in a significantly shorter timeframe than is currently possible. The ability to perform MR calculations using an exhaustive set of search models will offer a timesaving of weeks to months in a typical successful structure determination. Challenging structure determinations by MR currently can take more than 6 months, therefore it is extremely useful to know as quickly as possible when the MR approach might fail, and thus when to pursue alternative methods.

4. Acknowledgments

We thank the NHMRC, ARC, Victorian Partnership for Advanced Computing, and the Victorian Bioinformatics Consortium for funding and support. AMB is an NHMRC Senior Research Fellow.

5. 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] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, The Protein Data Bank. Nucleic Acids Res 28 (2000) 235-42.
[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, L.C. Storoni, and R.J. Read, Likelihood-enhanced fast translation functions. Acta Crystallogr D Biol Crystallogr 61 (2005) 458-64.
[5] J.C. Whisstock, and A.M. Lesk, Prediction of protein function from protein sequence and structure. Q Rev Biophys 36 (2003) 307-40.
[6] M. Gerstein, A. Edwards, C.H. Arrowsmith, and G.T. Montelione, Structural genomics: current progress. Science 299 (2003) 1663.
[7] A. Yee, K. Pardee, D. Christendat, A. Savchenko, A.M. Edwards, and C.H. Arrowsmith, Structural proteomics: toward high-throughput structural biology as a tool in functional genomics. Acc Chem Res 36 (2003) 183-9.
[8] R.I. Sadreyev, and N.V. Grishin, Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds. BMC Struct Biol 6 (2006) 6.
[9] A. Andreeva, D. Howorth, S.E. Brenner, T.J. Hubbard, C. Chothia, and A.G. Murzin, SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32 Database issue (2004) D226-9.
[10] 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.