# A Prototype System for Grid-Based Cancer Biomedical Informatics

Srivatsava Ranjit Ganta  Anand Sivasubramaniam  Raj Acharya
Department of Computer Science and Engineering
Pennsylvania State University
University Park, PA 16801, USA
{*ranjit,anand,acharya*}*@cse.psu.edu*

## 1. Introduction

Biomedical Informatics deals with the study and analysis of disease related data to aid drug discovery and help find better treatment procedures. The data includes heterogeneous information such as patient records, clinical observations and experimental genomic data. These data sets are distributed among various hospitals, diagnosis and research centers that are geographically separated and independently controlled. Consequently, researchers at these centers work with islands of data and informatics tools, a situation that impedes a global study of the disease. This scenario poses the requirement for a common infrastructure that facilitates collaborative sharing of data and analysis applications among the biomedical research community. Grid technology has been successfully applied to provide computational environments for such scenarios in various scientific fields including physics, astronomy, and more specifically, bioinformatics [3] [6]. Recently, several nationwide and international grids such as National Cancer Institute(NCI)'s CaGRID [5] [1] are under development to provide a grid-based collaborative computational infrastructure for biomedical research. However, these grids are aimed at specific escience domains that have different operational procedures and goals. This requires specialized modules over the grid to facilitate the development of escience applications over grids.

In this poster, we present a prototype escience grid-system for cancer biomedical informatics research. The system simulates a grid of nodes each corresponding to individual cancer research centers willing to share cancer research data in a collaborative fashion while maintaining independence. We identify some system functionalities required for performing biomedical research in a grid environment and present the design and implementation of solutions we developed. The prototype is loaded with data collected from some leading cancer research centers as part of Pennsylvania Cancer Alliance Bioinformatics Consortium(PCABC). We also demonstrate the grid-usage and information-sharing capabilities of the system by presenting results of studies aimed at capturing the global behavior of the disease data available on the grid.

## 2. System Architecture

In this section we give brief description of the design functionalities we offer on our system. The architecture is aimed at providing a layered set of services to facilitate data negotiation and application invocation over a specialized grid. The main set of services can be divided into two categories: 1. Application Management services and 2. Discovery and Negotiation services. The Application Management Services are responsible for invocation and management of CAA(Cancer Analysis Application) applications over the grid. The Cancer Analysis Application Service Factory (CAASF) deals with management of the available CAAs advertised at each host node. One of the key functionality requirements for a collaborative cancer data analysis environment is pipelining. The idea is to direct the results from one analysis application to the other by separating the data preprocessing modules. This helps the grid users to minimize effort and computation when using multiple applications in a pipelined fashion. The CAASF handles pipelining through a specialized pipeline specification file. This XML based file specifies the set of applications and the pre/post processing of data required to handle the analysis jobs.

The second category of services deal with CAA discovery and data negotiation services. Application discovery is provided by a persistent metadata handling service, Cancer Analysis Application Metadata Manager (CAAM) that maintains the metadata corresponding to applications advertised at each grid-node. A typical cancer analysis application consists of pruning the input data set, setting the input parameters and visualizing the results. To prune the data set required for the analysis, we run a specialized service called the CDPS (Cancer Data Prune Service) that allows the client to filter the available data sources based on certain criteria. During this stage, we also implement a domain-specific service called the Cancer Data Negotiator (CDN) that facilitates negotiations on the extent of data sharing between two grid nodes. This service helps individual nodes to control the amount of data and application sharing over the grid.

## 3. Results

To demonstrate the functionalities offered by our system we run a set of cancer data studies aimed at understanding the global behavior of the cancer data available on the grid. We choose Information-Fusion based applications that we developed in [2] to emphasize the data and computational sharing capabilities offered by an escience grid system. The data used includes over 6000 samples of prostate, breast and melanoma cancer related data collected by the leading cancer research centers such as Penn State Cancer Institute, The Wistar Cancer Institute, Univ of Pittsburgh Cancer Center, Fox Chase Cancer Center as part of The Pennsylvania Cancer Alliance for Bioinformatics Consortium (PCABC). The following is a brief description of the sample studies:

1. Disease Global Statistics: The application is aimed at helping the individual cancer research centers compute global disease statistics based on the various data sets available on the grid. These studies could help identify the behavior of the disease with respect to attributes such as race, geographic location etc.

2. Clustering Diverse Data Sets: The application is based on an algorithm we developed in [4] to cluster diverse biomedical data sets. These studies are aimed at better identification of genes responsible for the disease.

## 4. Conclusion

In this poster, we present a prototype escience grid-system for cancer biomedical informatics research. We identify some domain-specific design functionalities required by an escience grid system aimed at biomedical research and present the design and implementation of the solutions we developed. We demonstrate the grid-usage and information-sharing capabilities of the system by presenting results of studies aimed at capturing the global behavior of the disease data available on the grid. We are currently working towards the methodologies to evaluate the performance of our system.

## References


[1] D. Fenstermacher, C. Street, T. McSherry, V. Nayak, C. Overby and M. Feldman. "The Cancer Biomedical Informatics Grid". *In the Proceedings of IEEE Engineering in Medicine and Biology*, Shanghai, China, September 1-4, 2005.

[2] S. R. Ganta, J. Kasturi, J. Gilbertson, and R. Acharya. "An Online Analysis and Information Fusion Platform for Heterogeneous Biomedical Informatics Data". *In the Proceedings IEEE Symposium for Computer Based Medical Systems*, Dublin, Ireland, June 23-25 2002, pp. 153-158.

[3] Y. Teo, X. Wang and Y. NG. "GLAD: a system for developing and deploying large-scale bioinformatics grid". *Bioinformatics*, 2004.

[4] J. Kasturi and R. Acharya. "Clustering of Diverse Genomic Data using Information Fusion". *Bioinformatics Journal*, 21(4): 423-429, 2005.

[5] CaBIG - Cancer Biomedical Informatics Grid. *https://cabig.nci.nih.gov/overview* , July 23 2004.

[6] NC BioGrid. *http://www.ncbiogrid.org/*.