Sindh Univ. Res. Jour. (Sci. Ser.) Vol.43 (2) 161-168 (2011)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Grid and Distributed Public Computing Schemes for Structural Proteomics: A Short Overview

A.A. SHAH, I.A. ISMAILI, L.D. DHOMEJA, D. BARTHEL*, AND N. KRASNOGOR*

imdad.ismaili,lachhman@usindh.edu.pk , (dxb, nxk)@cs.nott.ac.uknatalio.krasnogor@nottingham.ac.uk

Institute of ICT, University of Sindh, Pakistan

+Corresponding author: A.A. SHAH, e-mail: azhar.shah@usindh.edu.pk, Cell. 92-3473707691

Received 28rth June 2011 and Revised 11th July 2011)

Abstract: Grid and distributed public computing schemes has become an essential tool for many scientific fields including bioinformatics, computational biology and systems biology. The adoption of these technologies has given rise to a wide range of projects and contributions that provide various ways of setting up these environments and exploiting their potential resources and services for different domains of applications. This paper aims to provide a distilled overview of some of the major projects, technologies and resources employed in the area of structural proteomics. The major emphasis would be to briefly comment on various approaches related to the *gridification* and parallelization of some flagship legacy applications, tools and data resources related to key structural proteomics problems such as protein structure prediction, folding and comparison. The comments are based on theoretical analysis of some interesting parameters such as performance gain after *gridification*, user level interaction environments, workload distribution and the choice of deployment infrastructure and technologies. The study of these parameters would provide a basis for some motivating justification needed for further research and development in this domain.

Keywords: Grid Computing, Distributed Computing, Structural Proteomics

1. INTRODUCTION

A thorough overview of recent scientific literature related to life sciences provides evidence that the Grid and distributed public computing schemes have become essential tools for disciplines such as bioinformatics, computational biology and systems biology. This may be perhaps due to the interdisciplinary, problem-rich and resource-hungry nature of these disciplines coupled with the ever growing number and size/complexity of biological databases and analytical tools. For example, there are about one thousand biological databases dispersed around the world and several thousands of analytical tools packed as different suites and libraries (e.g. European Molecular Biology Open Software Suite (EMBOSS)) (Rice *et al.,* 2000). These are made available by different institutes and organizations as web services to be added to custom programs and applications. The availability of these large-scale databases and analytical tools has given rise to a number of institutional, national and international grid-related projects and applications that explore the exploitation of proper grid infrastructures and enabling technologies for life sciences (e.g. myGrid (Stevens et al., 2003), EGEE (Gagliardi *et al.,* 2005), Swiss BioGrid (Michael *et al.,* 2006). These dedicated grid projects build their infrastructure setup through the federation of different *institutionally-owned* resources and allow their respective research communities for the authenticated and authorized exploitation of shared resources. There are also some other projects that build their infrastructure setup around *publicly-owned* unused computing resources which are voluntarily provided throughout the world (aka distributed *public computing schemes)* such as the World Community Grid that supports the Human Proteome Folding Project, Folding@Home (Pande *et al.,* 2007), Predictor@Home (Taufer *et al.,* 2005) and Rosetta@Home etc. All these projects attempt to provide an integrated environment for e-Science through appropriate layers of service infrastructure as shown in **(Fig. 1)**.

This paper focuses on the exploration of some of these flagship projects in terms of their use in key research areas of structural proteomics that deals with the high-throughput determination of protein 3D-structures, their comparison and functional annotation. A particular focus is to find out and compare various approaches related to the gridification/parallelization of some flagship legacy applications, tools and data resources by analyzing key parameters such as job/data distribution and management, user level interaction environments, deployment technologies and infrastructures, and the effect of gridification on overall performance of the system. The output of these findings is summarized in section **2,** whereas Section **3,** provides

*School of Computer Science, University of Nottingham, NG81BB, UK

+An earlier version of this paper was presented at ISPA07, Niagra Falls, Canada.

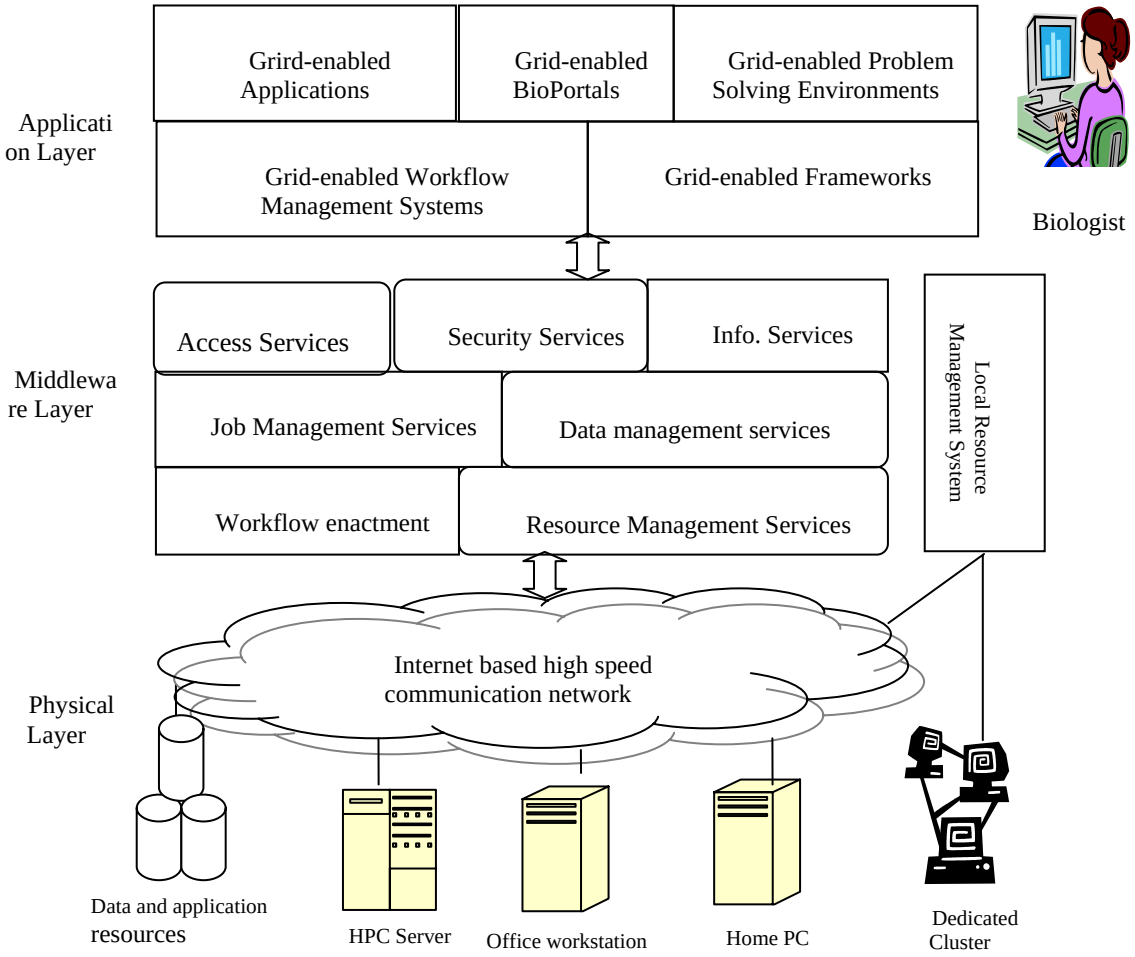

Fig. 1. Components and Services of a Generic BioGird Infrastructure (reproduced from (Shah *et al.*, 2010)

further information on middleware tools/approaches for grid-based management of structural proteomics data. Finally section 4 concludes with some comments on the indication and description of potential future trends related to the further exploitation of the Grid and its enabling technologies for structural proteomics.

2. MAJOR PROJECTS AND APPLICATIONS

Even though a very large number of protein primary structures (sequences) are known, he number of their corresponding 3D-strcutures (secondary or tertiary structures) is lagging far behind. For example, there are about 200,000 known protein sequences as compared to just 33,000 protein structures. The reason behind this sequence-structure gap is due to the difficulties associated with experimental structure determination methods such as X-ray crystallography and NMR spectroscopy. As secondary and tertiary structures are more helpful in tracing the evolution and function of the protein as well as in rational drug design, in order to reduce the gap between known sequences and known structures, computational approaches have been proposed for the prediction of these structures from a given protein sequence. As all these approaches are based on some form of modeling (such as *ab initio* or *de novo* protein modeling and comparative protein modeling techniques such as homology modeling and protein threading) and rely on multi-scale optimization techniques to optimize various model parameters (e.g. energy minimization), the availability of powerful computing facilities is essential. That is, structural proteomic methodologies require huge computational power and reliable access to various distributed and (often) heterogeneous biological databases and analytical tools in order to properly and accurately predict the structure from a given sequence or compare thousands of models against a target structure. Therefore, many research groups in this field such as Baker Laboratory at University of Washington, The Scripps Research Institute (TSRI) at California, Pande Group at Stanford University, High Throughput Computing Group at Osaka

University and others have started to make use of the Grid and distributed computing environments. We discuss some of these initiatives in the following sections.

2.1 Protein Structure Prediction

Protein structure prediction has remained an open problem in the field of structural proteomics (Zhang *et al.,* 2005). New methods are being explored and investigated at various research institutes and groups throughout the world. Evaluation of the quality and performance of these methods is carried out every two years through the *Critical Assessment of Techniques for Protein Structure Prediction* (CASP) competition. In order to provide best predicted results for a target protein, the use of grid and distributed computing public schemes has been successfully demonstrated through various projects. For example, researchers at TSRI (*The Scripps Research Institute)* have developed a distributed public computing based protein structure prediction super computer (*Predictor@Home*) using *Berkley Open Infrastructure for Network Computing* (BOINC) software. The BOINC software provides a set of different daemons that can be used for job submission, distribution and management as well as for data access, integration and storage in a distributed global grid environment consisting of publicly devoted heterogeneous PCs interconnected through the internet (aka *CPU-scavenging or cycle-scavenging*). The predictor itself consists of a set of complex protocols with increasingly sophisticated models that rely on standard software tools such as BLAST, SAM-T02, PSIPRED, MFOLD simulation (for conformational sampling) and CHARMM (for molecular simulations). It is reported that during the 6th *Critical Assessment of Protein Structure Prediction Methods* (CASP) competition 6786 users participated in the Predictor@Home project and contributed a total compute time of about 12 billion seconds, the equivalent of about 380 years of computation on a single desktop machine, within just 3 months time. This computation power had been exploited for appropriate conformational sampling and refinement of the predicted structures of 58 CASP6 targets. The quality of the predicted structures utilizing the public computing infrastructure was compared with results using a dedicated local cluster (64 nodes, 2.4 GHz Pentium Xeon processors, 1GB RAM, 1GB Ethernet network). The results of the comparison indicate that the vastly larger distributed computing power afforded by the BOINC implementation resulted in far better predictions than using the dedicated cluster. Another grid based approach that enhances the quality and performance of structure prediction has been demonstrated in (Susumu et al., 2005). It builds on the standalone web server named ROKKY (designed at Kobe University) that was ranked 2nd best prediction web server in the fold recognition category of CASP6 experiment. ROKKY uses a combination of standard analysis tools (PSI-PLAST and 3D-Jurry) and the *Fragment Assembly Simulated Annealing* (FASA) technique using the SimFold (Chikenji *et al.,* 2003) software package. In order to further enhance the quality of prediction and performance, a grid-based workflow design and control tool was added that allows the end-user to create/design a structure prediction experiment and submit it for execution on the Grid. As compared to *Predictor@Home*, the addition of a GUI-enabled workflow in ROKKY facilitates the dynamical interaction between the user and the prediction experiment during its execution. That is, the user can modify input parameters and/or methods based on the real-time inspection/monitoring of the current predicted results. It has been reported (Susumu *et al.,* 2005) that for target T0198, the workflow-based prediction gave a faster result that was closer to the target structure compared to employing a non-workflow based prediction, which uses simple batch files for job submission. This illustrates the importance of allowing the user to dynamically interact with the "production pipeline" even when the software is being distributed across the grid.

Likewise, another ambitious project, *Encyclopedia of Life* (EoL), attempts to predict structural information for all the proteins in all known organisms. The estimated computation time required for annotation of about 1.5 million sequences (as of 2003) using a pipeline of computational tools (such as TMHMM, PSORT, SignalP, WU-BLAST, PSI-BLAST and 123D) has been approximated to be 1.8 Million CPU hours (more than 300 years!) on a single 1.8 GHz CPU. In order to facilitate this task, a grid-based workflow management systems has been proposed and demonstrated (Birnbaum et al., 2004), which builds on the *AppLeS Parameter Sweep Template* (APST) technology providing an appropriate application deployment logistic and an adaptive scheduling and execution environment. The workflow was tested by running more than 54,000 proteome annotation jobs requiring 13670.5 CPU hours during the four days of the Super Computing Conference (SC'03) on a grid testbed. This consisted of 215 CPU nodes managed at ten different sites having different operating systems and local resource management software. Further details of some grid-based protein structure prediction applications are presented in **(Table 1)**.

Table 1. Grid-based applications for *protein structure prediction*

Project/ Application	Grid technologies and tools	Job/data distribution	Effect of gridification
ProtFinder (Herrera *et al.*, 2006; Huedo *et al.*, 2004) Prediction of thermodynamic properties of orthologous proteins	Globus based *GridWay Framework* that uses adaptive scheduling for dynamic grids Condor/G based GRAM GIIS Server for resource discovery, GASS and GridFTP for data handling User interaction with job submission agent through API or command line.	Prediction of 88 sequences of the Triose Phosfate Isomerase enzyme was carried out in parallel by submitting an array job with 88 parallel tasks specified in a *Job Template* File.	Jobs were run on 64 heterogeneous nodes located at different sites of IRISGrid and EGEE project. It took an average of about 43 minutes for the entire experiment
PSA/GAc (Tanimura *et al.*, 2004) Parallel Simulated Annealing using Genetic crossover	*NetSolve* based client-server application model through GridRPC API NetSolve agent keeps the service registry and monitoring information Client queries the agent and then communicates with the introduced server through GridRPC API for user interaction	Several iterated *simulated annealing* calculations were distributed in parallel for execution on NetSolve servers, whereas *genetic algorithm crossover* was performed at the client side in order to reduce the communication delays	It took about 16 minutes for the prediction job with population of 16 and crossover interval of 100 MCsweep

2.2 Protein Folding

Unlike the problem of protein structure *prediction*, where the goal is to obtain the final configuration of a given protein sequence, the problem of protein *folding* is to determine the dynamical aspects of the process involved. Folding is a thermodynamically driven process taking a few micro seconds, in which a protein adopts its native state. A proper understanding of this process would shed light into many issues at the core of biotechnology, such as the design of new proteins with a desired functionality, the understanding of some incurable diseases such as cancer or neurodegenerative diseases (e.g. Alzheimer's, Creutzfeldt-Jakob disease(CJD), Cystic fibrosis (CF), Huntington's disease (HD) and many other practical implementations of nanotechnology. The computational technique that helps in understanding the folding process uses simulations that require extremely high computational power far beyond the limits of any single traditional super computer or local cluster. It has been demonstrated in the *Folding@Home* project that this requirement can be met with a world wide distributed public-resource computing network that interconnects thousands of loosely coupled heterogeneous publicly-owned and voluntarily devoted PCs. *Folding@Home* uses an *'ensemble dynamics'* algorithm that performs *M* independent simulations with the same amino acids coordinates but with different velocities on *M* distributed processors such that each simulation starts with a slightly different initial condition and pushes the system through a free energy minimization process. This algorithm gives an *M* times speedup for the simulation of folding dynamics and thus avoids the overall waiting in free energy minima. Similarly, the process can be repeated in order to effectively handle multiple free energy barriers (multiple translations for complex folding dynamics). *U*sing a modified version of the Tinker molecular dynamics code, *β-hairpin* and *villin* were simulated and their folds successfully determined. Further details on a selection of grid-enabled protein folding applications are presented in **(Table 2)**.

Table 2. Grid-enabled applications for *Protein Folding*

Project/ Application	Grid technologies and tools	Job/data distribution	Effect of gridification
CHARMM (Natrajan *et al*., 2001) Chemistry at HARvard Molecular Mechanics	*Legion* grid operating system that provides process, files system, security services and resource management Simple command line interface with basic commands for job submission, monitoring and result visualization	Overall task: to study the energy and entropy of folded and unfolded states of a protein About 400 CHARMM jobs were run in parallel with different initial conditions	Grid test-bed :1020 nodes at 6 institutional sites 15% speedup in computational time.
CHARMM (UK *et al.*, 2002)	United Devices (UD) MetaProcessor (MP) platform for DesktopGrid Master/worker model Master (MP Server) controls and manages all the tasks and uses IBM DB2 for data storage Each worker runs a *UD Agent* with task API to run the task module and communicate with the server. API based user interaction	The overall job was to obtain efficient protein folding of 56 residue protein src-SH3 domain with different algorithmic approaches such as best-first, depth-first and breadth-first. The job is distributed into set of 50 work units (*work pool*) each having the same protein conformation but different random seed number. Each work-unit consists of 100,000 simulation steps	A work-pool consisting of 50 work-units was shot every 5 minutes on the heterogeneous platform of 45 desktop machines. Analysis of the results indicates that DesktopGrids are suitable for protein folding.

2.3 Protein Structure Comparison

Based on the diversity of the simulation results for a variety of molecules (*'from the nonbiological PPA helices to the 36-residue villin headpiece')* it has been observed that there is no single universal folding process and even sequences which fold to the same structure may have different folding processes. The comparison of protein three-dimensional structures based on a variety of similarity measures is a key component of the most challenging structural proteomic tasks, such as understanding the evolution of protein networks, protein function determination and, of course, protein folding and protein structure prediction.

As the number of known protein structures goes on increasing the size of their corresponding databases (such as PDB) also increases and hence, the process of structure comparison requires more efficient algorithms, which could exploit the power of web and grid computing technologies to provide accurate and optimal results with enhanced reliability and fault tolerance. One such approach has been demonstrated in (Ferrari et al., 2003), which employs a distributed grid-aware algorithm with indexing techniques based on geometric properties. Hash tables are used in order to partition the PDB database into sub-tables that can be dynamically updated with new entries in the PDB and provide fast and accurate comparison results. It used a Globus and MPICH based Grid testbed consisting of four nodes (each with 300 MHz CPU). Experiments were performed comparing a target against 19,500 PDB structures in about 19 seconds.

Another related approach is presented in (Barthel et al., 2006) describing a meta-server for *Protein Comparison, Knowledge, Similarity, and Information (ProCKSI),* integrating multiple protein structure comparison methods such as the *Universal Similarity Metric* (USM), the *Maximum Contact Map Overlap* (MaxCMO), and an algorithm for the alignment of distance matrices (DALI), amongst others. Additionally, it produces a consensus similarity profile of all similarity measures employed. The application runs on a mini-cluster and provides a web-based interface (http://www.procksi.net/) for job submission and result visualization. Further details of some grid-enabled applications for protein structure comparison are presented in **(Table 3)**.

Table 3. Grid-based protein *structure comparison*

Project/application	Grid technologies and tools	Job/data distribution	Effect of gridification
FROG (Park *et al.*, 2002; Park *et al.*, 2003) Fitted Rotation and Orientation of protein structure by means of real-coded Genetic algorithm	Ninf Grid RPC Master-slave model Asynchronous parallel programming in C language Web-based interactive user interface through NinfCalc tool.	The initial population is generated on the master node, which then copies three non-redundant parents to each node in the grid repeatedly.	With a generation size of 2000, population size of 100 and a crossover rate of 50, the comparison of 1j7n_d4 and 4tli took about 97 minutes on a single machine where as it takes only 17 minutes on a grid testbed of 16 nodes.
PROuST (Cannataro *et al.*, 2005) Ontology and workflow based grid-enablement of PROuST application for protein structure comparison.	PROTEUS problem solving environment with Globus-based grid infrastructure. Unified Modeling Language (UML) activity diagram workflow language specification Graphical user interfaces for workflow composition, browsing, selection and result visualization.	The PROuST application is divided into independent phases such as pre-processing, similarity search and structural alignment Each phase is implemented as an independent sub-workflow/software component	The independent sub-workflows are stored in the PROTEUS workflow meta-data repository, which has been used for the development of overall grid-aware PROuST application

3. GRID-BASED DATA MANAGEMENT APPROACHES

As important as computational horsepower, data sharing and re-using has become a fundamental aspect of almost all modern applications in the field of life sciences. This is due to the very large size and number of datasets used and the frequency by which these are updated. There have been several contributions targeted at the provision of some universal grid-based data standards, models, tools and technologies. Some of these efforts are summarized in **(Table 4)**.

Table 4. Grid-based data management approaches for structural proteomics

Project/application	Middleware technologies	Databases
ISPIDER (Zamboulis *et al.*, 2006) *In Silico Proteome Integrated Data Environment Resource*	*OGSA-DAI:* wrapping databases as web services *OGSA-DQP: distributed/parallel query processing* *AutoMed:* global schema transformation	*gpmDB*: global proteomics machine database *PEDRo*: Proteome Experimental Data RepOsitory *PepSeeker*: proteome peptide identification
e-HTPX (Allan *et al.*, 2004)	Globus Grid-FTP with web services (SOAP protocols) Web portal based user interface implemented with Java Server Pages (JSP), Java Beans and Servlets running on Apache Tomcat.	Automated collection of proteomics data from different academic projects and laboratories and its storage in to public databases such as Macromolecular Structure Database (MSD) and Protein Data Bank (PDB).

4. CONCLUSIONS AND FUTURE TRENDS

It has been observed from the reviewed literature that both Grid and distributed public computing schemes have been used successfully in the field of structural proteomics for both compute and data intensive applications. The former is powered by standard grid middleware technologies such as Globus, Legion, NetSolve, Ninf, myGrid, Condor/G etc., whereas the latter is powered by BOINC, UD MetaProcessor etc. In fact, the diversity of enabling technologies for grid and distributed computing makes it difficult for the developer to select most appropriate technological infrastructure with proven technological standards and tools. Various demonstrations reviewed in this paper are aimed at providing a roadmap in this dilemma. It has been observed that selection of an appropriate grid/distributed computing approach mainly depends on the nature of the application. For example, applications with an independent and parallel nature of jobs are more suitable for distributed computing based on publicly-owned resources. On the other hand, if the application requires some pre-determined and controlled quality of service in terms of data and process management with enhanced reliability and security, then organizational or cross-organizational grid infrastructure with standard middleware would serve in a better way. Furthermore, some middleware technologies such as Nimrod and Legion are considered as best middleware choices for applications with adjustable parametric data values. Similarly Condor/G and Proteus based grid workflow environments provide the highest level of abstraction

needed for an easy and efficient exploitation of grid resources from an ordinary user's point of view. However, in order to further extend the level of this abstraction, the future intake of proper grid infrastructure for structural proteomics should move towards the development of more interactive user interfaces and high level services in terms of portals, problem solving environments and workflow management systems. A more comprehensive and detailed review of web and grid technologies in the life sciences can be found in (Shah *et al.*, 2010).

5. ACKNOWLEDGEMENTS

N. Krasnogar acknowledges the BBSRC for funding project BB/C511764/1 and the EPSRC for project GR/T07534/01. A.A Shah acknowledges University of Sindh, Pakistan for scholarship FSCH/ SU/PLAN/F.SCH/794/2006.

REFERENCES:

Allan, R., R. Keegan, D. Meredith, M. Winn, G. Winter, (2004) e-HTPX– HPC, Grid and Web-Portal Technologies in High Throughput Protein Crystallography. Proceedings of the UK e-Science All Hands Meeting, Nottingham, UK.

Barthel, D., J. D. Hirst, J. Blacewicz, N. Krasnogor, (2006) Procksi: a metaserver for protein comparison using kolmogorov and other similarity measures. In Late Breaking Papers. European Conference on Computational Biology, Eilat, Israel.

Birnbaum, A., J. Hayes, W. Wilfred, (2004) Grid Workflow Software for a High-Throughput Proteome Annotation Pipeline. In: Konagaya, A., Satou, K. (eds.): Lecture Notes In Bioinformatics, Vol. (**3370**): 68–81.

Cannataro, M., M. Comin, C. Ferrari, C. Guerra, A. Guzzo, P. Veltri, (2005) Modeling a Protein Structure Comparison Application on the Grid Using PROTEUS. Lecture Notes in Computer Science, Vol. (**3458**): Springer Berlin/ Heidelberg, 75-85.

Chikenji, G., Y. Fujitsuka, and S. Takada, (2003) A reversible fragment assembly method for De Novo protein structure prediction. Journal of Chemical Physics, Vol. (**119**): 6895-6903.

Ferrari, C., C. G. Guerra, Zanotti, (2003) A grid-aware approach to protein structure comparison. Journal of Parallel Distributed Computing, Vol. (**63**): 728–737.

Gagliardi, F., B. Jones, F. Grey, M.E. Begin, M. Heikkurinen (2005) Building an infrastructure for scientific Grid computing: status and goals of the EGEE project. Philosophical Transactions: Mathematical, Physical and Engineering Sciences, Vol. (**363**): 1729-1742.

Herrera, J., E. Huedo, R.S. Montero, and I. M. Lorente, (2006) Benchmarking of a joint IRISGrid/EGEE Testbed with a Bioinformatics Application. Scalable Computing: Practice and Experience, Vol. (**7**): 25–32.

Huedo, E., S.M. Ruben, and M. L. Ignacio, (2004) Adaptive Grid Scheduling of a High-Throughput Bioinformatics Application. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Vol. (**3019**): 840-847.

Michael, P., M. Sergio, K. Peter, (2006) The Swiss BioGrid Project: Objectives, Preliminary Results and Lessons Learned. 2nd IEEE International Conference on e-Science and Grid Computing (e-Science 2006) Workshop on Production Grids. IEEE Computer Society Press, Amsterdam, Netherlands.

Natrajan, A., M. Crowley, D.N. Wilkins, M.A. Humphrey, A.D. Fox, A.S. Grimshaw, C.L. Brooks, (2001) Studying Protein Folding on the Grid: Experiences using CHARMM on NPACI Resources under Legion. In Proceeding of the HPDC Conference, San Francisco, CA, USA

V.S. Pande, I. Baker, J. Chapman, (2003) Atomistic Protein Folding Simulations on the Submillisecond Time Scale Using Worldwide Distributed Computing. Biopolymers, Vol. (**68**): 91–109.

Park, S.J., M. Yamamura, (2002) FROG (Fitted Rotation and Orientation of protein structure by means of real-coded Genetic algorithm): Asynchronous Parallelizing for Protein Structure-Based Comparison on the Basis of Geometrical Similarity. Genome Informatics, Vol. (**13**): 344-345.

Park, S. J., M. Yamamura, (2003) Two-layer Protein Structure Comparison. Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03).

Rice, P., I. Longden, A. Bleasby, (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics, Vol. (**16**): 276-277.

Shah, A.A., D. Barthel, P. Lukasiak, J. Blacewicz, and N. Krasnogor, (2010) Web & Grid Technologies in Bioinformatics, Computational Biology and Systems Biology, Current Bioinformatics, Vol. 3 (**1**): 10-31/

Stevens, R.D., A.J. Robinson, and C.A. Goble, (2003). myGrid: Personalized bioinformatics on the information grid. Bioinformatics, Vol. (**19**): 302-304.

Susumu, D., F. Kazutoshi, M. Hideo, N., Haruki, S. Shinji, (2005) An Empirical Study of Grid Applications in Life Science Lesson learnt from Biogrid project in Japan. International Journal of Information Technology, Vol. (**11**): 16-28.

Tanimura, Y., K. Aoi, T. Hiroyasu, M. Miki, Y. Okamoto, J. Dongarra, (2004) Implementation of Protein Tertiary Structure Prediction System with NetSolve. Proceedings of the High Performance Computing and Grid in Asia Pacific Region, Seventh International Conference on (HPCAsia'04)

Taufer, M., C. An, A. Kerstens, and C.L. Brooks, (2005) Predictor@Home: a protein structure prediction

supercomputer based on public-resource computing. Proceedings of 19th IEEE International Parallel and Distributed Processing Symposium, ISBN: 0-7695-2312-9.

UK, B., M. Taufer, T. Stricker, G. Settanni, A. Cavalli (2002) Implementation and characterization of protein folding on a desktop computational grid - is CHARMM a suitable candidate for the united devices metaprocessor?. Technical Report 385, ETH Zurich, Institute for Computer systems.

Zamboulis, L., H. Fan, K. Belhajjame, J. Siepen, (2006) Data Access and Integration in the ISPIDER Proteomics Grid. Lecture Notes in Computer Science, Vol. (**4075**): Springer Berlin / Heidelberg, 3-18.

Zhang, Y. and J. Skolnick, (2005) The protein structure prediction problem could be solved using the current PDB library. Proceedings of US National Academy of Science, Vol. (**102**): 1029-1034.