

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

Abstract

Biomedical Informatics deals with the study and analysis of disease related data such as patient records, clinical observations and genomic experimental data to aid drug discovery and better treatment procedures. A global study of these information sources requires a computational environment that facilitates collaborative sharing of data and informatics tools. Grid technology has been successfully applied as the computing paradigm for such scenarios in various scientific fields. In the context of biomedicine, several nationwide and international grids such as The National Cancer Institute(NCI)'s CaGRID [8] are under development to provide a collaborative computational infrastructure. Based on these backbone grids, there is a need for specialized grid-based systems to solve problems involved in biomedical information sharing and analysis. In this paper, we present a prototype design and implementation for such a system aimed at prostate cancer research. We identify some domain specific issues in building such a system and present the design and implementation of our solutions. We run examples of information analysis methods used by the biomedical informatics research community and report the results we obtain for data collected from the leading cancer research centers in the state of Pennsylvania.

1. Introduction

Biomedical Informatics deals with the study and analysis of disease related data such as patient records, clinical observations and experimental genomic data. The goal here is to mine through these heterogeneous information sources to help find better treatment procedures and aid drug discovery. The data sets involved in biomedical informatics studies can be broadly divided into two categories: Clinical and Genomic. Clinical data consists of data collected at various stages of disease diagnosis and treatment. It mainly consists of patient related information such as patient demo-

graphics, clinical observations and treatment records. Genomic data includes more complex and huge experimental data sets like gene expressions, DNA and protein sequences etc. 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 [2], astronomy [11] and more specifically, bioinformatics [12]. The systems range from large-scale grids such as [15] to more specialized systems [16] that solve domain-specific problems over the grid. A grid-programming toolkit to facilitate the development and deployment of bioinformatics applications is proposed in [5]. With the recent increase of focus on Problem Solving Environments (PSE) over grids, domain specific PSEs such as [3] are being built. On the other hand, there has been a lack of large-scale grid infrastructures that focus on solving problems for biomedicine and lifesciences. However, recently, several nationwide and international grids such as National Cancer Institute(NCI)'s CaGRID [8] are under development to provide a collaborative computational infrastructure for biomedical research. These large-scale grids are aimed at specific domains that have different operational procedures and goals. This brings out a need for specialized grid based systems to cater for the domain-specific requirements of these grid infrastructures.

In this paper, we present a prototype system that provides a domain-specific grid system for cancer biomedical informatics research. Though we focus on cancer related research, the issues are largely the same for biomedical informatics studies for other diseases. We identify some domain-specific functionalities required by a grid computational en-

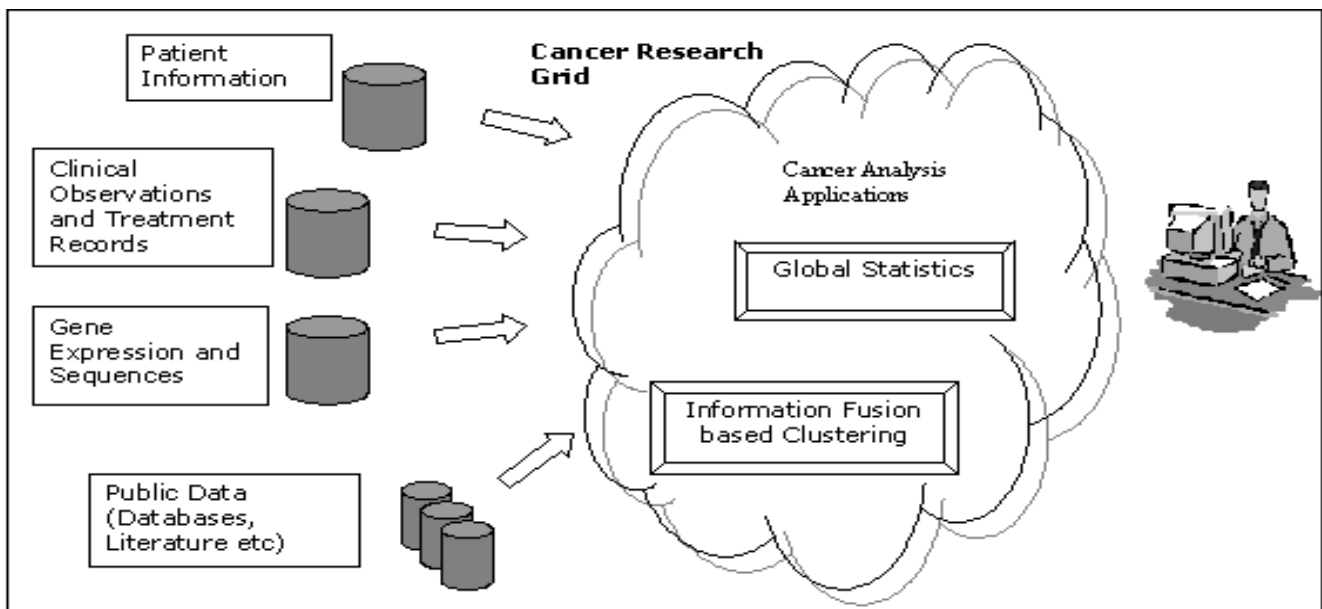


Figure 1. Information Fusion Based Cancer Biomedical Informatics

environment and present the design and implementation of solutions we developed. We run examples of some commonly used analysis paradigms on data obtained from the leading cancer research centers in the state of Pennsylvania. The rest of the document is organized as follows. Section 2 motivates the need for a biomedical analytical system. The architectural design is presented in section 3 followed by some sample results in section 4. Section 6 provides the conclusion.

2. Motivation

Cancer is one of the most deadly diseases in the history of mankind. Every year thousands of people worldwide die of this disease. It affects various parts of the human body and spreads abnormally killing the tissues and finally resulting in death. Various kinds of data are collected during the course of the disease diagnosis through patient treatment. Data collection starts with patient information such as age, sex, race and family history of prostate cancer etc. This is done mostly in paper form and in some centers collected as a relational database. Various clinical observations are then recorded to mark the stage of the disease. This is followed by the treatment procedure and the progress at different stages of the treatment. The data sets collected so far can be categorized as clinical data. Orthogonal to this, samples collected from the patients are used to conduct several genome-wide experiments in order to study behavior of the disease. These data sets are typically much larger and complex than clinical data. In addition to clinical and genomic

data sets information from publicly available literature and databases are also used in biomedical informatics.

The biomedical informatics community uses numerous analysis techniques to mine through these data sets independently. Typically, analysis studies are conducted in two overlapping methodologies. These approaches deal with a majority of either clinical data or genomic data. However, because of close inter-relationships among them, a global study of the disease requires multiple instances of these data sets. Recently, there has been a focus on using Information Fusion based techniques to mine through these data sets. These paradigms have been proven to provide a new outlook for information analysis in the biomedical scenario. The basic idea is to use data from disparate and individually controlled systems to mine biomedical data. Grid based technology offers the necessary platform to perform collaborative analysis in such scenarios. Figure 1 gives an overview of this environment.

In this paper, we use two information fusion based paradigms to demonstrate grid-based information analysis for biomedical informatics. These approaches have been proven and are currently used by the biomedical research community through a publicly available portal [4]. The first method is the “Disease Demographics Analysis” which mainly involves the collection of global statistics of disease behavior. For this purpose, we divide the data sets into categories that are linked with each other in a specific manner. The global statistic is defined as a combination of a ‘fact’ category, a set of ‘dimension’ categories, and a statistical function. For example, given a geographical region

such as North America, a global statistic of interest would be “The average age at which a North American male of African American origin is diagnosed with prostate cancer”. In this example, the ‘age-at-diagnosis’ is the fact category and ‘race’ and ‘geographic location’ of the patient are dimension categories. A qualified set of such statistics can help researchers understand the spread of the disease. Similar queries can be posed on the numerous parameters in clinical data such as tissue information, treatment procedures and also genomic experimental data such as microarrays and sequences. This kind of analysis, though not computationally complex, involves a large number of data sets. We employ the methodology that we used in [4] to implement this technique.

The second facet of biomedical informatics is the gene expression data analysis. This methodology is used in drug discovery to determine the effects of various stages of the disease on the genes. Clustering [15] is used to analyze gene expression data to identify genes that behave similarly. In this prototype, we use an information fusion based clustering technique we developed in [6] to represent this family of analysis. This is based on a Self-Organizing Map [16] cluster algorithm to identify clusters of genes simultaneously based on their similarity in expression as well as other information sources. The additional information could be drawn from genomic sequences and clinical readings of the tissue. The algorithm is also capable of weighing the data sources, if needed, in order to produce clusters with greater similarity of genes within one data source when compared to the other data sources. It has been observed in [10] that the clinical behavior of prostate cancer is linked to underlying gene expression differences and hence, such an analysis could lead to better identification of genes that anticipate the clinical behavior of the disease.

3. Architecture

The overall layered architecture of our system is presented in figure 2. The design is aimed at providing a set of grid services to facilitate collaborative sharing of Cancer Analysis Applications (CAA). The main set of services can be divided into two categories: 1. Application Management services and 2. Discovery and Negotiation services. We assume a CaGRID [8] based data sharing environment with each node on the grid considered as a data source for cancer Bioinformatics Infrastructure Objects (CaBIO) data objects. CaBIO is the set of object definitions specified by the NCI and is the standard for biomedical cancer research data. Our system deals with other kinds of data sources by using a set of wrappers to map them into CaBIO specifications. We also assume that the querying and handling of these CaBIO data objects are handled by CaGRID services.

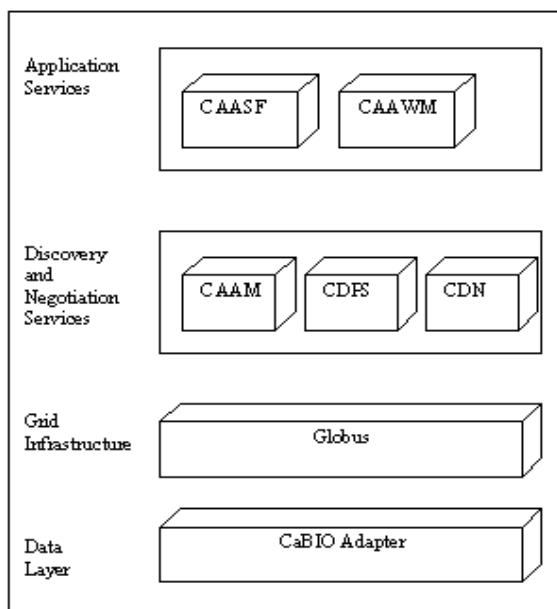


Figure 2. Architecture

3.1. Application Management Services

The Application Management Services are responsible for invocation and management of CAA applications over the grid. The design of these services follows Open Grid Services Architecture-Data Access and Integration (OGSA-DAI) based grid services. The Cancer Analysis Application Service Factory (CAASF) is a persistent service running on each node of the cancer grid. It deals with management of the available CAAs advertised at each host node. This is based on extensions to standard interfaces like the GridServiceFactory of the OGSI specification. When an invocation is received, the CAASF spawns a transient instance of Cancer Analysis Application Service (CAAS) to handle the request. The CAAS handles the interface with the client and other services throughout the lifetime of the invocation. The CAAS in turn invokes the corresponding CAA based on the request received through the service factory.

One of the key functionality requirements for a collaborative analysis environment is pipelining of results. The idea is to direct the results from one analysis application to the other by separating the data preprocessing modules. This helps users to minimize effort and computation when using multiple applications in a pipelined fashion. The CAASF handles pipelining requests through a specialized pipeline specification file. This XML based file specifies the pipeline of applications and the pre/post processing of data required to handle the analysis jobs. Figure 3 shows an example pipeline specification. This example formulates a pipeline request to run analysis applications A1 and A2 on the data source D1. Followed by a pre-processing of the

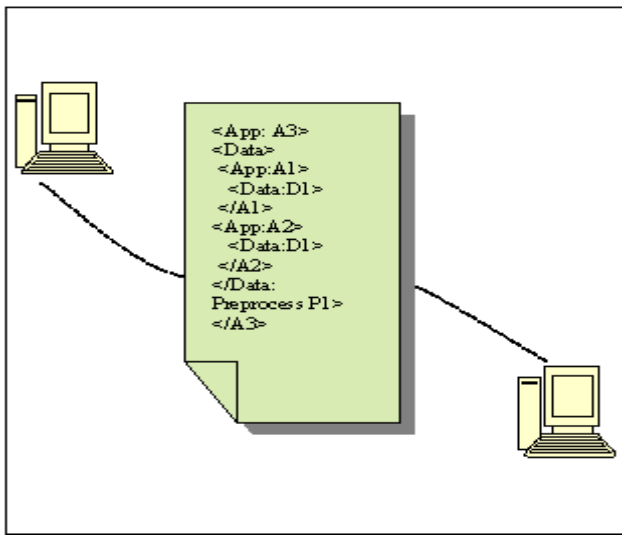


Figure 3. Application Pipelining

results using P1 and finally directing the results as input to application A3.

The Cancer Analysis Application Workflow Manager (CAAWM) is responsible for the workflow monitoring within the system. This consists of a persistent service, Cancer Analysis Application Workflow Manager Service Factory (CAAWMSF) that handles the requests received from CAASF. For each invocation it receives, the CAAWMSF spawns a transient service Cancer Analysis Application Workflow Manager Service (CAAWMS) which is responsible for the execution of the job on the available grid resources.

3.2. Discovery and Negotiation Services

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. This module interacts with the CAASF to maintain the list of available applications and provides an interface for the clients to query the metadata.

A typical cancer analysis application consists of pruning the input data set, setting the input parameters and visualizing the results. In this prototype we do not deal with the visualization of the results. To start with the data search and prune, we assume that the underlying grid-technology (CaGRID) offers a way to identify CaBIO data sources available on the grid. 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 deals with the control of collaborative sharing possible through the grid.

As mentioned in the previous sections, cancer biomedical data is controlled by independently operated hospitals and research centers. In a conventional collaborative sharing environment each node specifies the data sharing boundaries for each class of data it offers based on the sensitivity and importance of the data set. This is done based on a digital specification for each of the shared data objects, indicating whether it is available or not. The CDN helps extend these digital boundaries by allowing two grid nodes to negotiate on the extent of data access allowed for each other. This is implemented using an Access Control List (ACL) based structure for each advertised data object which maintains the data access privileges for all other nodes in the system. A data request would also include a list of non-advertised data sources offered by the invoker along with the standard parameters. In response, the service node can make decisions on sharing its own non-advertised data objects with the invoker.

4. Results

The proposed system was implemented on a prototype grid with basic services to support the design. The grid was loaded with cancer research data collected by the leading cancer research centers in the state of Pennsylvania as part of The Pennsylvania Cancer Alliance for Bioinformatics Consortium (PCABC). The participating members include Penn State Cancer Institute, The Wistar Cancer Institute, Univ of Pittsburgh Cancer Center, Fox Chase Cancer Center. The data includes over 6000 samples of prostate, breast and melanoma cancer related data along with gene expression and publicly available literature.

For the sake of clarity, we only present results obtained for data sets related to prostate cancer. Figure 5 presents a visualization of the results obtained by running a global statistic query on the number of cases diagnosed with prostate cancer with respect to the patient's age-at-diagnosis. The results are further categorized based on the race to which the patient belongs to. We observed that for any age-group, more number of people of Caucasian origin were diagnosed with the disease in comparison to other races. This is not necessarily an indication of the global behavior of the disease since this observation is limited to the data sets considered. Nevertheless, using more grid nodes could help identify such global trends more accurately.

For the second facet of the data analysis as mentioned section 2 we present results obtained by using information fusion based clustering algorithm on yeast gene expression

data and DNA sequence data. The result is shown in figure 4. Clustering of gene expression data using motif factor frequency data derived from sequence information yielded ‘better’ clusters when compared with conventional clustering. We refer the reader to [6] for explanation of the results.

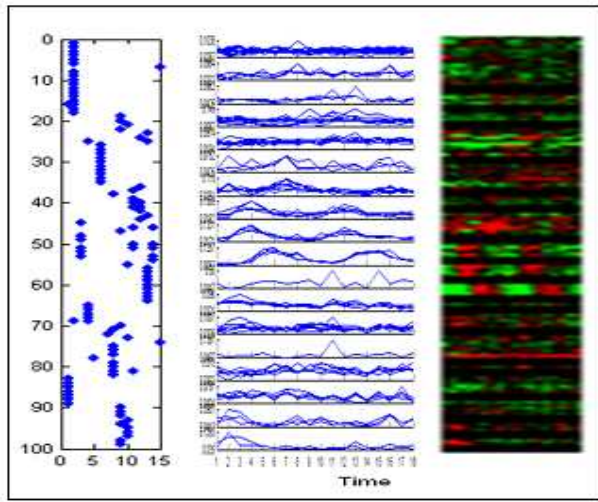


Figure 5. Clustering using gene expression and sequence data

5. Conclusion

In this paper, we presented a prototype grid-based system for cancer biomedical research. The system facilitates collaborative sharing of data and analysis applications over a domain specific grid. Grid-Based biomedical informatics applications comprise of a combination of tasks that need high computational power and handle high data distribution. We identified and proposed solutions for some domain specific problems related the grid computational environment for biomedical analysis applications. We believe that such problems need to be dealt with to make specialized grid systems as the next generation computational platforms for science. We are currently working on methodologies to evaluate the system performance. Our future work includes deploying the system on the NCI Cancer Grid infrastructure.

References

- [1] V. Boccia, M. R. Guarracino, L. D’Amore and G. Lacetti. “A Grid Enabled PSE for Medical Imaging: Experiences on MediGrid”. *In the Proceedings IEEE Symposium for Computer Based Medical Systems*, Dublin, Ireland, June 23-25 2002, pp. 529-536.
- [2] J. Bunn and H. Newman. *Data-intensive grids for high-energy physics. John Wiley and Sons, Inc, New York, 2003.*
- [3] M. Cannataro, C. Comito, F. Schiavo and P. Veltri. “Proteus, a Grid based Problem Solving Environment for Bioinformatics: Architecture and Experiments”. *In the Proceedings IEEE Computational Intelligence Bulletin.*, Portland, OR, May 3-10 2003, pp. 125-137.
- [4] 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.
- [5] P. V. Jithesh et al. “GeneGrid: Grid Based Solution for Bioinformatics Application Integration and Experiment Execution”. *In the Proceedings IEEE Symposium for Computer Based Medical Systems*, Dublin, Ireland, June 23-25 2002, pp. 523-528.
- [6] J. Kasturi and R. Acharya. “Clustering of Diverse Genomic Data using Information Fusion”. *Bioinformatics Journal*, 21(4): 423-429, 2005.
- [7] Y. Liu et al. “Grid-BLAST:Building A Cyberinfrastructure for Large-scale Comparative Genomics Research”. *in Proceedings Cancer Cell*, March 2002, pp.1(2):203-9.
- [8] W. Sanchez, B. Gilman, M. Kher, S. Lagau and P. Covitz. “CaGRID White Paper”. <https://cabig.nci.nih.gov/guidelines/documentation/CaGRIDWhitepaper.pdf>, July 23 2004.
- [9] A. Simpson, D. Power, M. Slaymaker, and E. Politou. “GIMI: Generic Infrastructure for Medical Informatics”. *In the Proceedings IEEE Symposium for Computer Based Medical Systems*, Dublin, Ireland, June 23-25 2002, pp. 564-566.
- [10] D. Singh et al. “Gene Expression correlates of clinical prostate cancer behavior”. *In Proceedings Cancer Cell*, March 2002, pp.1(2):203-9.
- [11] W.T.Sullivan III et al. “A new major SETI project based on Project Serendip data and 100,000 personal computers”. *In Proceedings of the Fifth International Conference on Bioastronomy*, Capri, Italy, 1997.
- [12] Y. Teo, X. Wang and Y. NG. “GLAD: a system for developing and deploying large-scale bioinformatics grid”. *Bioinformatics*, 2004.

- [13] E. Domany. "Cluster Analysis of Gene Expression Data". *Journal of Statistical Physics*, Vol. 110, Nos. 36, March 2003.
- [14] T. Kohonen. "Self-organizing maps". *Springer*, Berlin, 1995.
- [15] NC BioGrid. <http://www.ncbiogrid.org/>.
- [16] Grid BLAST. <http://keck1.biotec.uiuc.edu:8080/gridblast/index.html>.

CASE_ID by RACE and AGE

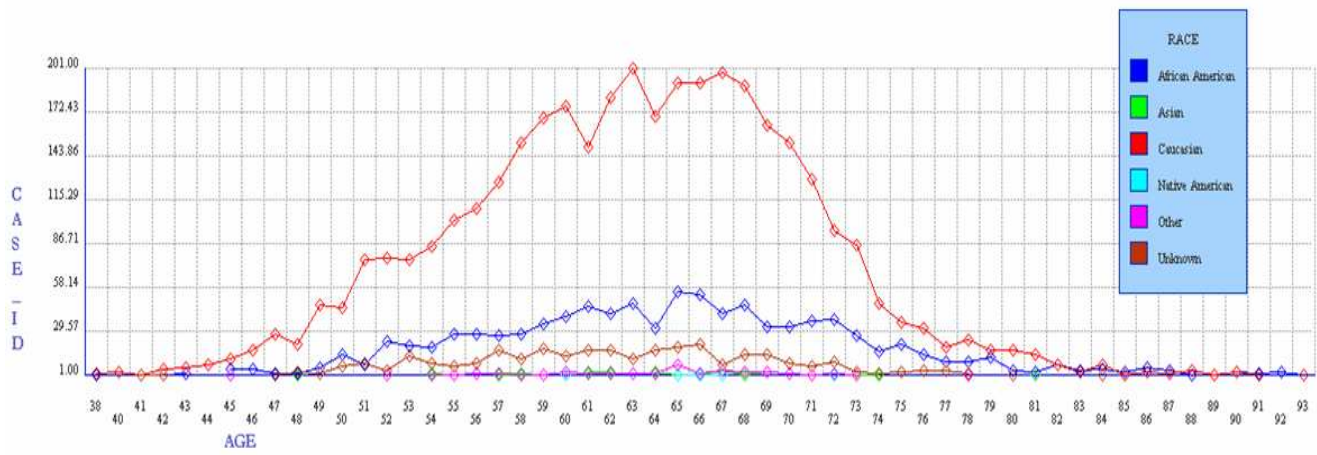


Figure 4. Sample Disease Demographics Result : Number of patients diagnosed with respect to age at diagnosis