# BIOSIMGRID: A DISTRIBUTED DATABASE FOR BIOMOLECULAR SIMULATIONS

Bing Wu[1,2], Kaihsu Tai[1], Stuart Murdock[3], Muan Hong Ng[4], Steve Johnston[4], Hans Fangohr[4], Paul Jeffreys[2], Simon Cox[4], Jonathan Essex[3] and Mark S.P. Sansom[1,]*

[1]Department of Biochemistry, University of Oxford, [2]e-Science Centre, University of Oxford, [3]Department of Chemistry, University of Southampton, [4]e-Science Centre, University of Southampton

**to whom correspondence should be addressed: mark.sansom@biop.ox.ac.uk*

**Keywords**: Grid, Biomolecular Simulation, Molecular Dynamics, Protein, OGSA, OGSA-DAI, Distributed Database, Web Service, Middleware, Digital Certificate

## Abstract

Biomolecular simulations provide data on the conformational dynamics and energetics of complex biomolecular systems. We aim to exploit the e-science infrastructure developing in the UK to enable large scale analysis of the results of such simulations. In particular, the BioSimGrid project (www.biosimgrid.org) will provide a generic database for comparative analysis of simulations of biomolecules of biological and pharmaceutical interest. In order to address the challenges of distributed computing on large amounts of simulation data the project is building an open software framework system based on OGSA (Open Grid Service Architecture) and OGSA-DAI (Open Grid Service Architecture Database Access and Integration). The system will have a service-oriented computing model using Grid-based Web service technology to deliver data analysis. Data mining services will be provided for the biomolecular simulation and structural biology communities. To address the security problem of the heterogeneous BioSimGrid environment, a Grid certificate-based and a user/pass based authentication mechanism will be integrated across the system. Furthermore, a distributed authorization and accounting mechanism is discussed to enhance the security.

## 1. Introduction

Biomolecular simulations enable us to explore the conformational dynamics of complex molecules such as proteins and nucleic acids. However, at present there are considerable problems in comparing the results of multiple simulations, and in integrating simulation results with other (experimentally-derived) sources of data. This problem will become more pressing if simulation studies are to match up to post-genomic approaches such as high throughput protein crystallography.

As protein structure determination becomes more automated, and as advances are made in structural bioinformatics [1] and computational biology, it will become increasingly important that biomolecular simulations do not exist as a standalone physicochemical analysis of single systems, but rather that they become embedded in a matrix of computational an experimental studies of proteins (see Fig. 1). Furthermore, it will be essential to provide data retrieval and analysis tools that are accessible to a wide community of structural and cell biologists, not jut to simulation specialists. It is in this context that the BioSimGRID project is being developed. The overall aim of this project is to exploit the developing e-science infrastructure in the UK to enable large scale analysis of the results of biomolecular simulations. In particular the project will provide generic procedures for comparative analysis of simulations of complex biological macromolecules and systems. We also wish to integrate simulation data with those emerging from post-genomic approaches to structural biology.

In this paper we provide a brief overview of biological simulation and molecular dynamics. This is followed by a description of the BioSimGRID database, of the overall software architecture of the project, and of the database distribution. Application development is described using an example work flow. We conclude with some details of security implementation in the project, and a brief discussion of future directions.

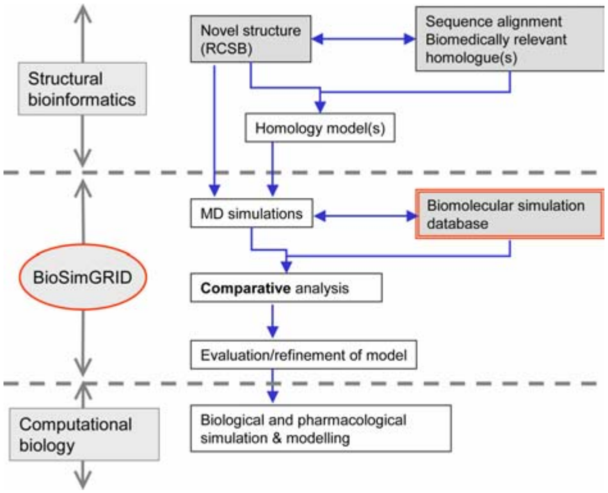

Structural bioinformatics
Novel structure (RCSB)
Sequence alignment Biomedically relevant homologue(s)
Homology model(s)
MD simulations
Biomolecular simulation database
BioSimGRID
Comparative analysis
Evaluation/refinement of model
Computational biology
Biological and pharmacological simulation & modelling

*Figure 1: BioSimGRID in the context of structural bioinformatics and computational biology.*

## 2. Biological Simulation and Molecular Dynamics

Molecular simulations with atom-level resolution, first performed more than 25 years ago, have now entered the mainstream of biological research [2]. In particular, the molecular dynamics (MD) method, which solves the Newtonian equations of motion for a atoms interacting via an empirical classical forcefield, has become a staple for investigating the nanosecond to microsecond dynamics of a wide range of biopolymers, including DNA, proteins and membranes.

MD has benefited considerably from improvements in computer technology. As computers become faster, biologists have become able to explore larger molecules for longer timescales. Furthermore, the advent of cheap commodity cluster ('Beowulf') computing has had a significant impact on the numbers of research groups undertaking biomolecular simulation studies being undertaken. Currently, a typical simulation may have a system size of ~100,000 particles (atoms), and a nanosecond timescale simulation may requires ~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.

The status quo for the archiving of these data is far from optimal. Typically, data is archived in an *ad hoc* fashion at the level of individual laboratories. Furthermore, the reporting of the simulation *metadata* is by traditional journal article publishing, and can be prone to omission of technical details in the interests of brevity. 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, and from biology in general, where new information is derived by comparisons between the results of individual research endeavours. The BioSimGrid project [3] aims to provide a possible solution to this difficulty by providing a framework to enable and facilitate comparisons of biomolecular simulations.

As a case study of the information that can arise from comparisons between simulations, we will take an example of a family of simulations that have been performed in one laboratory, thus enabling comparisons to be made using current technologies. Glutamate receptors (GluRs) are the major excitatory neurotransmitter receptors in mammalian brains. Structural biology studies have revealed that the neurotransmitter-binding domains of mammalian glutamate receptors share a common fold with a bacterial GluR [4, 5] and with bacterial periplasmic binding proteins. Comparative MD simulation studies of mammalian GluR2 [6], of bacterial GluR0 (Arinaminpathy, Sansom and Biggin, unpublished results) and of a bacterial periplasmic binding protein [7] have revealed that these functionally disparate proteins also share a common pattern of change in dynamics upon neurotransmitter/ligand binding. The extension of this comparative approach to further members of the periplasmic binding protein family will reveal to what extent conformational dynamics are conserved across a family of proteins with a similar protein fold. However, generalising this approach across multiple protein folds would be somewhat challenging without a simulation database and interrogation tools.

## 3. A Bimolecular Simulation Database

In an ideal world all simulation data would be available to all interested parties. However, at present simulation data reside in the 'home' laboratory and are not accessible to other research groups. Indeed, even within the home laboratory, pressures on disk storage in the past have been such that once initial analysis is complete and papers have been published, data are archived to tape and sometimes lost.

One solution would be to deposit all simulation results in a centralised database such as the RCSB Protein Data Bank [8, 9]. However, in reality the amounts of data are such that a centralised database and rapid access are problematic. Thus, even though the cost of data storage continues to fall, managing this volume of data centrally is non-trivial. Furthermore, there are the problems of physically maintaining and curating a centralised database. However, using the Grid technologies [10, 11] we have an opportunity to draw together distributed collections of simulation data in disparate formats, whilst maintaining a centrally accessible meta-database.

The BioSimGRID project will establish a formal database for biomolecular simulations within the UK, increasing collaboration via a distributed computing environment. There are three levels of data existing in the database (see Fig. 2):

- Raw data: generated by biomolecular simulations;
- 1st level metadata: describing the generic properties of raw simulation data, such as data location, simulator configuration, etc;
- 2nd level metadata: describing the results of generic analyses of simulation data. This will be produced by a suite of generic analysis tools and will provide simulation data 'kite-marks'.

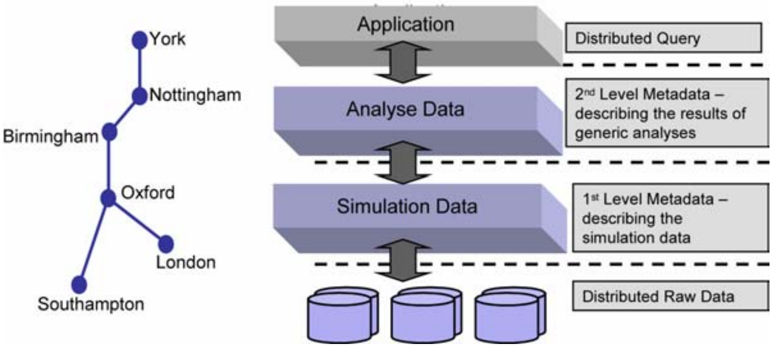


*Figure 2: An overview of the BioSimGRID database*

Once all these data are in place, software tools will be developed for interrogation and data-mining across the entire distributed database. In this way we will enable data-mining across all simulations in the database. This will also democratise simulation data by providing relevant metadata, including links to structural biology and genomics database entries, thus facilitating access to and understanding of biomolecular simulation results for non-specialists.

## 4. System Architecture

An overview of the BioSimGRID system architecture is given in Fig. 3. The project is building an open software framework system based on OGSI (Open Grid Services Infrastructure) and OGSA (Open Grid Service Architecture) [12, 13], the *de facto* standards in Grid computing. These standards are implemented in the middleware, namely Globus Toolkit 3 (GT3) [14], which provides a community-based, open-architecture, open-source set of services and software libraries enabling applications to handle distributed and heterogeneous computing resources as a single virtual machine through Grid/Web services.

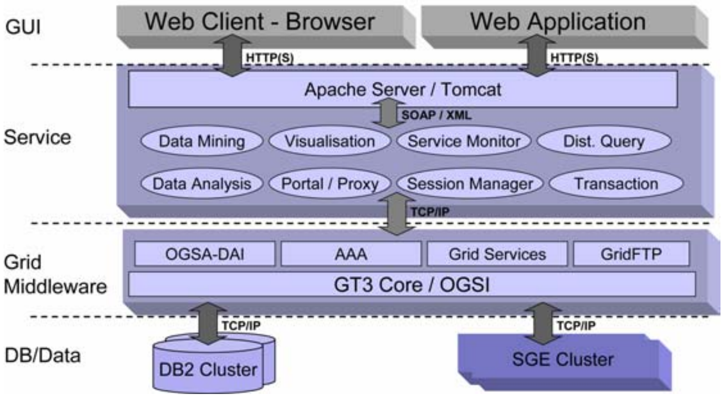


*Figure 3: System architecture*

The framework system of BioSimGRID has a service-oriented multi-tier computing model. The system architecture currently consists of the following components.

*GUI:* An HTTP(S)-based web client provides user interaction with the system. The client can be either a standard web browser or web-based application. The use of the web interface eliminates development and maintenance of client software. The user can interact with the BioSimGrid from anywhere and with “anything” (laptop, PDA, etc.).
*Service:* This tier is dedicated to deliver data analysis and data mining services to the biomolecular simulation and structural biology communities through Grid-based Web services. There are also supporting services such as monitoring, transaction, and distributed query services. The protocols used here are XML/SOAP.
*Grid Middleware:* This tier contains the central element of the Grid middleware GT3 Core, which provides the core services and capabilities required to construct a computational grid. On top of the GT3 Core, there is a set of components that implement basic services, such as security, resource, management, database access and communications.
*Database/Data*: Data and database resources are distributed across collaborating sites. The Grid middleware enables the access of these resources transparently in a format of virtual machine. Within the multi-tier system architecture, the associated applications can be developed independently as distributed services and integrated into the system as well as using off-the-shelf middleware components. This also improves the system scalability and flexibility.

## 5. Database Distribution

To address the challenges of distributed computing on large amounts of simulation data (we estimate an initial size of >2 TB storage for ~1000 trajectories), the project uses a leading commercial database, namely IBM DB2 Universal Database Enterprise Server [15]. A prototype database system has been implemented in a distributed environment in two universities (Oxford and Southampton). This will be rolled out to a further four collaborating universities (Birkbeck College, London; Birmingham; Nottingham; and York) in the UK within the next 12 months. In the longer term, a number of possible data curation solutions will be investigated in order to maintain large amounts of distributed simulation data for extended periods of time (e.g. decades).

We intend that the future database distribution will be based on OGSA-DAI (Open Grid Service Architecture Database Access and Integration) [16] technology. This project is co-developed by UK e-Science centres (Edinburgh, Manchester and Newcastle) and industrial partners (IBM, Oracle and Microsoft). OGSA-DAI defines open standards and open source based uniform service interfaces for accessing heterogeneous database/data resources within OGSA using OGSI. Through the OGSA-DAI interfaces distributed and heterogeneous data resources can be accessed and controlled as though they were a single logical resource. The concept of OGSA-DAI is based on delivering database as Grid/Web services using standard services, i.e. *OGSA-DAI = DBMS + XML + Distributed SQL*. By using OGSA-DAI, BioSimGRID applications will deliver services based on distributed queries with minimal programming efforts. The security, transaction management, distributed database access and job management are integrated internally into the BioSimGRID applications. Furthermore, heterogonous databases (DB2, Oracle9i [17], etc.) can be used in the project without the need for any changes to the application code.

## 6. Application Development and Example Work Flow

Prototype V1.0 of BioSimGRID has two types of GUI clients: a web portal client and an application client. These will provide software tools for interrogation and data-mining across the entire distributed database, and a set of generic analysis tools for biomolecular simulations. The object of applications is to be able to perform different analysis techniques across a range of different biosimulation trajectories. Many of the individual analysis techniques are already commonly in use. We will enable analyses to be performed on many trajectories with ease and with confidence that the analysis applied to each trajectory is the same.

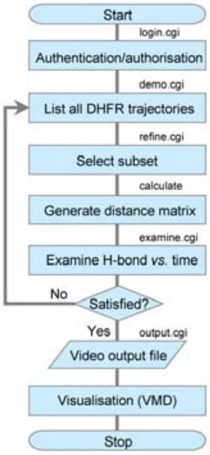


*Figure 4: An example of a simple BioSimGRID work flow*

Fig. 4 is an example of a work-flow for the current implementation of BioSimGRID applications. The client uses a web-based front-end to login to the system. Once the client passes the authentication and authorisation, (s)he can browse the available trajectories in the BioSimGRID database by selecting from a list of metadata option on a web form. After the input from the user selection, a set of variables, which we call the database query handler (this consists of a trajectory id, frame selections and an analysis choice) is sent back from the client to the web server. The web server passes on this handler to the application server to start up an analysis. After the analysis is complete, the simulation results are displayed on the browser. If the user is satisfied with the analysis results, (s)he can choose to download selected simulation data which are sent to the client site for visualisation. This will automatically invoke VMD (Visual Molecular Dynamics) [18, 19] to play the simulation. The user can also choose to discard the analysis and select different trajectories.

The current implementation of the prototype tools is based on Python [20] – a powerful object orientated language which is becoming popular within the biosimulation community. There is also a popular biosimulation library called MMTK (Molecular Modelling Toolkit) [21, 22] which is written in Python and which can provide many complex functions. The prototype tools we have incorporated include surface and volume, internal angle, and RMSD calculations of protein molecules using the MMTK Library. In addition to MMTK-based tools we have also written our own analysis tools, including molecular visualisation methods, and geometrical calculations. Molecular visualisation is performed with VMD, a standard package widely used within the biosimulation community that is capable of providing both static representations and animations of molecules. As data analysis is fundamental to this project we have included facilities for viewing numerical data. As well as producing PNG format representations of graphs we will enable use of Grace [23] to facilitate customisation of data analysis results. Both data and image files may be downloaded to the client's machine.

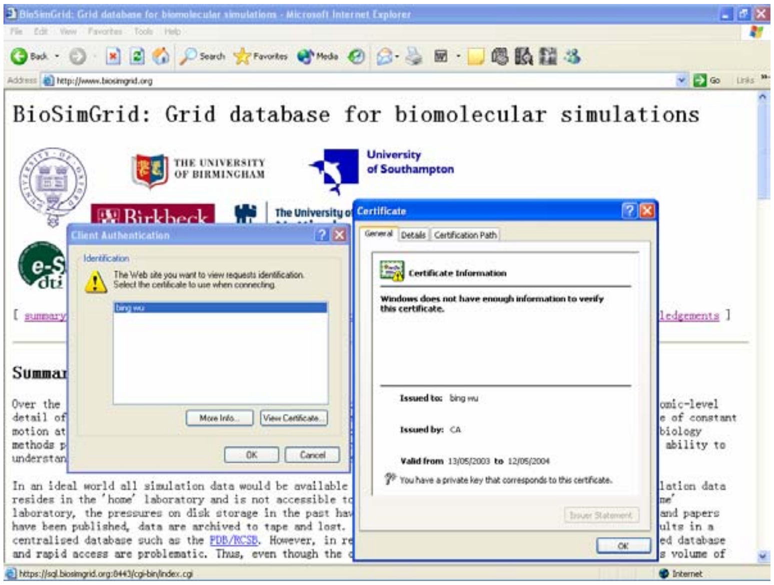
Our future plans include enabling acceptance of user-contributed analyses, thus aiding the continuous development of the toolkit. We also intend to enable deposition of BioSimGRID calculated results into the database for future reference.

## 7. Security Implementation

The security implementation of the system will be based on three core components: Authentication, Authorisation and Accounting.

### 7.1 Authentication

To guarantee high security and easy accessibility to the heterogeneous BioSimGrid environment, two levels of authentication infrastructures have been implemented in the system. The first level is based on a digital certificate. A Grid certificate-based authentication mechanism (OpenCA) has been integrated across the system. This is based on PKI (Public Key Infrastructure) and X.509 digital certificate [24] technology. When a user wants to access specific BioSimGrid services, the subject of his/her X.509 personal digital certificate is verified against the one stored in the corresponding database and only if the security check is successful can the user have the access to the services.


BioSimGrid: Grid database for biomolecular simulations - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.biosimgrid.org
Go Links
THE UNIVERSITY OF BIRMINGHAM
University of Southampton
Birkbeck
The University of
e-Science dti
Client Authentication
Identification
More Info... View Certificate...
OK Cancel
Certificate
General Details Certification Path
Certificate Information
Issuer Statement
OK
https://sql.biosimgrid.org:8443/cgi-bin/index.cgi
Internet

BioSimGrid: Grid database for biomolecular simulations

The Web site you want to view requests identification. Select the certificate to use when connecting.

bing wu

Windows does not have enough information to verify this certificate.

Issued to: bing wu

Issued by: CA

Valid from 13/05/2003 to 12/05/2004

You have a private key that corresponds to this certificate.

*Figure 5: digital certificate based authentication*

The second security level is that of user/pass based authentication. This is designed for those who do not have digital certificates installed on their client machines. This level of authentication enables web access of the system via a public pc anywhere in the world.

All the web accesses are based on Secure Socket Layer (SSL) via HTTPS. This means each packet is individually encrypted against the public key in the TCP/IP level. The authentication is currently based on user credential delegation implemented using MyProxy [25].

### 7.2 Authorisation

Authorisation of the web portal is handled by the authorisation component of the system. BioSimDB has an internal security database containing the necessary user account information. This component queries the security database to retrieve the user account information such as username, organisation, and access level. Once this information is confirmed, the user will have access his/her permitted area(s).

### 7.3 Accounting

All the transactions of the system are logged in the database. To maintain efficiency of the database, a distributed accounting mechanism will be developed as part of the system to enhance the security as well as to trace the system usage. User activities will be stored in the nearest account database of the system. To retrieve the accounting information of a particular user, an accounting component will be developed. This has to be based on distributed queries across the entire database.

## 8. Summary and Future Developments

The BioSimGRID project is still at an early stage of its development. In particular we need to refine the database schema (not discussed here), with particular attention to simulation metadata, and to develop methods for data deposition and for quality control of simulation data. This will enable us to run initial comparative analyses on complex simulations (e.g. of biological membranes) in order to evaluate the strengths and weaknesses of the current prototypes in real world applications.

## Acknowledgements

Many thanks to our BioSimGRID partners (L. Caves, C. Laughton, D. Moss and O. Smart) for their numerous inputs to this project. BioSimGRID is funded by BBSRC and DTI. Research in MSPS's laboratory is funded by BBSRC, EPSRC, MRC and the Wellcome Trust. Our thanks to all of our colleagues in the Oxford and Southampton simulation labs, to Matthew Dovey and Ivaylo Kostadinov in the OeSC, and to our colleagues in the Southampton Regional e-Science Centre for their encouragement and advice.

## References

[1] Bourne, P.E. and Weissig, H. (2003) Structural Bioinformatics, Wiley-Liss, Hoboken.
[2] Karplus, M.J. and McCammon, J.A. (2002) Nature Struct. Biol., 9, 646-652.
[3] http://www.biosimgrid.org
[4] Armstrong, N., Sun, Y., Chen, G.-Q. and Gouaux, E. (1998) Nature, 395, 913 - 917.
[5] Mayer, M.L., Olson, R. and Gouaux, E. (2001) J. Mol. Biol., 311, 815-836.
[6] Arinaminpathy, T., Sansom, M.S.P. and Biggin, P.C. (2002) Biophys. J., 82, 676-683.
[7] Pang, A., Arinaminpathy, Y., Sansom, M.S.P. and Biggin, P.C. (2003) FEBS Lett., (*in press*) http://www.elsevier.nl/febs/14/175/article.html.
[8] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) Nucl. Acids Res., 28, 235-242.
[9] http://www.rcsb.org/pdb/
[10] Berman, F., Fox, G. and Hey, T., Eds., *Grid Computing: Making the Global Infrastructure a Reality* (Wiley, 2003).
[11] Foster, I. and Kesselman, C., Eds., *The GRID: Blueprint for a New Computing* (Morgan-Kaufmann, 1999).
[12] Foster, I., Kesselman, C., Nick, J. and Tuecke, S., The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration, Global Grid Forum (2002).
[13] Grimshaw, A. (2002) The Evolution of Grid Standards, Avaki Corporation.

[14] http://www.globus.org
[15] http://www-3.ibm.com/software/data/db2/udb/
[16] http://www.ogsa-dai.org
[17] http://www.oracle.com/ip/deploy/database/oracle9i/
[18] Humphrey, W., Dalke, A. and Schulten, K. (1996) J. Molec. Graph., 14, 33-38.
[19] http://www.ks.uiuc.edu/Research/vmd/
[20] http://www.python.org/
[21] Hinsen, K. (2000) J. Comp. Chem., 21, 79-85.
[22] http://dirac.cnrs-orleans.fr/MMTK/
[23] http://plasma-gate.weizmann.ac.il/Grace/
[24] Tuecke, S., Engert, D., Foster, I., Thompson, M., Pearlman, L. and Kesselman, C., Internet X.509 Public Key Infrastructure ProxyCertificate Profile, IETF (2001).
[25] http://www.ncsa.uiuc.edu/Divisions/ACES/MyProxy/