The CSBG - LSU Gateway: Web based Hosted Gateway for Computational System Biology Application Tools from Louisiana State University

Eroma Abeysinghe
Science Gateways Research Center,
Pervasive Technology Institute,
Indiana University
eabeysin@iu.edu

Michal Brylinski
Department of Biological Sciences,
Louisiana State University
michal@brylinski.org

Marcus Christie
Science Gateways Research Center,
Pervasive Technology Institute,
Indiana University
machrist@iu.edu

Suresh Marru
Science Gateways Research Center,
Pervasive Technology Institute,
Indiana University
smarru@iu.edu

Marlon Pierce
Science Gateways Research Center,
Pervasive Technology Institute,
Indiana University
marpierc@iu.edu

ABSTRACT

Science gateways are identified as an effective way to publish and distribute software for research communities without the burden of learning HPC (High Performance Computer) systems. In the past, researchers were expected to have in-depth knowledge about using HPC systems for computations along with their respective science field in order to do effective research. Science gateways eliminate the need to learn HPC systems and allows the research communities to focus more on their science and let the gateway handle communicating with HPCs. In this poster we are presenting the science gateway project of CSBG (Computational System Biology Group - www.brylinski.org) of Department of Biological Sciences with Center for Computation & Technology at LSU (Louisiana State University). The gateway project was initiated in order to provide CSBG software tools as a service through a science gateway.

CCS CONCEPTS

• **Computing methodologies → Simulation evaluation**; • **Software and its engineering** → *Software design engineering*;

KEYWORDS

Apache Airavata, Science Gateway, Computational System Biology, Bioinformatics

ACM Reference Format:
Eroma Abeysinghe, Michal Brylinski, Marcus Christie, Suresh Marru, and Marlon Pierce. 2018. The CSBG - LSU Gateway: Web based Hosted Gateway for Computational System Biology Application Tools from Louisiana State University. In *PEARC '18: Practice and Experience in Advanced Research Computing, July 22–26, 2018, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3219104.3229245

PEARC '18, July 22–26, 2018, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-6446-1/18/07.
https://doi.org/10.1145/3219104.3229245

1 INTRODUCTION

The Computational System Biology Group (CSBG) at LSU with consultation and development support from the Science Gateways Community Institute (SGCI) [4] are developing the sciencegateway.brylinski.org gateway for existing and new users of CSBG tools. Currently the group has about half a dozen software tools developed and in production. They are compiled and installed in LSU cluster (mike.hpc.lsu.edu) with plans expanding to the SuperMIC XSEDE cluster (smic.hpc.lsu.edu). The gateway is envisioned to be practical in terms of reaching large number of users and managing them through the gateway as well as a way to reduce the number of errors and issues of individuals compiling and installing the software on their own.

2 CSBG TOOLS

CSBG offers a variety of tools for Structural Bioinformatics to support the prediction of protein structure and function from raw sequence data Table 1. eThread is a meta-threading procedure to accurately identify templates for protein modeling and functional annotation. This tool was shown to outperform single-threading approaches generating models correctly at the fold level for the majority of targets and detecting many facets of protein function even in a low sequence identity regime. Improved template selection by eThread motivated us to develop eFindSite, a method for ligand-binding site and residue prediction. eFindSite employs various machine learning techniques to efficiently integrate structural and evolutionary information. It has been shown to provide more accurate annotations than other methods for ligand-binding site and residue prediction. An example of the modeled structure of a gene product from the human proteome by eThread with a drug-binding site annotated by eFindSite is shown in Figure 1.

eSimDock is a similarity-based docking tool employing nonlinear machine learning-based scoring functions to improve the accuracy of ligand ranking and binding pose prediction. Importantly, the performance of eSimDock is largely unaffected by the deformation of ligand binding regions; thus it represents a practical strategy for across-proteome virtual screening against protein

Table 1: CSBG tools for Structural Bioinformatics

Application	Purpose
eThread	Protein structure modeling.
eFindSite	Drug-binding site detection.
eSimDock	Ligand docking.
GeauxDock	Ligand docking.
eRank^{PPI}	Protein-binding site detection.
eRank^{PPI}	Modeling of quaternary structures.

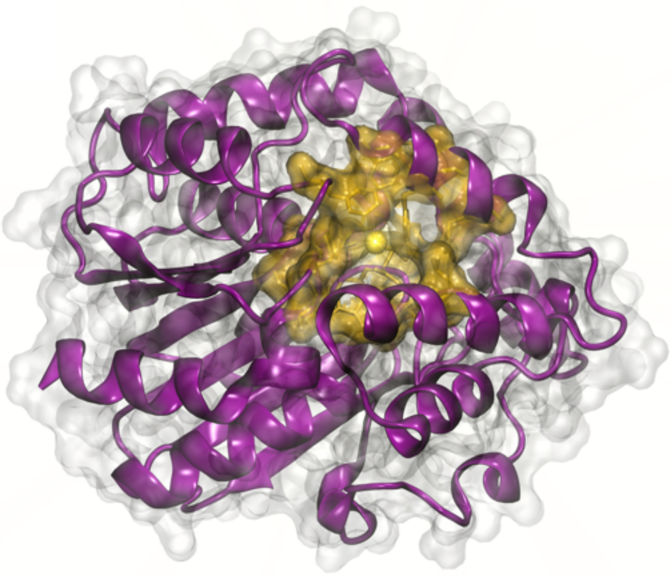


Figure 1: Example of a protein structure modeled with eThread (purple) and putative drug-binding site predicted with eFindSite (gold)

models. Moreover, we developed another ligand docking algorithm, GeauxDock, which uses a descriptor-based scoring function integrating evolutionary constraints with physics-based energy terms. GeauxDock is well suited for proteome-scale applications taking advantage of the increasingly growing protein sequence and structural data.

eFindSite^{PPI} and eRank^{PPI} are programs to predict protein-binding interfaces and assembly dimer structures. eFindSite^{PPI} uses the 3D structure of a target protein, remotely related templates and machine learning to predict binding residues. A unique feature of eFindSite^{PPI} is its capability to detect specific molecular interactions at the interface. The eRank^{PPI} is an algorithm for the identification of near-native conformations generated by protein docking using experimental structures as well as protein models. It employs multiple features including interface probability estimates calculated by eFindSite^{PPI} and a novel contact-based symmetry score. Both programs were demonstrated to consistently outperform other algorithms, offering a high accuracy in the exhaustive structure-based reconstruction of protein-protein interaction networks across proteomes.

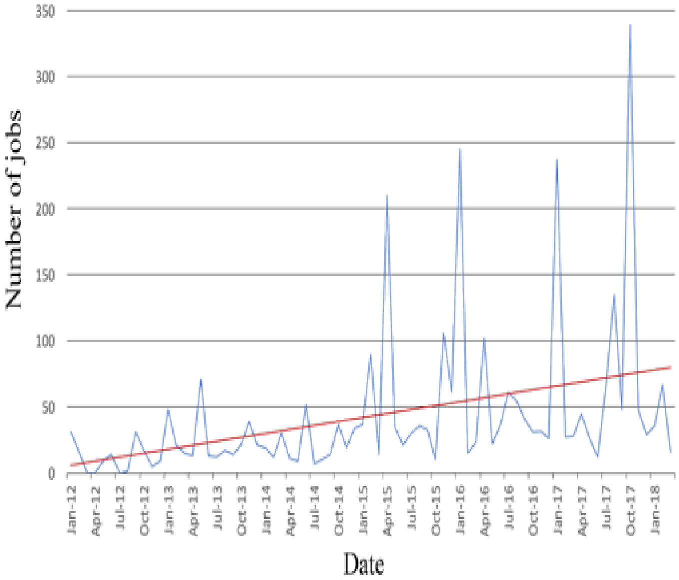


Figure 2: Number of jobs submitted by external users each month (blue). Red is the trendline.

3 CSBG USAGE

There is a growing interest in our tools from the research community. Existing web servers have been used by other groups to process 3,103 jobs since 2012. Figure 2 shows the number of jobs submitted each month. For instance, eThread and eFindSite have been used 754 and 1,476 times, respectively. The majority of users are experimental groups with limited experience in processing jobs on HPC machines. Science gateways provide an easy access to CSBG tools, which can be used to support a number of projects ranging from across-proteome function inference to drug discovery.

With the science gateway, the tools will be made available for classroom environments as needed. Managing and making these tools available for large number of users is easier in gateway environment as opposed to giving direct access in the HPC resources.

4 CSBG GATEWAY WITH APACHE AIRAVATA SERVICES

4.1 Requirement & Implementation

CSBG contacted SGCI for support for gateway project. Through SGCI, the project was assigned to the SciGaP [7] gateway team. The multi-tenanted SciGaP gateway platform provides hosted, managed gateway portals for science communities with data storage for user data files. SciGaP provides primary services required by a science gateway, user identity management, accounts, authorization, and access to multiple high performance computer (HPC) resources from campus, national, and international resource providers. The SciGaP platform uses Apache Airavata [5] middleware for computational job construction, submission to HPCs and to manage computational execution data files, inputs and outputs. The CSBG Gateway consists of a web client for end users to communicate with the LSU cluster and XSEDE cluster and dedicated data storage for

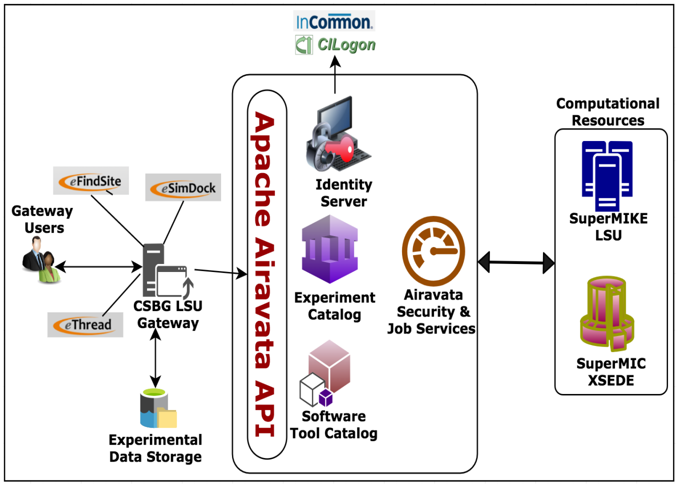
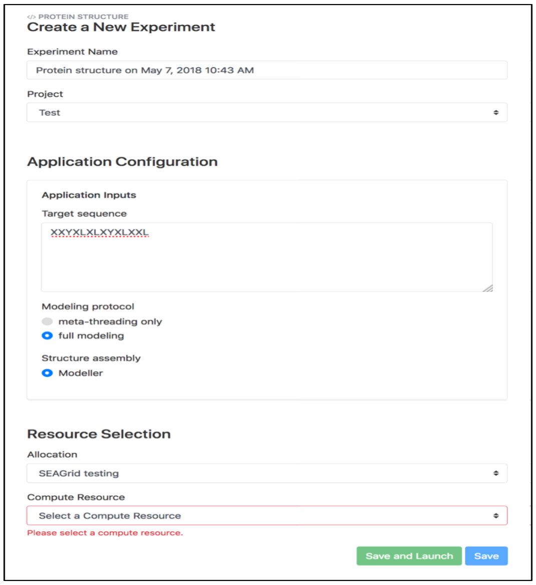


Figure 3: CSBG LSU Gateway, Data storage with Apache Airavata & Computational Resources


</> PROTEIN STRUCTURE

Create a New Experiment

Experiment Name

Protein structure on May 7, 2018 10:43 AM

Project

Test

Application Configuration

Application Inputs

Target sequence

XXYXLXLXYXLXXL

Modeling protocol

○ meta-threading only

● full modeling

Structure assembly

● Modeller

Resource Selection

Allocation

SEAGrid testing

Compute Resource

Select a Compute Resource

Please select a compute resource.

Figure 4: eThread Experiment Launch Page.

gateway user data. The gateway client, gateway storage resource and Apache Airavata middleware are hosted at Indiana University. Figure 3 depicts the aforementioned set up of gateway web client, gateway middleware components with HPC clusters.

4.2 Planning and Configuration

CSBG required customizations on top of the default gateway web client provided by the SciGaP services. Customizations mainly focused on minimizing erroneous data uploading for bioinformatic softwares. Users uploading erroneous data can result in HPC resource and time wastage and also blocking the other users from using the HPCs due to traffic. To do so, it was discussed and designed to validate data inputs prior to experiment submission in the gateway. The 'Experiment' is the record created and launched using 'Experiment Launch' page (Figure 4) in the gateway in order to create a job script and submit the job to the HPC cluster. Validations are done when user adds the inputs required for software execution. Instead of uploading a single large input file with all the data required to execute the software, the input is absorbed in the formats of input text, selecting from option buttons, checkboxes, etc. Options are nested; based on previously selected options, the current options can change. The CSBG gateway supports configuring the bioinformatic softwares with required validations. Gateway administrator can add validations for each input field at the time of configuring the application. Gateway is currently configured to run bioinformatic computations on LSU SuperMike-II cluster and can expand to any other campus, XSEDE or cloud clusters.

4.3 User Access

The CSBG gateway has the option of letting users use already existing organizational accounts (e.g.: Google, campus login, etc...) to gain access or create a new gateway user account. Apache Airavata uses open source Keycloak [2] identity management system for user account creation and to authorize and authenticate. CILogon [1] is the federated authentication system used to enable using other organizational logins to create a gateway account. Gateway administrators can decide on the level of access to the gateway that they want to grant users.

Another requirement for CSBG Gateway is to have 'guest', or 'anonymous' gateway login. These accounts are to be used by potential users who want to explore the gateway or researchers or reviewers of journal and paper submissions. Such user accounts ensure access to the gateway without revealing personal data such as name, email, and institute. While such accounts are practical from guest or temporarily user point of view, the gateway administrator would want to provide restricted access on using SuperMike-II for test type computations. As a result, the gateway development team is currently working on providing restricted access to the gateway in terms of software and HPC usage.

4.4 Secure Communication

Apache Airavata uses SSH key based communication with HPCs (computational clusters) and user data storage at the gateway client side. Gateway Credential Store [3] allows the gateway administrator to generate public-private key pairs for secure SSH communications. Gateway administrators can create as many as needed, either use one key for all the clusters and data storage resource communications or have dedicated keys created for each cluster and data storage resource. These channels are used for computational job script transfer from Apache Airavata middleware to the HPC working directory, input and output data transfers from gateway data storage to the cluster.

4.5 Bioinformatic Computations

Two bioinformatic tools, eThread and eFindSite by CSBG are currently available within the gateway for users. In order to use them users need to provide the required inputs and launch an 'Experiment' in the gateway. The experiment will submit a job script with all required information to run together with data files uploaded by the user. Once the job is submitted user can view the status through 'Experiment Summary' page. Apart from viewing the status, user can cancel a running experimental job, clone an old experiment or share the experiment with other gateway users from the experiment summary page.In the CSBG LSU gateway data sharing [6] is available at two levels, individual experiment level and project level. Project is a collection of experiments within the gateway.

5 OUTREACH EVENTS

We plan to disseminate gateways to a broad research community. The online resources will include comprehensive sets of manuals, tutorials and various case studies to encourage and help others adopt our tools in their projects. We will not only actively promote scientific gateways at various conferences, but also use these tools in the classroom. Protein structure modeling and functional inference will be included in several courses currently offered at LSU for undergraduate and graduate students. These efforts will increase the student awareness of the importance of computational approaches to modern biological research.

6 CONCLUSIONS

The CSBG group will keep enhancing and improving the current software tools and they will be made available from sciencegateway.brylinski.org. Currently the computations are possible in SuperMike-II and it is planned to be available on XSEDE SuperMIC resource. Moving forward, it will be expanded and users will have the option of selecting where to run the jobs from a wide range of HPCs ranging from campus to national, XSEDE resources. The gateway is planned to be introduced and used in classroom environments and also expanding the horizon with new users who needs to use bioinformatics softwares without the hassle of needing to learn ways of HPCs.

ACKNOWLEDGMENTS

The Louisiana State University - Computational System Biology Group Science Gateway was implemented with the support from the Science Gateways Community Institute (NSF award 1547611). Development of the Apache Airavata used to develop the science gateway is supported by NSF award 1339774.

REFERENCES

[1] Jim Basney, Terry Fleury, and Jeff Gaynor. 2014. CILogon: A federated X. 509 certification authority for cyberinfrastructure logon. *Concurrency and Computation: Practice and Experience* 26, 13 (2014), 2225–2239.

[2] Marcus A Christie, Anuj Bhandar, Supun Nakandala, Suresh Marru, Eroma Abeysinghe, Sudhakar Pamidighantam, and Marlon E Pierce. 2017. Using Keycloak for Gateway Authentication and Authorization. (2017).

[3] Thejaka Amila Kanewala, Suresh Marru, Jim Basney, and Marlon Pierce. 2014. A credential store for multi-tenant science gateways. In *Cluster, Cloud and Grid Computing (CCGrid), 2014 14th IEEE/ACM International Symposium on*. IEEE, 445–454.

[4] Katherine A Lawrence, Michael Zentner, Nancy Wilkins-Diehr, Julie A Wernert, Marlon Pierce, Suresh Marru, and Scott Michael. 2015. Science gateways today and tomorrow: positive perspectives of nearly 5000 members of the research community. *Concurrency and Computation: Practice and Experience* 27, 16 (2015), 4252–4268.

[5] Suresh Marru, Lahiru Gunathilake, Chathura Herath, Patanachai Tangchaisin, Marlon Pierce, Chris Mattmann, Raminder Singh, Thilina Gunarathne, Eran Chinthaka, Ross Gardler, et al. 2011. Apache airavata: a framework for distributed applications and computational workflows. In *Proceedings of the 2011 ACM workshop on Gateway computing environments*. ACM, 21–28.

[6] Supun Nakandala, Suresh Marru, Marlon Piece, Sudhakar Pamidighantam, Kenneth Yoshimoto, Terri Schwartz, Subhashini Sivagnanam, Amit Majumdar, and Mark A Miller. 2017. Apache Airavata Sharing Service: A Tool for Enabling User Collaboration in Science Gateways. In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*. ACM, 20.

[7] Marlon Pierce, Suresh Marru, Mark A Miller, Amit Majumdar, and Borries Demeler. 2013. *Science Gateway Operational Sustainability: Adopting a Platform-as-a Service Approach*. Technical Report.