

An Online Analysis and Information Fusion Platform for Heterogeneous Biomedical Informatics Data

Srivatsava Ranjit Ganta¹, Jyotsna Kasturi¹, John Gilbertson², Raj Acharya¹
¹Dept of Computer Science and Engineering, Pennsylvania State University,
²University of Pittsburgh Cancer Institute
Email: {ranjit, jkasturi, acharya}@cse.psu.edu, GilbertsonJR@upmc.edu

Abstract

Current research in biomedical informatics involves analysis of multiple heterogeneous data sets. This includes patient demographics, clinical and pathology data, treatment history, patient outcomes as well as gene expression, DNA sequences and other information sources such as gene ontologies. Analysis of these data sets could lead to better disease diagnosis, prognosis, treatment and drug discovery. However, the extent of knowledge that can be extracted from individual data sets is limited. Recently, there has been an initiation on techniques that analyze genomic data sources in an integrated manner through information fusion. This places a need for an online platform to analyze biomedical informatics data sets using these techniques. We present here a preliminary report on an online data warehouse to perform data exploration and analysis across heterogeneous biomedical informatics data sets with the aid of information fusion. The prototype platform is available at <http://biogeowarehouse.cse.psu.edu>.

1. Introduction

Biomedical informatics in generic terms deals with the study and analysis of clinical and biological data leading to better disease diagnosis, treatment and drug development. This involves information sources like patient demographics, tissue information, gene expression, sequence and other sources like the gene ontologies. The goal is to discover hidden trends and patterns that could be used to verify more complex hypothesis.

These information sources are highly heterogeneous and are usually analyzed independently. However, the amount and richness of the knowledge that can be extracted from each of these information sources separately is limited. This is because of the fact that they are related in an intrinsic way from a global point of view. For example, tissues from patients are used as the source for a set of genes in a microarray experiment. Using the clinical pathology information of these tissues in the analysis of the microarray might be useful to find hidden trends that are otherwise undetectable.

In recent times, techniques such as [3] have been proposed to signify the use of multiple data sources to perform information fusion based analysis of data sets. The basic idea behind these techniques is to use multiple data sets simultaneously to extract knowledge. This places a need for a platform to facilitate such analysis through an explorative means for biomedical informatics data sets.

In this paper, we present a preliminary report on an online data warehouse to perform fusion based data visualization and analysis across heterogeneous biomedical informatics data sets through information fusion. We use prostate cancer data to demonstrate the functionalities of the system. The rest of the paper is organized as follows: Section 2 gives an overview of data warehousing and the techniques employed to demonstrate information fusion on biomedical informatics data. Section 3 gives an

outline of the system architecture. Section 4 presents snapshots of results obtained for some example queries. The last section provides a conclusion followed by the future work.

2. Background

Data warehousing is a technology that provides a database platform for analytical purposes. The basic idea is to pool in multiple data sets onto one database platform that serves analysis queries. Our platform is an online data warehouse that provides tools for performing fusion based data analysis and visualization. It provides two main functionalities: 1. A platform that serves as a one-point access to biomedical informatics data sets related to various diseases 2. An environment for data visualization and analysis with the aid of information fusion. On this platform, we demonstrate information-fusion using two techniques 1. Multidimensional Visualization, and, 2. Fusion Based Clustering. The rest of this section gives a brief overview of these techniques.

2.1. Multidimensional Visualization

Multidimensional Visualization is based on the data cube model introduced by Gray et al [2] for running analysis queries on databases. Several logical models like [1] have been proposed formalizing the same concept in various frameworks. The basic idea is that information can be thought of as an n -dimensional ‘cube’ with some attributes as ‘dimensions’ and some as ‘facts’. Using a data cube, the ‘facts’ are visualized along the ‘dimensions’. There has been some earlier work in Tao et. al. [5] for simultaneous visualization of genotypic and phenotypic datasets. We use a similar approach and a novel factor in our approach is defining the ‘cube’ over multiple data sets and applying it to visualize both patient demographic and genomic data such as gene expression information.

In our approach, the queries are formulated by first selecting the subset of the information sources to be considered for the analysis. Once this done, the user selects a certain attribute(fact) from one of the data sets as the focus of analysis and some attributes(dimensions) along which the fact needs to be analyzed. Based on this selection, the system presents the user with an initial visualization of the corresponding information cube. The user is then allowed to further explore this by performing certain *operations*:

Summarize: This operation takes in the current information visualization and summarizes it based on a hierarchy defined over the dimension. A hierarchy is a tree based grouping of all possible values for a given dimension. For example, gene ontology can be looked as a hierarchy defined on genes based on the cellular functions they belong to. Using this operation the user can view the fact values at different levels of dimension values.

Detail: This operation is the transpose of summarize. It presents the user with a detailed view of the cube. In effect, it is used to go down the hierarchy defined of the dimension.

2.2. Fusion Based Clustering

Gene clustering is the most widely used techniques in bioinformatics to understand gene functions. However, current work on this solely use gene expression data obtained from microarray experiments as the data source. The question is whether one can use other datasets

to cluster genes with similar function. We developed a simple clustering mechanism to cluster genes from both gene expression as well as their motif frequency information. The goal of this technique is to allow clustering of multiple heterogeneous data together. The basic idea is to extend a simple clustering algorithm such that multiple data sets are taken into consideration. Due to space constraints we only provide an overview of our approach. The key challenges involved are 1. Determining the extent of influence of each data set on the clustering 2. Measuring the distance between cluster centroid and the data vectors 3. Mechanism to adjust the centroids in each iteration. Our algorithm extends the basic Self-Organizing Map [6] and starts by initializing the cluster centroids randomly. Based on the user input, the algorithm picks one of the data sets for each iteration and adjusts the centroids based on the distance measure corresponding to that data set. This is done until there is little change in the position of the centroids. The platform provides this as a tool and enables clustering on the chosen data sets and parameters. Clustering multiple data sets leads to better tighter clustering and biomedical studies such as [4] observe 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. System Architecture

Figure 1 presents our system architecture. The system is built based on the standard 3-tier architecture: the presentation layer, the application logic and the database layer. This design offers the flexibility and resistance to future additions. The presentation layer essentially deals

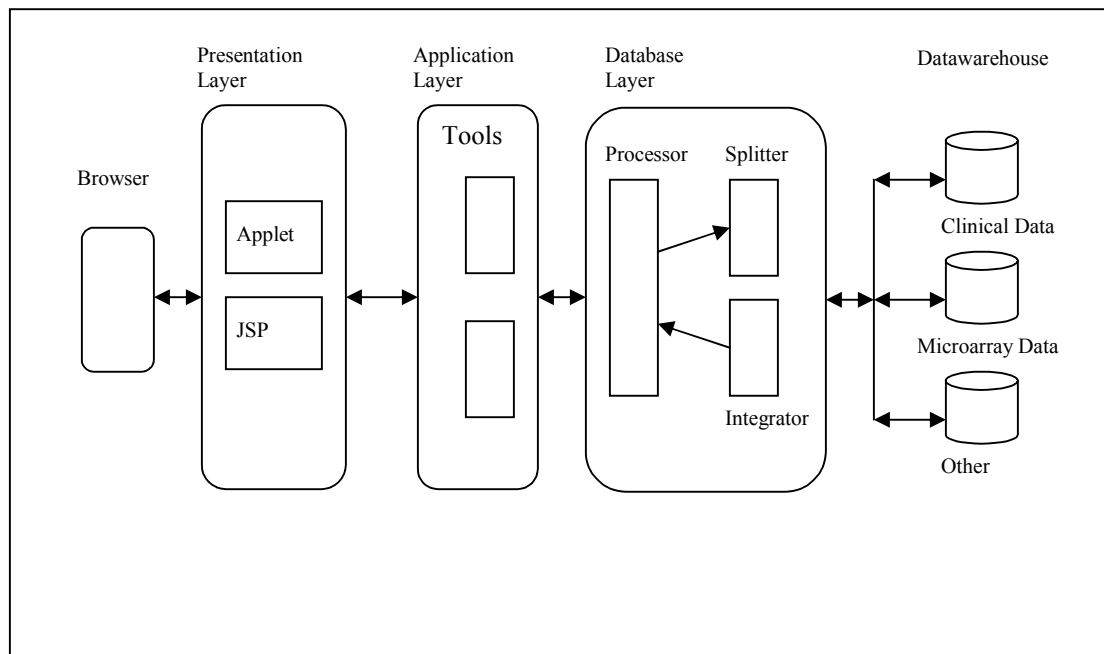


Figure 1: System Architecture

with the presentation of results from various information fusion based techniques. It is currently implemented using java applets along with jsps to provide the required visualization functionalities. The application logic layer deals with the functionalities of the various tools offered through the system. It implements various information fusion based algorithms using the data access functions offered by the database layer. The database layer takes care of the

necessary query processing capabilities required for integration from heterogeneous data sets stored on the data warehouse. The processor accomplishes this by sending in pre-processed queries to the splitter which directs the data requests to appropriate section of the data warehouse. The data integrator collects the data and processes it to cross link each of the data items obtained and submits it to the processing engine. The final results are submitted to the presentation layer for visualization.

4. Results

In this section, we present the functionalities of the system using prostate cancer data sets. Prostate cancer is the most commonly diagnosed non-skin cancer in the United States with an estimated 198,100 new cases and 31,500 deaths in 2001 [5]. One in six American men develops prostate cancer in the course of their lifetime. Data collection starts at the diagnosis stage with the demographic information such as patient's age, sex, and the family history of prostate and other cancers. Study of this information set gives a global view of the trends in various disease related attributes and insights into spread of the disease in different races and geographic locations. Examples of some generic queries on this data set are:

1. How many number of patients under the age of 50 were diagnosed with prostate cancer?
2. What percentage of the patients recorded belong to the African American community?

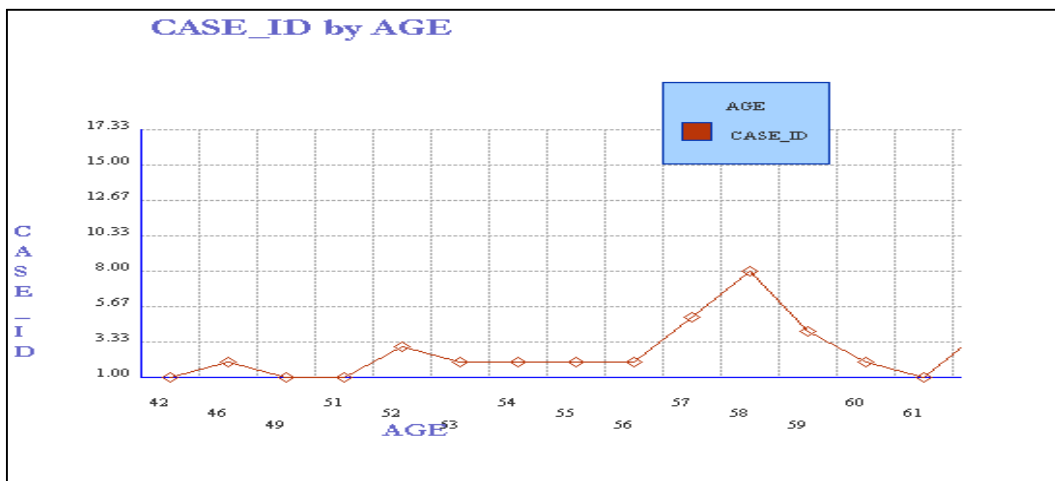


Figure 2: Number of patients with specific age diagnosed with prostate cancer

Figure 2 gives a snapshot of the result obtained for query 1 3 by using the Multidimensional Visualization tool. Figure 3 shows the same result after executing the

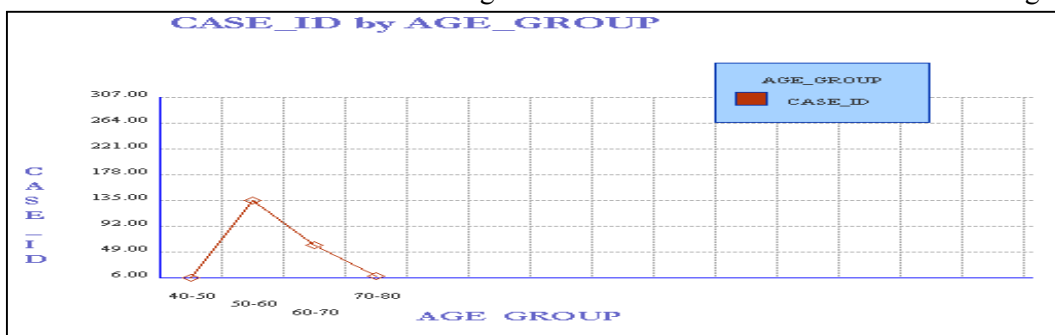


Figure 3: Number of patients belonging to various age groups diagnosed with prostate cancer

summarize operation by defining a hierarchy on the age dimension by grouping them into various risk groups. It can be observed those patients belonging to the age-group 50-60 are at maximum risk of the disease. During the diagnostic process pathologic studies are done to categorize the type and aggressiveness of the tumor (histologic type and grade) as well as the extent of tumor spread (stage). In most cases the disease is cured by the initial therapy, however, in some cases the tumor recurs. Documentation of the time and extent of this recurrence is important in better understanding the biologic nature of aggressive tumors. Example queries on these data sets are:

3. What are the average Gleason-Grade values observed in patients categorized with different Pathologic T-stages?

4. What is the average tumor size for cases with Gleason Grades of greater than 3?

Figure 4 gives a snapshot of the results obtained for query 3. Gene expression and sequence studies are done on some of the diagnostic tissue samples. Understanding the relationships between initial grade and stage, therapy and recurrence (and outcome) and

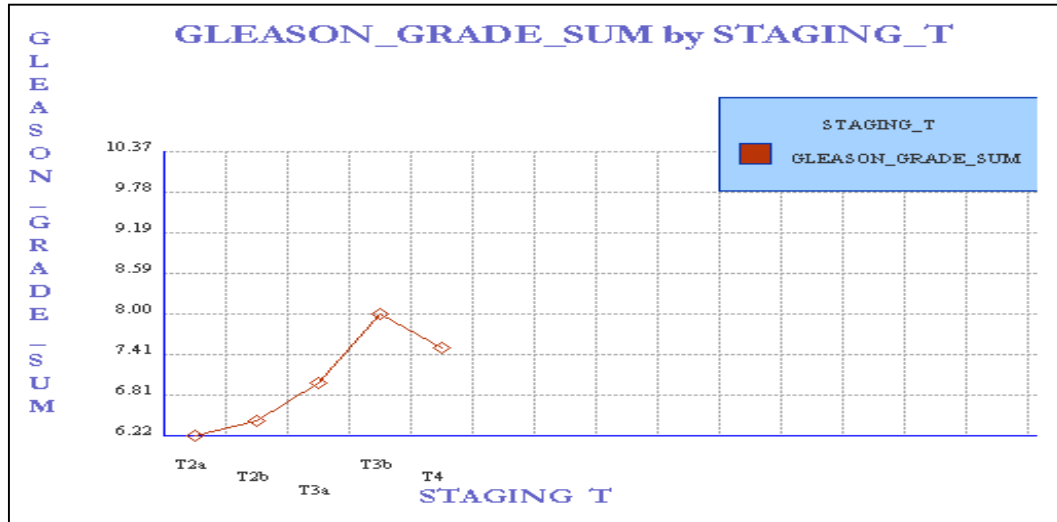


Figure 4: Average Gleason-Grade-Sum values observed in cases belonging to various T-Stagings.

how this correlates with gene expression and sequence studies is a major area of biomedical informatics research. Example queries are:

5. Output the average expression vector for patients with a certain PSA value.

6. Based on certain ontology for cellular functions, give me the average expression vector of all the genes corresponding to a particular cellular function.

Figure 5 gives a snapshot of the results obtained for query 5. The results of the above mentioned queries offer an explorative method to visualize multiple heterogeneous data sets.

To demonstrate the fusion based clustering technique, we used motif frequency data as the additional data source to cluster gene expression data from [4]. Figure 6 gives a snapshot of the result obtained. The result revealed that tighter gene clusters were obtained through fusion based clustering and also that one of the gene clusters obtained was biologically proven to have similar gene function. Due to space constraints, we leave further description of the result for future work.

5. Conclusion

We present here a fusion based analytical and visualization platform for biomedical informatics. Though the presentation focused on issues related to prostate cancer, the principles can be easily extended to other disease related data sets. The idea of explorative analysis using information fusion could serve as a preprocessing step for verifying other complex hypotheses. Our future work involves extending the system to a distributed environment and automated discovery of interesting patterns. This project is funded, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds.

7. References

- [1] R. Agrawal, A. Gupta, S. Sarawagi. "Modeling Multidimensional Databases", Proc of 13th Intl Conf on Data Engineering, ICDE 1995.
- [2] J. Gray et al. "Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross- Tab, and Sub Totals," *Data Mining and Knowledge Discovery*, 1(1), 1997, pp. 29-53.
- [3] I. Holmes, W.J. Bruno, "Finding Regulatory Elements Using Joint Likelihoods for Sequence and Expression Profile Data," *ISMB*, 2000, pp.202-210.
- [4] D. Singh et al. "Gene Expression correlates of clinical prostate cancer behavior," *Cancer Cell*, March 2002, pp.1(2):203-9.
- [5] Y. Tao, C. Friedman, Y.A. Lussier, "Visualizing Information across Multidimensional Post-Genomic Structured and Textual Databases," *Bioinformatics*, Dec 2004.
- [6] Kohonen, T. Self-organizing maps. Springer, Berlin, 1995.

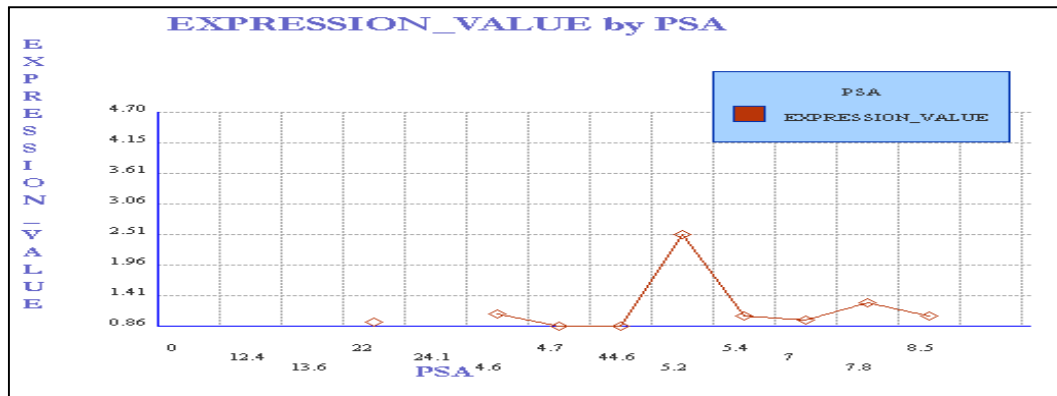


Figure 5: Average expression value of a gene with respect to the PSA value of the corresponding tissue

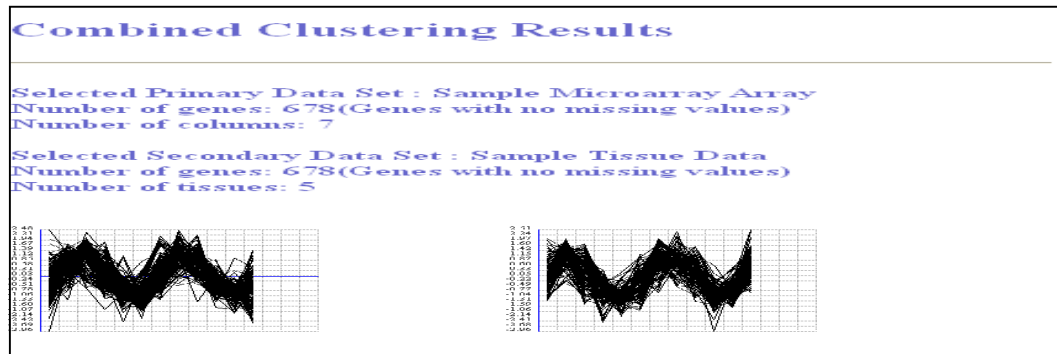


Figure 6: Clustering typical Microarray data with sample Gleason Sum Score values as the additional data set