In Proceedings of UK e-Science All Hands Meeting, 2005
pp 577-580. ISBN 1-904425-53-4.

Towards A Grid-Enabled Biomolecular Simulation Database

Bing Wu[1], Kaihsu Tai[1], Muan Hong Ng[2], Steven Johnston[2], Stuart Murdock[2], Hans Fangohr[2], Mark S.P. Sansom[1], Jonathan Essex[2], Paul Jeffreys[1] and Simon Cox[2]

[1]University of Oxford and [2]University of Southampton

Abstract

The overall aim of the BioSimGrid project is to build a distributed data and computing platform to enable large scale analysis of the results of biomolecular simulations. To address the challenges of distributed computing on large amounts of simulation data, we have built an open software framework system based on off-the-shelf middleware, SRB (Storage Resource Broker) and the distributed enterprise database, Oracle 10g. We describe the implementation details: architecture, data distribution, security, and implemented analysis modules. We outline two examples, among many more, where BioSimGrid is a necessary tool.

1. Background

Biomolecular simulations enable us to explore the conformational dynamics of complex molecules such as proteins, membranes and nucleic acids. In particular, molecular dynamics (MD) [1] is widely used to investigate nanosecond to microsecond dynamics for a wide range of biomolecules. Currently, a typical simulation generates large amount of digital data.

The overall aim of the BioSimGrid project [2] is to build a distributed data and computing platform to enable large scale analysis of the results of biomolecular simulations. In particular the project will establish generic procedures for comparative analysis of simulations of biomolecules. We also wish to integrate simulation data with those emerging from post-genomic approaches to structural biology.

2. Challenges

Computer technology advances enable computational biologists to simulate larger molecules for longer timescales. Currently, a typical simulation may have a system size of ~100,000 particles (atoms), and a nanosecond timescale simulation may require ~1,000,000 timesteps (i.e. iterations of integrating the equations of motion). Such a simulation would take a few weeks on between ~8 and ~64 processors (depending upon the efficiency of the simulation code and protocols employed) and could generate gigabytes of data for subsequent analysis and visualisation.

Furthermore, data is archived in an *ad hoc* fashion at the level of individual laboratories and consequently, even medium-scale comparisons between multiple simulations are not possible unless the simulations are performed within a single research group. This excludes simulation results from the domain of structural bioinformatics, where new information and knowledge is derived by comparisons between the results of individual research endeavors.

Given the context, it is impossible to perform a simulation comparison across any two research groups within the BioSimGrid collaborators (see the BioSimGrid location topology in Figure 1) using conventional methods.

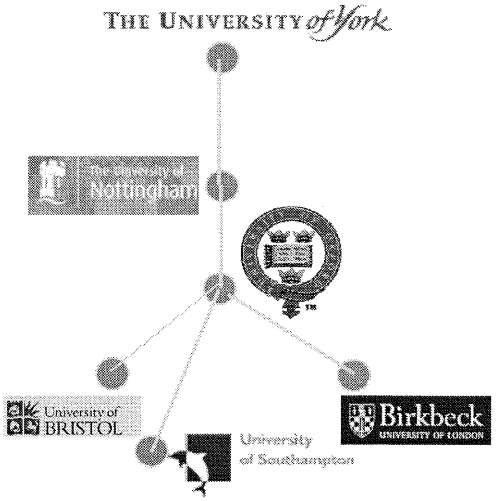

Figure 1. The BioSimGrid Collaborators

3. Solution – Grid-Enabled Architecture

To address the challenges of distributed computing on large amounts of simulation data (over 2 TB storage for 1000 trajectories), the BioSimGrid project has built an open software framework system based on off-the-shelf Grid middleware, SRB (Storage Resource Broker) [3] and the distributed enterprise database, Oracle 10g [4]. Using the power of the Grid [5] users are able to access data efficiently and effectively, thus enabling them to take a comparative, i.e. genuinely biological, approach to analysis of simulation data. The Grid infrastructure of the distributed system has been implemented in 6 universities (Oxford, Southampton, Birkbeck College, London; Bristol, Nottingham and York) in the UK. We use SRB to distribute large amount raw simulation data across our collaboration sites. Each BioSimGrid site hosts 2.5 TB data storage sharing with other sites. Then we use Oracle 10g database to store small amount of metadata (~20 GB) and this is distributed over 6 sites for easy local access and better application performance.

The current implementation of BioSimGrid is based on multi-tier open architecture. We have 4 tiers of applications:

- GUI: HTTP(S)-based web client, provides user interaction with the system. The client can be either a standard web browser or web-based client. The use of the web browser eliminates development and maintenance of client software.
- Service: This tier is dedicated to delivering data and application services to GUI clients. The top parts are application servers which handle communications between web applications and services. Underneath this is the AAA (Authentication Authorisation and Accounting) service. There are also supporting services such as monitoring, transaction, and distributed query services.
- Grid Middleware: We build our distributed data solution based on SRB (Storage Resource Broker). All the data access requests are handled transparently by the Data Engine module.
- Data/Database: Oracle 10g has been deployed as the core database. Meta data resources are distributed across collaborating sites. The flat files of simulation data is distributed via SRB.

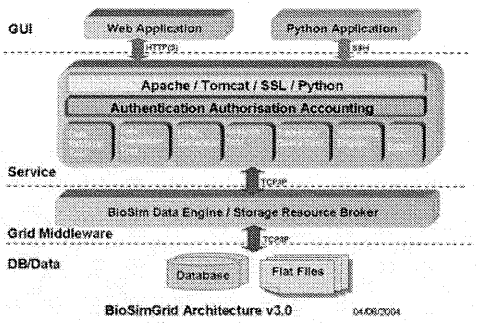

Figure 2. The BioSimGrid Open Architecture

4. Security

Security is a critical element of the project as we have a distributed system implementation. The system needs a secure and robust environment to guarantee smooth access to various BioSimGrid services. In order to achieve this, a number of security mechanisms have been integrated into the system. We have implemented PKI (Public Key Infrastructure) and X.509 Digital Certificate [6,7] based authentication. Two kinds of UK e-Science certificates (X.509 certificates) have been used in the system: host certificates for the BioSimGrid servers hosting resources and user certificates for users to access the BioSimGrid resources.

We provide two levels of authentication methods for a user to access BioSimGrid. The first is a secure and robust authentication mechanism based on digital certificates using Grid certificates issued by UK e-Science Certification Authority (CA). The other level of authentication is a user/password based authentication. While user/password based authentication enables wider access of the portal via a public PC from anywhere in the world, users of certificate based access can have more enhanced level of security for access and authentication.

5. Example Applications

Biomolecular applications are the key of the BioSimGrid project. We establish a formal database for biomolecular simulations within the UK, increase collaboration via a distributed computing environment. The current BioSimGrid supports a wide range of analysis tools including: RMSD, RMSF, Internal Angles, Molecular Surfaces & Volumes, Principal Component Analysis, Average Structures, Centres of Mass, Geometry, Inter-atomic

distances, Secondary Structure analysis, and Entropy calculations. All these analysis tools are available via either the BioSimGrid web portal [8] for easy access or a Python based client for expert users. We are currently using these tools to develop a set of standard analyses which may be used to provide an overall measure of the quality of a simulation (Murdock et al., ms. in preparation).

We outline two examples, among many more, where BioSimGrid is a necessary tool: The first is the protein Nitrogen Regulatory Protein C (NTRC) and deposited in BioSimGrid for analysis and archiving (Kinns et al., unpublished results). NTRC plays a key role in the regulation of nitrogen metabolism, undergoing a change in conformation upon phosphorylation by a kinase. Molecular dynamics simulations at a range of temperatures have been used to explore this conformational change, Figure 3.

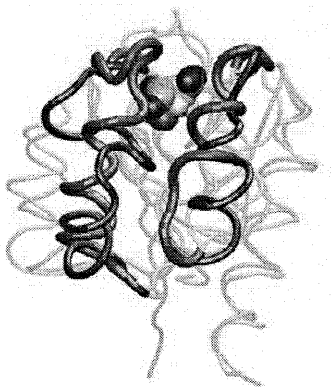

Figure 3. The native (blue) and phosphorylated (red) conformations of NTRC

The second is the comparison of the active site dynamics of four hydrolases: AChE, acetylcholinesterase; OMPLA, an outer-membrane phospholipase A; OmpT, an outer-membrane protease; and PagP, an outer membrane acyl transferase (Tai et al., ms. in preparation). The dynamics of these enzymes are described in 17 MD trajectories, from two different laboratories and three different researchers. Structural data show that these enzymes share some similarities in their active sites (a triad of amino acid side chains suggested to be involved in their catalytic mechanisms). The outer membrane proteins are all β-barrels, but otherwise their structures are unrelated. The acetylcholinesterase shares no structural similarities with the other enzymes apart from its active site. Thus, we wish to explore the extent to which active site dynamics are preserved within this group of enzymes.

In addition to the biological differences, this test case also provides a microcosm of a more general challenge for comparative simulations. The AChE simulations were performed at University of California, San Diego; the OMPLA and other outer membrane protein simulations in Oxford. Normally, such data would never "meet" and would reside in the separate laboratories. Furthermore, the simulations were run using different programs and protocols and the data are in different formats. Rather than re-run one (or both) of the simulations (which would consume many days of costly supercomputer time), we are using BioSimGrid to make the comparison. Thus this apparently simple test case enables us to test many aspects of the underlying methodology. We have used the BioSimGrid toolkit to generate a simple metric for active site integrity, and have compared how this evolves in time for the four enzymes. Preliminary results from this analysis reveal the sensitivity of the metric to the simulation conditions and the oligomerization state of the protein (in the case of OMPLA).

6. Summary

The BioSimGrid project is still developing and deploying wider scale of its applications and integrating with other systems. We are currently working on automating the process of depositing data directly to BioSimGrid from simulations running on NGS (National Grid Service) [9] clusters.

Acknowledgements

We would like to thank our BioSimGrid collaborators (Charles Laughton, Adrian Mulholland, Leo Caves, David Moss and Oliver Smart) for their input to this project. Our thanks to all of our colleagues in the Oxford and Southampton simulation labs, to Matthew Dovey in the OeSC, and to our colleagues in the Southampton Regional e-Science Centre for their encouragement and advice. Thanks also to Marc Baaden for providing access to the OMPLA and OmpT data, and to Katherine Cox for the PagP data.

BioSimGrid is funded by BBSRC and DTI.

References

[1] Karplus, M.J. and McCammon, J.A. (2002). Nature Struct. Biol., 9, 646-652.

[2] http://www.biosimgrid.org

[3] http://www.sdsc.edu/srb/

[4] http://www.oracle.com/database/db_collateral-.html

[5] Foster, I. and Kesselman, C., Eds. (1999). The GRID: Blueprint for a New Computing, Morgan-Kaufmann.

[6] Tuecke, S., Engert, D., Foster, I., Thompson, M., Pearlman, L. and Kesselman, C. (2001). Internet X.509 Public Key Infrastructure ProxyCertificate Profile, IETF.

[7] ftp://ftp.isi.edu/in-notes/rfc2459.txt

[8] https://portal.biosimgrid.org

[9] http://www.ngs.ac.uk/