GLYCOMICS WORKBENCH: A GRID TECHNOLOGY-BASED WORKBENCH FOR GLYCOME ANALYSIS

National University

Arun K. Datta
National University, San Diego, CA

Nitin Sukhija
Slippery Rock University of Pennsylvania, Slippery Rock, PA

Slippery Rock University Established 1889

A significant number of software tools and databases have been developed over the years for glycome analysis. However, full utilization of some of those useful tools is restricted due to the fact that those are not accessible via Web. This also restricts the semantic analysis of a vast amount of experimental data that were generated under the Consortium of Functional Glycomics (CFG), which is now hosted at the National Center for Functional Glycomics. These authors proposes the development of a grid technology-based infrastructure, termed 'Glycomics Workbench', to integrate such useful computational tools and resources that can better serve the Glyco-community. Grid technology offers multiple advantages including high scalability and Web accessibility. Grid infrastructure includes grid services, grid computing, and data grid. Grid computing provides accessibility to High-Performance Computing, such as, XSEDE. Grid services built on Open Grid Collaborating Environments (OGCE) are based on several Web service technologies. Our earlier work on neoGrid development was built on OGCE. Data grid is a commodity grid that can host exabytes of data that has become essential for glycome analysis. Our work on C-Grid development was to fulfill that need. This proposed Glycomics Workbench development, however, needs active participation of researchers for creating 'molecule page' of their interest, an effort that can be achieved through a Consortium.

MOTIVATION

- A significant number of databases and software tools have been developed over time that are useful for Glycome Analysis (1). However, accessibility of some of these tools are restricted due to the technology platform used. These tools are not accessible via Web.
- This project is to utilize Grid Technology that will provide Web accessible framework, which can be easily scalable (2,3).
- API will be created to provide Stand alone software tools accessible.
- User friendly Front-end-interface will be created for databases and made them accessible.
- Relevant educational and Training materials will also be provided for the end users.

OBJECTIVE

✓ These PIs are engaged in developing tools and technologies that can benefit Higher Education and Research communities.
✓ Developing a grid-based portal to store, manage and share large amounts of distributed health research related data in a data grid for further analysis by the researchers in a collaborative environment.
✓ Use of Grid Technology to develop solutions
✓ Develop APIs and user-friendly GUIs for accessing the software tools and databases.

BACKGROUND

- Earlier, neoGRID system was developed that provides this incentive to develop this Workbench.
- The neoGRID was designed to offer an HPC-supported collaborative environment for the researchers from multidisciplinary scientific fields to gather data (such as, C-Grid and other data resources), and analyze using workflows developed by using various workflow management tools.
- OGCE , which was used earlier for CHOIS, will be used for this development (3).
- Developed and deployed a data-grid, termed C-Grid for storing and managing big data (5).
- Recently, Metabolomics Workbench has been launched that will provide some guidance for this development.
- This system will be launched with collaborative support.

PROBLEM DESCRIPTION

✓ A significant number of software tools and databases have been developed that can't be accessible by a majority of Glycoscientists.
✓ Presently, these are not available via Web .
✓ Some efforts are underway by commercial sources (Google, GitHub, etc.) to provide the source code (e.g., GlycoWorkbench).

SOLUTION

- Developing Glycomics Workbench as an IT solution for addressing large amount of data generated by glycome analysis.
- OGCE will be used that provides scalability and Web access for useful software tools and databases
- C-Grid to store massive amounts of data generated from the analysis
- APIs will be developed for access
- XSEDE for Grid Computing and related services

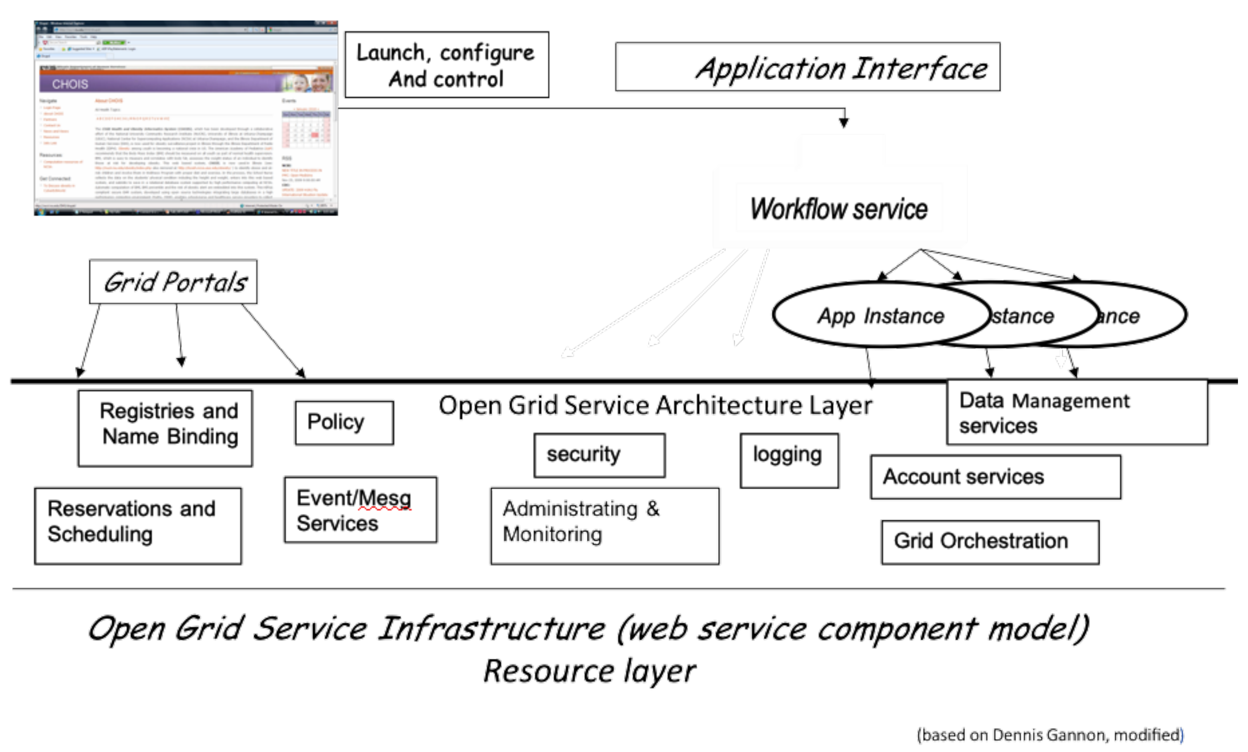

Figure 1. The proposed CI-supported Grid-based Glycomics Workbench for supporting glycome research. This has three fundamental components: Software-as-a service, C-Grid for distributed storage of data and grid computing supported by XSEDE. This is also designed to support education and training that will be stored in the C-Grid.

C-GRID

✓ C-Grid, the Community Health Grid web portal, serves as a gateway for the collaborative institutions and organizations to utilize cyberinfrastructure supported XSEDE resources for data analysis and helps the users via ez-iRODS to create and manage "virtual data collection" that can be stored in heterogeneous data resources across distributed network (5).
✓ Remote management of this data grid is performed using iRODS, the Integrated Rule-Oriented Data System, which is a middleware developed by the Data Intensive Cyber Environments (DICE) research group, and collaborators

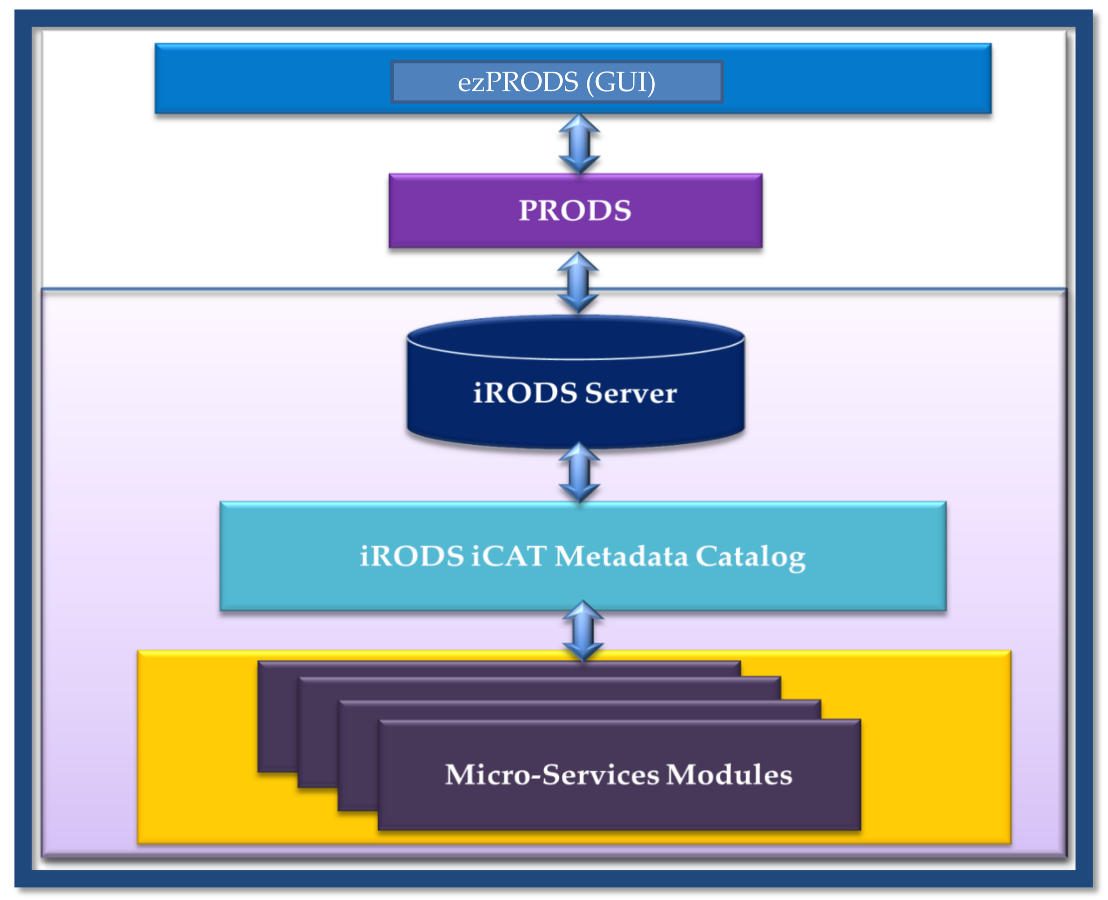

Figure 2. C-Grid or Community Grid Portal. Portal framework

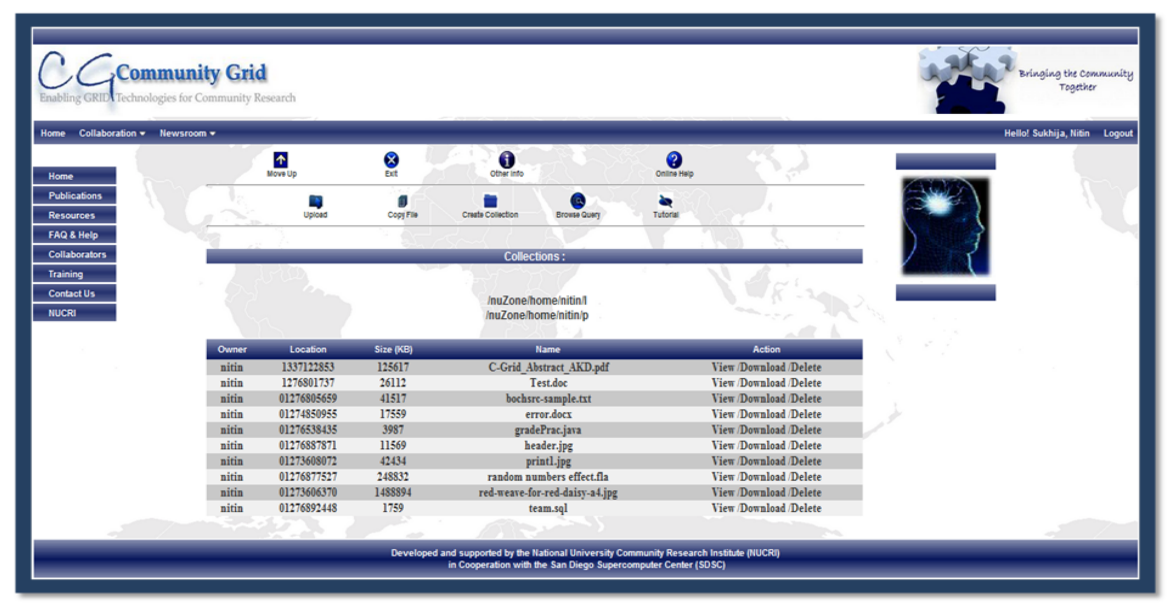

Figure 3. C-Grid or Community Grid Portal. Portal interface for viewing collections in C-Grid after role-based access with authentication.

✓ GUI component, termed ez-PRODS, has been created as a component of G-Grid to interact with iRODS located in the data grid using PRODS API.

✓ Data collected using mobile applications can be used for health monitoring while data on educational training events can be saved for a longitudinal study.

✓ Video and other big data can be saved in C-Grid

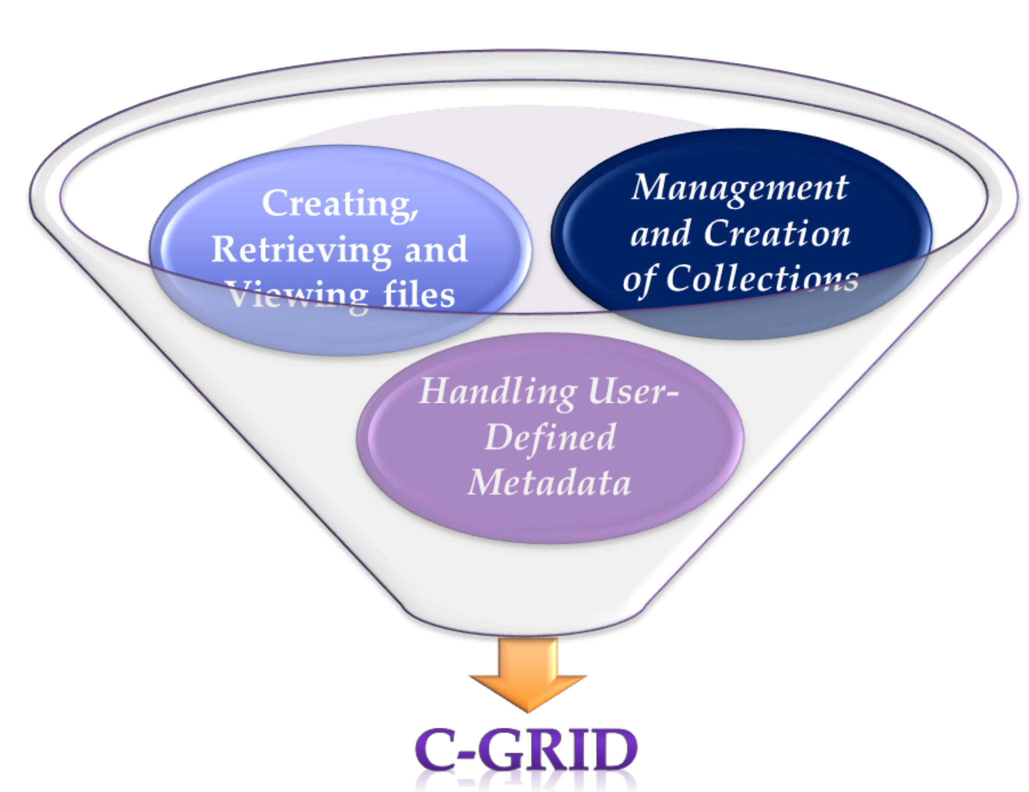

Figure 4. Core functionality offered by C-grid web portal

CONCLUSION

- C-GRID acts as a distributed computing environment and data management system for sharing resources, data and computing power with others.
- Provides collaborative data sharing and maintenance of distributed storage resource collections.
- C-Grid portal addresses the challenge of dealing with the problem of scalability of data and the data visualizations.
- Grid technology-based workbench can provide Web accessibility and is highly scalable.

FUTURE WORK

This proposed framework is under development to include necessary components and make it available to the scientific community in near future.

ACKNOWLEDGEMENT

The authors acknowledge the support of XSEDE. This project is supported by internal resources.

REFERENCES

1. Datta, AK. (2020). Glycobioinformatics in deciphering the Mammalian Glycocode: Recent Advances (2019). Nova Publishers, *in press*.
2. Nitin Sukhija and Arun K. Datta (2012). Grid Technology for Community Health Research. XSEDE 2012: eXtreme Digital Discovery, July 17 - 20, Chicago (IL).
3. Arun K. Datta, Victoria Jackson, Radha Nandkumar, Jill Sproat, Weimo Zhu , Heidi Krahling (2010). CHOIS: Enabling grid technologies for obesity surveillance and control. In 'Healthgrid Applications and Core Technologies (Eds. , T. Solomonides, I. Blanquer, V. Breton, T. Glatard, and Y. Legre), vol 159, p191-202, IOS Press, Washington D.C. ISBN 978-1-60750-582-2.
4. Arun K. Datta, Becca Eller and Anthony P. Sinai (2011). neoGRID: Enabling access to teragrid resources - Application for sialylmotif analysis in the protozoan Toxoplasma gondii. TeraGrid 11: Extreme Digital Discovery, July 18-21, Salt Lake City, Utah.
5. Nitin Sukhija, and Arun K. Datta (2013). C-Grid: Enabling iRODS-based Grid Technology for Community Health Research. Information Technology in Bio- and Medical Informatics, Lecture Notes in Computer Science (M. Bursa, S. Khuri, and M. E. Renda, Eds.), LNCS 8060, pp 17-31, Springer-Verlag, Berlin, Heidelberg, ISBN 978-3-642-40092-6); also, see: ACM ISBN 978-1-4503-6446-1/18/07. https://doi.org/10.1145/3219104.3229247