

Fusion Based Analytical and Visualization Techniques for Biomedical Informatics

Srivatsava Ranjit Ganta¹, Jyotsna Kasturi² PhD, Raj Acharya¹ PhD

¹Department of Computer Science and Engineering
Pennsylvania State University
University Park, PA 16802

²Non-Clinical Biostatistics,
J&J PRD, 1000 Rt. 202 S,
Raritan, NJ 08869

Corresponding Author:
Srivatsava Ranjit Ganta
ranjit@cse.psu.edu
337 IST, University Park, PA 16802
Phone: 814-863-4254
Fax: 814-865-3176

Abstract:

Current research in biomedical informatics involves the study and analysis of multiple heterogeneous datasets. This includes disparate information sources such as patient demographics, clinical and pathology data, treatment history, patient outcomes as well as gene expression, DNA sequences etc. Knowledge discovery, and data visualization studies on all these information sources is important in finding better solutions for Biomedical research. However, the extent of knowledge that can be extracted from individual datasets is limited. The successes of recent studies that analyze biomedical data in an integrated manner suggest that use information fusion based techniques to extract rich knowledge from these datasets. In addition, the advent of nationwide grid infrastructures such as CaBIG (Cancer Bioinformatics Grid) allows researchers to share disparate biomedical data and furthers the call for information fusion based analytical and visualization tools. In this paper, we present an online data warehouse platform to perform fusion based data analysis and visualization across heterogeneous biomedical informatics datasets. Our system is aimed at researching existing methodologies to study such information sources and extend them to operate on multiple datasets and make them publicly available as a toolset on the data warehouse. The prototype system is available at <http://biogeowarehouse.cse.psu.edu>.

I. INTRODUCTION

BIOMEDICAL informatics deals with the study and analysis of highly heterogeneous clinical and genomic data. Clinical data consists of patient demographics, diagnosis information, treatment records and follow-up data while Genomic data consists of more complex data such as gene expression, DNA sequences etc. Study and analysis of these information sources could lead to better disease diagnosis approaches, treatment procedures and drug development. These data sources are distributed among various hospitals, research centers and government agencies that are controlled independently. Hence, researchers have so far been limited to islands of data and informatics tools. This scenario poses a serious challenge for studies that aim to research diseases from a global point of view. The heavy distribution of data and tools impose restrictions on the extent of data analysis and visualization possible. However, with the advent of nationwide grid systems such as CaBIG (Cancer Bioinformatics Grid) [1], sharing data and tools across organizational and even international boundaries has been made feasible. Such infrastructures facilitate a comprehensive collaborative research platform wherein researchers have access to multiple datasets and share analysis tools. This scenario provides an opportunity for integrated studies that take more information sources into consideration than so far.

Data analysis in biomedical informatics is usually carried out using techniques such as clustering, analytical querying, visualization etc. However, these studies mainly involve exclusively either only clinical or genomic datasets. For example, application of clustering techniques has by far been limited to gene expression data. Recent initiation by studies such as one done by Holmes et. al.[6], Tao et. al.[11] indicate that this traditional

approach is limited in the knowledge extraction process. This indicates the need for *Information Fusion* based approaches to data analysis and visualization. ‘Information Fusion’ involves *fuzing* multiple, heterogeneous information sources to perform knowledge extraction. The idea is that the knowledge that could be extracted from multiple information sources put together yields insights which are otherwise not possible.

In this paper, we present an online data warehouse platform to perform fusion based data analysis and visualization across heterogeneous biomedical informatics datasets. Figure 1 gives an overall view of the goal. The platform is aimed at providing two main functionalities: 1. A data warehouse that serves as a one-point access for cancer biomedical informatics data, 2. An information fusion environment that provides a suite of tools for the users. We present the techniques we developed to incorporate clinical as well as genomic datasets in data analysis and visualization. We also demonstrate the toolkit functionalities using prostate cancer data collected from the leading cancer research centers in the state of Pennsylvania as part of The Pennsylvania Cancer Alliance for Bioinformatics Consortium (PCABC) [7] in addition to some publicly available datasets. The rest of the paper is organized as follows: Section II introduces the currently used methodologies for performing data visualization and analysis and motivates the need for information fusion in such techniques. Section III provides a brief background on data warehousing and the system architecture. This is followed by the suite of techniques/tools available on the platform: 1. Multidimensional Visualization, 2. Correspondence Analysis, 3. Fusion Based Clustering. Section IV provides the summary and conclusion.

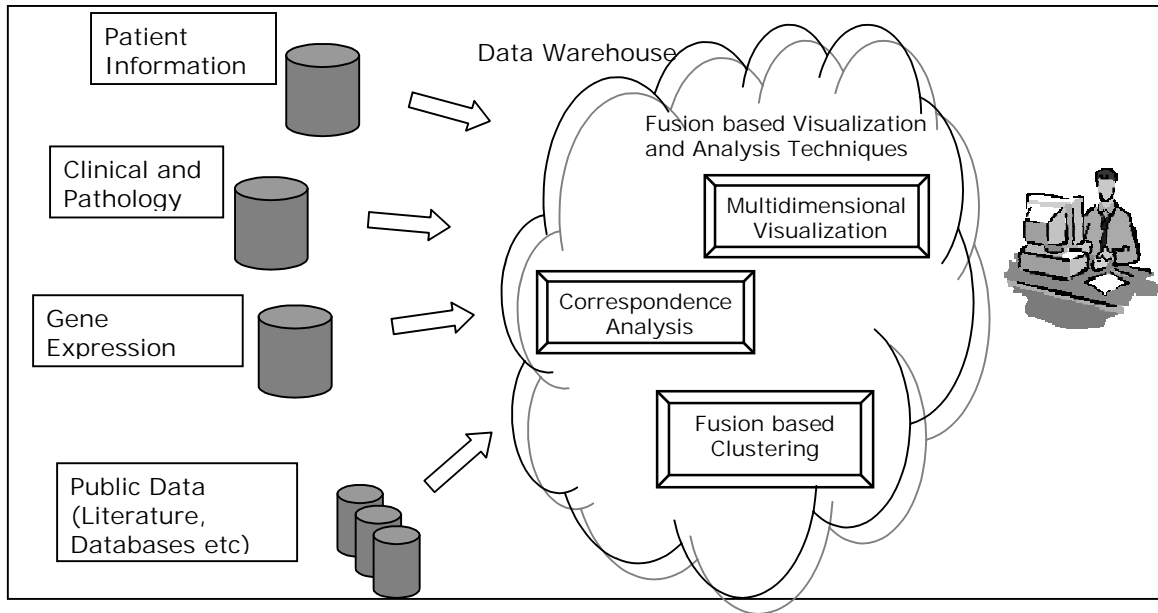


Figure 1. Platform Overview

II. MOTIVATION

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 [8]. One in six American men develops prostate cancer in the course of their lifetime. Biomedical data collection starts with patient demographics data which comprises of the patient's age, sex, family history of cancer etc. Visualization of this information gives a global view of the trends in various disease related attributes and insights into the spread of the disease among different races and geographic locations [11]. Some example queries are:

1. **How does the age-at-diagnosis for prostate cancer patients vary with their race?**
2. **What percentage of the diagnosed patients recorded to have a family history of prostate cancer?**

While the above mentioned scenario deal exclusively with clinical information, the studies can be extended to genomic datasets including gene expression and sequence

data. Note that such queries span multiple data sources via indirect linkage of records.

Example queries are:

3. **How does the gene expression vector for patients vary with their Gleason score?**
4. **Based on certain ontology for cellular functions, what is the average expression vector of all the genes corresponding to a particular cellular function?**

The understanding and interpretation of these associations helps identify disease hot-spots, geographical-spread patterns and other global insights [2][5]. The goal behind the “Multidimensional Visualization” tool is a system that facilitates the formulation of such queries across multiple datasets and use data visualization to interpret the result. The tool also provides a mechanism to further explore the basic visualization obtained through various operations that can be performed on the visualization. Section IIIA describes the Multidimensional Visualization technique and presents some sample results and interpretation.

Clinical and genomic data sources *correspond* to each other in different ways because of the inherent linkage among them. For example, clinical data of certain diseases indicate that patients from certain races have lesser age-at-diagnosis than when compared to others. In this sense, the average-age-of-diagnosis of the patient corresponds to the race the patient belongs to. This may indicate that people of certain races may be prone to a disease much earlier than people of other races. Simple correspondence such as this among multiple data sources could be detected through multiple queries (finding average age-at-diagnosis for datasets from different hospitals and comparing these with the race of the patient). However, other complex correspondence relationships among indirectly linked data cannot be detected easily through querying. Thus, there is a need for visual

exploratory technique to help identify and quantify such relationships. The “Correspondence Analysis” tool aims to help researchers detect correspondence relationships among various biomedical informatics datasets through a novel application of a visual exploratory technique proposed by Greenacre et.al. [13]. The functionalities and usage of the tool is discussed in Section IIIB.

Another key facet of biomedical informatics studies involves mining gene expression data to identify disease biomarker genes [8][12]. This is usually carried out using clustering techniques such as Hierarchical, K-means etc [4]. However, the application of clustering techniques so far has been limited to gene expression data. Recent work by Holmes et. al. [6] motivates the use of multiple datasets through to obtain better clustering results. This suggests the inclusion of additional information sources such as DNA sequence data and clinical data to cluster genes with similar function. We developed a simple fusion based clustering algorithm to cluster multiple datasets. The technique is based on Self Organizing Map (SOM) clustering technique and extends it in several ways to incorporate multiple information sources. This technique is made available on the system as “Fusion based Clustering” tool and provides a means to cluster gene expression data along with other data sources such as DNA sequence based information. Section IIIC elaborates on the technique and presents some sample results.

With this motivation, we now proceed to present an overview of our system followed by the techniques/tools made available on the platform.

III. PLATFORM

In this section, we first present a brief background on data warehousing and then our

system architecture. “Data warehousing” encompasses the architectures, algorithms and tools for bringing together selected data from multiple databases and information sources. Traditionally analytical queries on clinical and genomic data are run done by the *lazy* or *on-demand* approach which involves a two step process: 1. Accept a query, determine the appropriate set of information sources to answer the query and then fire the sub-queries to corresponding data sources. 2. Retrieve the results back from each repository and compute the final answer for the user. The disadvantage of this approach is that data is not retrieved until a query is fired and involves some delay. Datawarehousing involves the alternative approach of prefetching the data so that specific analysis queries can be answered in an optimized way. This approach yields better results than the *on-demand* approach when the system is targeted at specific data analysis and exploration operations.

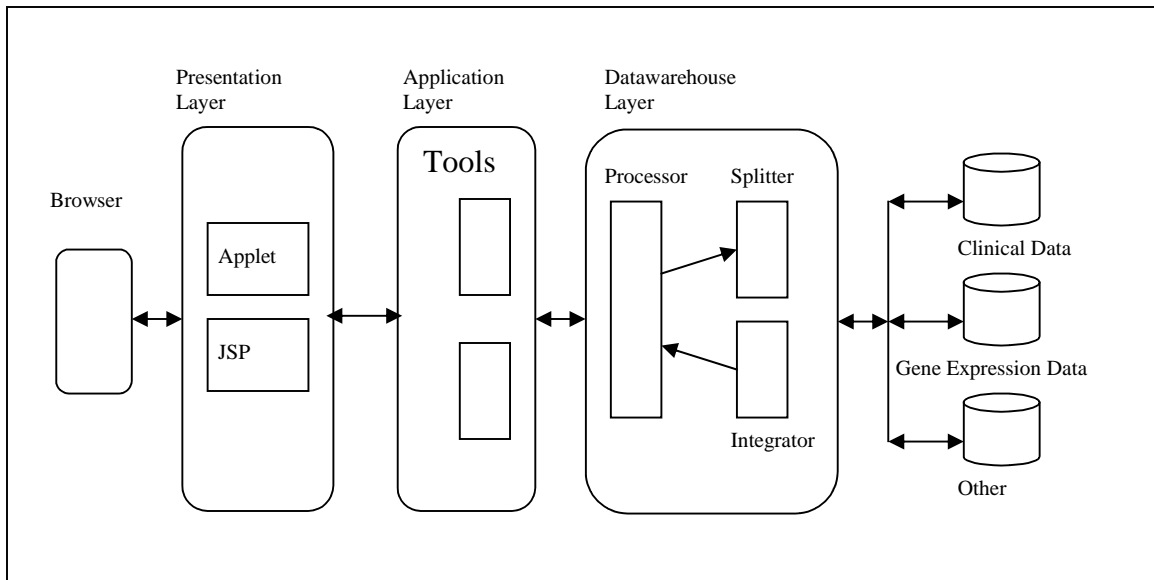


Figure 2. System Architecture.

Our system architecture is based on a 3-tier design and is presented in Figure 2. It consists of three layers: The Presentation layer, the Application layer and the Datawarehouse layer. The presentation layer essentially deals 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 implementation of the various tools offered through the system. The datawarehouse layer takes care of the necessary query processing capabilities required for integration from heterogeneous datasets stored on the data warehouse. The Processor module 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. This design offers the flexibility and robustness required for extending the functionalities offered on the system.

The rest of the section presents the information fusion tools made available on the platform. We use prostate cancer related datasets collected from the leading cancer research centers in the state of Pennsylvania as part of Pennsylvania Cancer Alliance for Bioinformatics Consortium (PCABC) and some publicly available data to demonstrate the functionalities of the system.

A. *Multidimensional Visualization*

The ‘Multidimensional Visualization’ tool facilitates the visualization and exploration of *associations* among multiple biomedical datasets. As motivated in Section II, the goal is to design a system that helps formulate association queries and present a visualization of the result. Our solution for this is based on the multidimensional data cube model proposed by Gray et. al. [9]. The basic idea is to model the query using *facts* and *dimensions* as the building blocks. The *facts* are the ‘data-of-interest’ i.e. the data to be analyzed with respect to the dimensions. A data cube can has multiple attributes as facts for the same set of dimensions. Each of the dimensions and facts belong to specific domains. One can roughly define the cube as a set of tuples of the form: $\{(d_1, d_2, \dots, d_n, f_1, f_2, \dots, f_m), \dots\}$

Where d_i is the dimension value and f_i is the fact value and there are n dimensions and m fact values in each tuple. Each dimension d_i belongs to the domain D_i .

Consider the query, “Visualize how the number of the prostate cancer patients vary with the Race of the patient and Age-Group of the patient”. In this case, the fact is “Number of Patients” and the dimensions are “Race” and “Age-at-diagnosis” of the patient. Consider another such query spanning gene expression and tissue data, “Visualize how gene expression for certain genes vary across tissue sample taken at

different disease stages”. In this case, the fact is “Gene Expression Value” and the dimensions are “Genes” and “Tissue Type”. Figure 3 illustrates the conceptual data cube for this query.

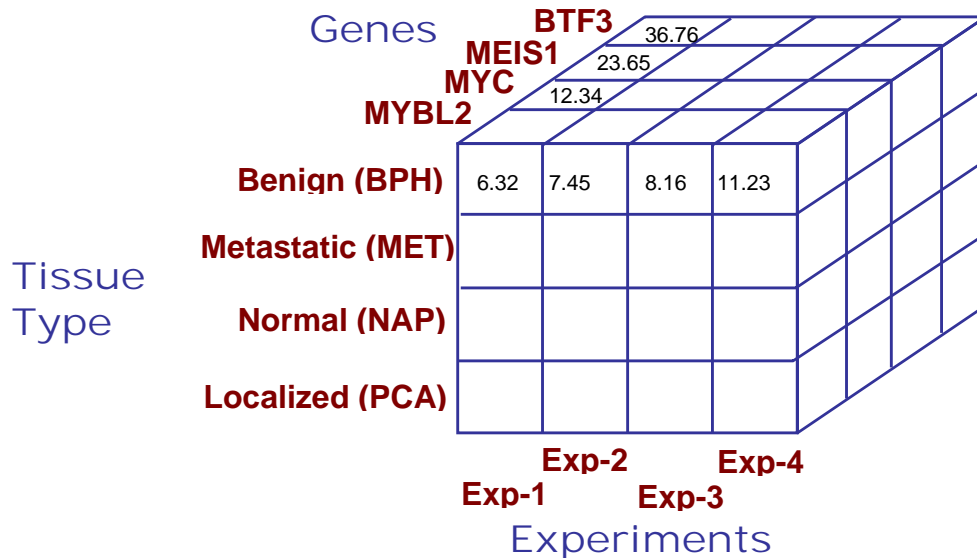
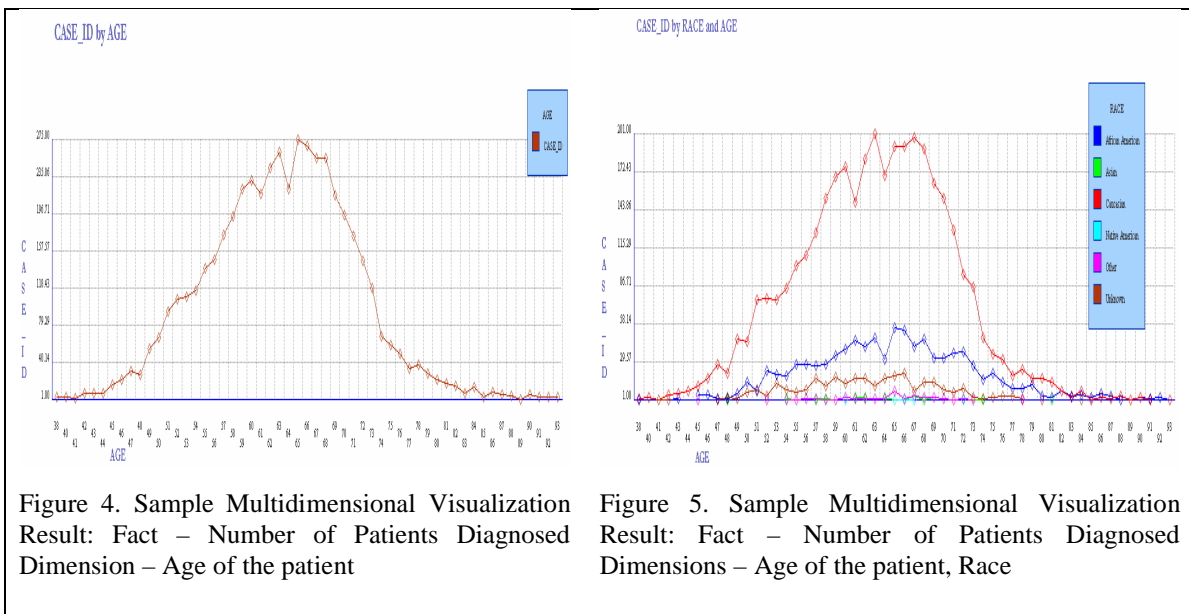


Figure 3. Multidimensional Data Cube

The tool consists of a query formulation module that helps the user formulate an input query and a visualization module that helps in the visual interpretation of the result. The queries are formulated by first selecting the subset of the information sources to be considered for the analysis. Once this is done, the user selects one attribute of the data as the fact from one of the subsets as the focus of analysis. The user then selects the attributes that are to be considered as dimensions. For this, the user can select from 1 to 3 attributes (the maximum number of dimensions supported is 3). The system then presents an initial visualization of the fact with respect to the dimensions. Figure 4 presents a snapshot of the result obtained for the above example query depicting how the number of prostate cancer patients varies with respect to the age of the patient. It can be observed that the graph peaks for the age values 63-67 and thus the user can conclude that most

prostate cancer diagnosis happens in this age group. At this point, the example uses only one dimension i.e. “Age” of the patient. The system allows the user to add more dimensions to the current visualization. Figure 5 depicts the result obtained by adding the dimension “Race” to the result obtained earlier in Figure 4.

The tool also facilitates a way to further explore this *basic* visualization. Looking at the same piece of information at different granularity levels is usually very useful in exploring the data. The tool facilitates this by allowing the user to perform certain *operations* on the visualization result. Our prototype system supports two main operations:



Summarize: The “Summarize” operation takes in the current information view and *summarizes* it based on a *value-hierarchy* defined over the dimension attribute. A hierarchy is a tree based grouping of all possible values for a given dimension into multiple levels. For example, gene ontology is a hierarchy defined on genes based on the cellular functions they belong to. The user can select either *pre-defined* hierarchies for

each of the dimension available on the system or choose to define specific hierarchies in a specific format. Using this operation the user can visualize the *fact* by going “up” the levels of dimension value-hierarchy. Figure 6 shows the conceptual view of summarization for the basic conceptual model shown in Figure 3. The genes dimension is summarized such that the genes are grouped into “Under-Expressed” and “Over-Expressed” categories. This helps visualize the difference in gene expression values between under-expressed and over-expressed genes. The operation uses simple aggregation functions such as Average, Maximum, Minimum etc.

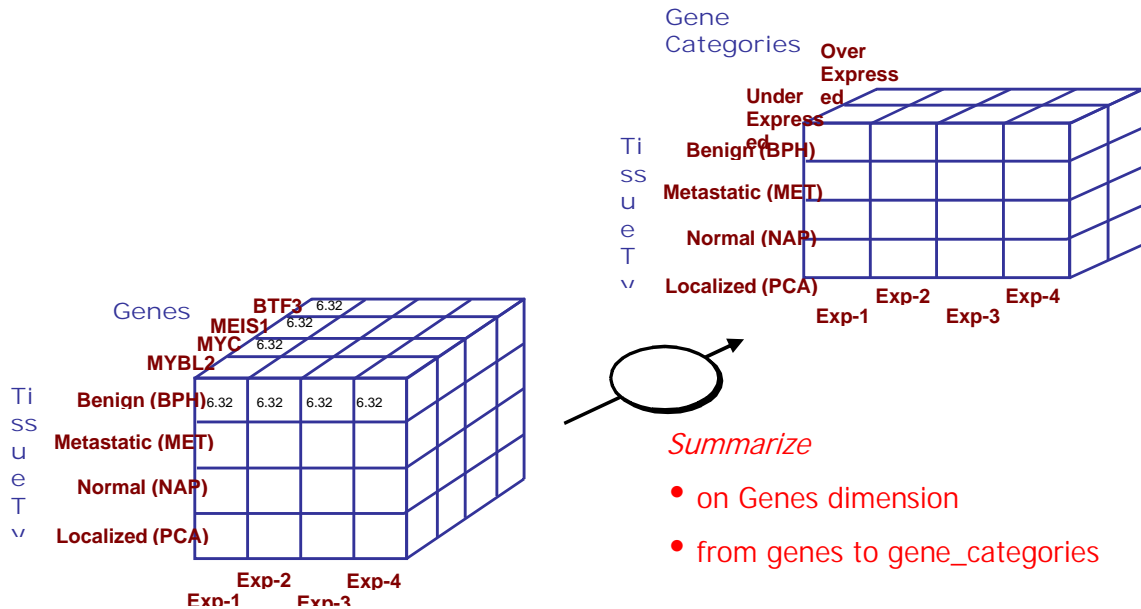


Figure 6. Summarize Operation

Detail: The “Detail” operation is the complement of the summarize operator. It is used to go *down* the levels of dimension hierarchy, thus providing a more detailed view of the data. Figure 7 depicts the conceptual view of the operation for the basic conceptual model shown in Figure 3.

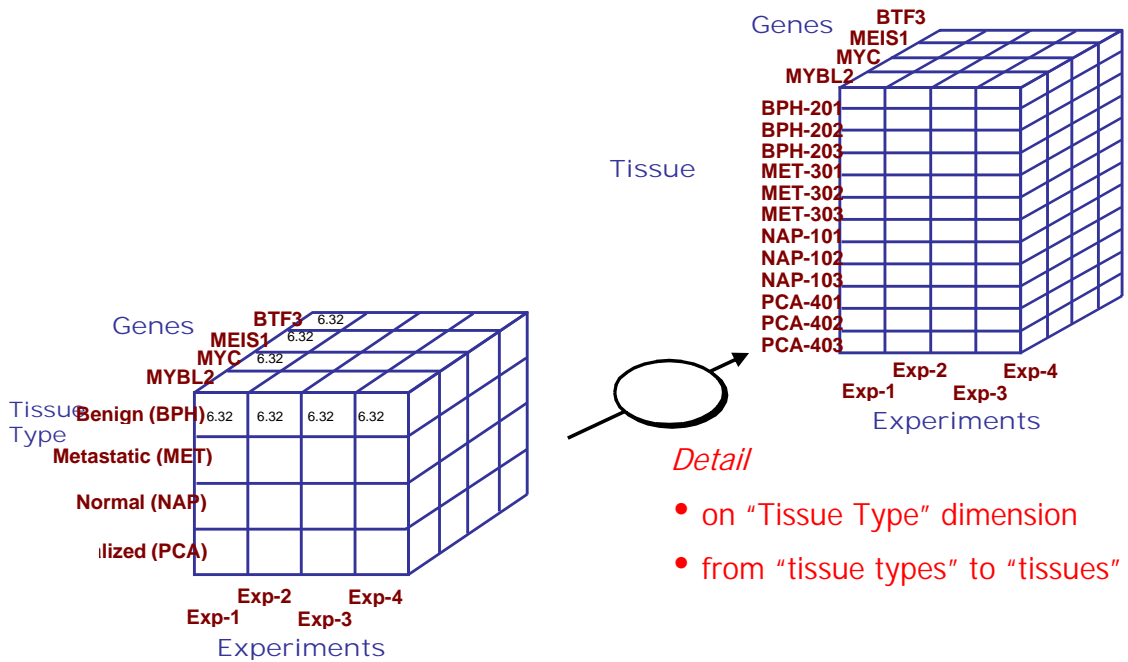


Figure 7. Detail Operation

The example query used so far in this section involves only patient demographics data. Association queries on multiple datasets can be posed in a similar manner by choosing one data set as the fact and other data set(s) as the dimensions. Fig. 8 shows a sample result that depicts the gene expression values of a set of genes along the clinical stages of the tissues from which the samples are taken in a study [12] involving the identification of biomarkers for prostate cancer. The clinical stages are coded as BPH- Benign, NAP- Normal Adjacent, PCA- Localized and MET- Metastatic. In the original study, it is found that the genes MYBL2 and MYC are over expressed in malignant tumors. Multidimensional Visualization confirms this conclusion and further quantifies the amount of over-expression observed in these genes when the summarize operation is done over the basic visualization as shown in Figure 9. Observe that this result visually indicates that the gene MYBL2 is much more dominant in metastatic tumors when compared to MYC. The same conclusion was hypothesized by the original study through

biological experiments.

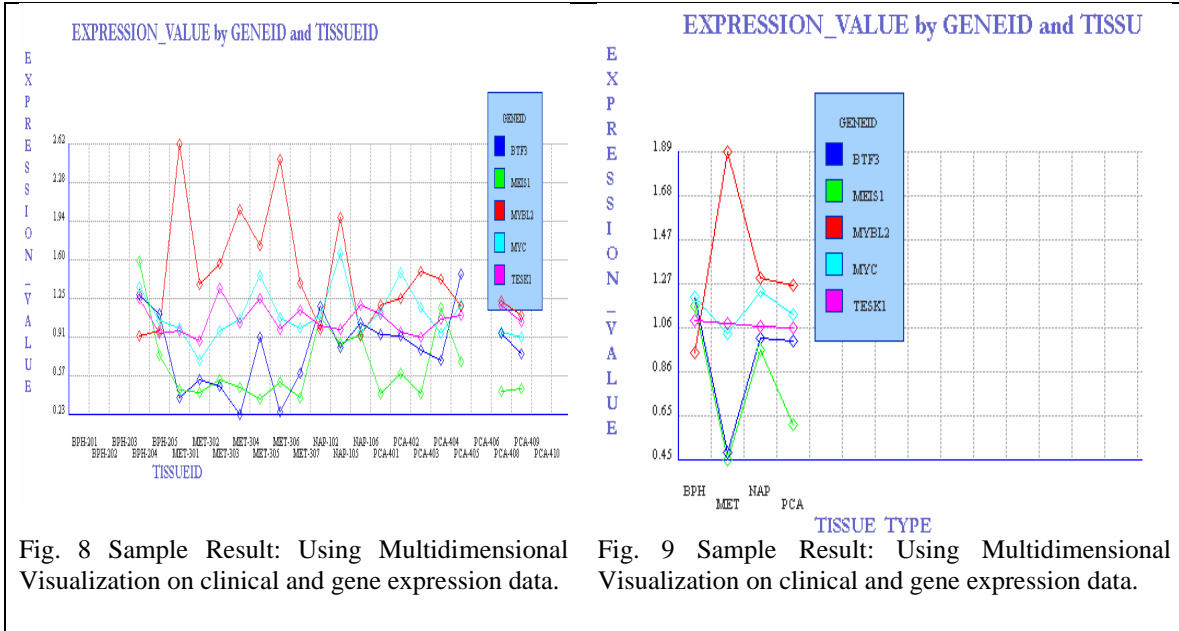


Fig. 8 Sample Result: Using Multidimensional Visualization on clinical and gene expression data.

Fig. 9 Sample Result: Using Multidimensional Visualization on clinical and gene expression data.

Multidimensional Visualization provides a simple, yet insightful method to visualize and explore the data available through the warehouse. The system offers users pre-defined aggregation operations and dimensional hierarchies and further allows the user to define his/her own operations and hierarchies.

B. Correspondence Analysis

Correspondence Analysis is a technique used to study the *correspondence* between disparate datasets that can be connected using a two-way table. The tool is based on a simple extension of the original technique proposed by Greenacre et al. [13]. The primary goal is to provide a system for *visual analysis* of the associations. Consider a simple two-way table constructed with “number of patients” as the content, and “age group of the patient” and “race of the patient” as row and columns. This can also be looked on as a 2-dimensional cube result from Multidimensional Visualization tool where “number of patients” is the fact, and “age group of the patient” and “race of the patient” are the

dimensions. Given this, consider the following query: “For patients belonging to race “A”, which is the most likely age group “G” at which they are diagnosed” from a global point of view of the data available on the platform. In other words, the user is interested in the association between the *profiles* of race “A” and age-group “G”. The central idea behind Correspondence Analysis is a *profile*. A two-way data subset consists of *row* profiles or *column* profiles. In the above example, race “A” could be a row profile and the age group “G” could be a column profile. The technique plots row and column profiles as points on a *symmetric map* which is interpreted as follows:

1. The origin corresponds to the *average profile* of the data.
2. It is comprised of the “optimal displays” of the row and column profiles, although strictly speaking these two sets of points occupy different spaces.
3. The distance between two points corresponds to the similarity/dissimilarity of the corresponding profiles.
4. The map is scaled such that row and column points are equally spread out along each principal axis (for a 2D plot, along the horizontal and vertical directions).
5. Although there is no direct interpretation of the distance between a row and a column point, there is certainly a joint interpretation of the row and column points with respect to the principal axes of the map.

Although Correspondence Analysis could be computed in multiple ways [13], we use a generalized Singular Value Decomposition (SVD) based algorithm to compute the point coordinates for the profiles. Algorithm 1 summarizes our method.

Input: Data Matrix $X(n \times m)$

Let $\mathbf{1}$ be a column vector of ones of appropriate order.

$$s = \mathbf{1}'X\mathbf{1}$$

Transform X by dividing each cell by the sum of elements $P = X/s$

Let $\text{diag}()$ create a diagonal matrix from a vector of some order

Row masses $r = P*\mathbf{1}$; $D_r = \text{diag}(r)$

Column masses $c = P'\mathbf{1}$; $D_c = \text{diag}(c)$

Perform generalized SVD on P obtain the point coordinates. This is achieved indirectly by performing ordinary SVD of Q given by

$$Q = D_r^{-1/2} P D_c^{-1/2}$$

Obtain normal SVD of Q , giving $Q = U D_\alpha V'$

Left generalized singular vector $A = D_r^{-1/2} U$,

Right generalized singular vector $B = D_c^{-1/2} V$ and

Generalized singular values $D_u = D_\alpha$

Output:

Row Coordinates $F = D_r^{-1} A D_u$

Column Coordinates $G = D_c^{-1} B D_u$

Algorithm 1. Generalized SVD based Correspondence Analysis

We skip further details here for the sake of simplicity and the interested reader is referred to [13] for a thorough exposition. We now demonstrate the functionalities of the tool through examples. Consider a 2D data cube along the dimensions “Race” and “Age” of the patient with “total number of patients” as the fact. Figure 10 is the snapshot of the result obtained by running the correspondence analysis tool on such an input. The following are some observations that can be drawn from this result.

- 1) The proximity of the profiles “Caucasian” and “50-60” to the centroid indicate a *typical* profile or the *average* profile. Based on this, one can hypothesize that middle aged and older Caucasians are more prone to prostate cancer as compared to other age groups.

- 2) The age groups (<40) and (80-90) which are far away from the origin, indicating that they are *atypical* profiles.

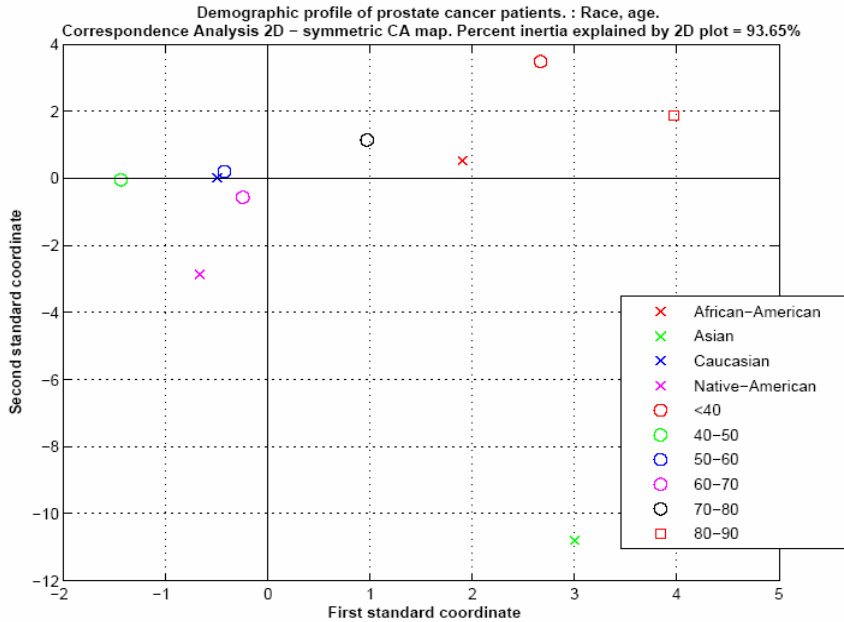


Fig. 10. Correspondence Analysis on Patient Demographic Data.

Similar studies can be made simultaneously on multiple datasets such as gene expression data and tissue sample data. Figure 11 displays the snapshot of the result obtained for the same dataset [12] we used earlier in section IIIA. Here, the input dataset contains gene expression profiles of 9984 human cDNA samples taken from tissues from different disease stages. The tissues are categorized as BPH : Benign Prostate Hyperlylsia; NAP : Normal Adjacent Prostate; PCA : Localized and MET : Metastatic. It can be observed that the benign and normal samples (BPH & NAP) lie to the right of the origin (w.r.t. the first principal axis) while the malignant states (PCA and MET samples) lie on the other side. Thus the visual interpretation indicates a separation between the benign and the malignant states along one of the principal axes. This observation was made by the original paper through biological experiments. The genes which lie to the left of the

origin tend to have a "positive association" with the malignant states. Such genes may be candidates for biomarkers especially if they are close to any of the clinical sample points in the graph. Ex. Hepsin, LIM(Enigma),PIM1,MYC. This conclusion was also actually hypothesized and verified in the original study [12]. Thus the Correspondence Analysis tool helps visualize and interpret the associations among clinical and genomic datasets in a comprehensible way.

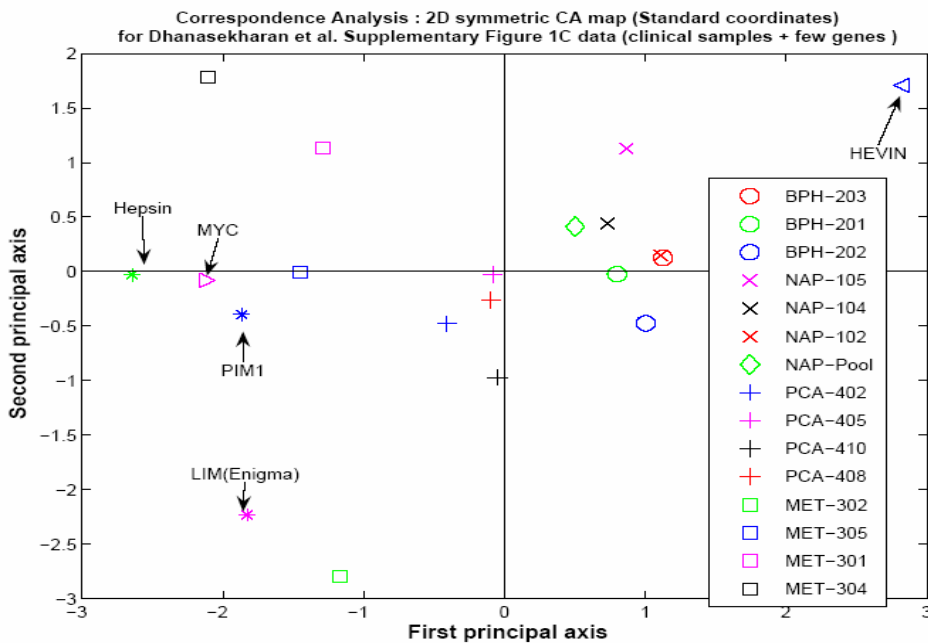


Fig. 11. Correspondence Analysis on Gene Expression Data.

C. Fusion Based Clustering

Clustering gene expression data is one of the most widely used analysis techniques in biomedical informatics. DNA microarrays are high-throughput experiments whereby it is possible to measure the changes in expression of several thousands of genes simultaneously. Cluster analysis of gene expression data is carried out by grouping genes based on similar expression profiles. Recent studies such as [6] motivate the inclusion of other information sources such as sequences, ontologies for

identifying gene clusters with similar biological functionality. Further, studies such as [8] indicate that including tissue data and clinical data in gene expression analysis might lead to better identification of biomarkers.

Our fusion based clustering algorithm is a simple extension of the basic Self-Organizing Map SOM [7] based clustering algorithm. It extends the basic clustering mechanism such that multiple datasets are taken into consideration. This is presented in Algorithm 2. The key challenges involved in designing such fusion based clustering algorithm 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. The algorithm starts by initializing the cluster centroids randomly. Based on the user input, the algorithm picks one of the datasets for each iteration and adjusts the centroids based on the distance-metric corresponding to that data set. This is done until there is little change in the position of the centroids. The algorithm is also capable of weighting 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.

The algorithm uses an iterative procedure by which the probability distribution of the data is reproduced as closely as possible. At each iteration step, a dataset is randomly selected based on the weighting scheme P . The chosen category r and its associated distance function d_r are used to train the network of neurons. The weights for the entire input tuple (of dimension $N_1+N_2+\dots+N_m$) are updated using the Kohonen learning rule [14], although the distances are calculated on each segment of the input vector independently using the appropriate distance-metric. Information-

theoretic similarity measures such as Kullback-Liebler are preferred for clustering of intensity values [15]. To measure similarity between genes based on frequency of motif occurrence, we use a measure based on the Extended Jaccard Similarity coefficient. We introduce a new method of assigning cluster memberships to data points based on weights assigned to each dataset through P . For each data point (gene), find the “closest” cluster centroid for each of the k datasets (using the corresponding coordinates and distance function). For each data point, the *likelihood* of each possible cluster is calculated by the sum of the weights assigned to each dataset indicating that particular cluster. The data point is finally assigned to the cluster with the maximum likelihood.

Let the number of datasets be k
Let N_g be the total number of data points (genes) to be clustered
Let the dimension size of each dataset be $N_1, N_2 \dots N_k$
Let D_i ($i = 1, \dots, k$) be the distance function associated with each dataset
Let $P = \{p_1, p_2, \dots, p_k\}$ be the weights/apriori probabilities assigned to each of the datasets such that $p_1 + p_2 \dots + p_k = 1$.

Initialize the N cluster centers: c_1, c_2, \dots, c_N . Normalize and transform the data independently for each dataset.

For iteration n (randomly permute the rows of the data every iteration):

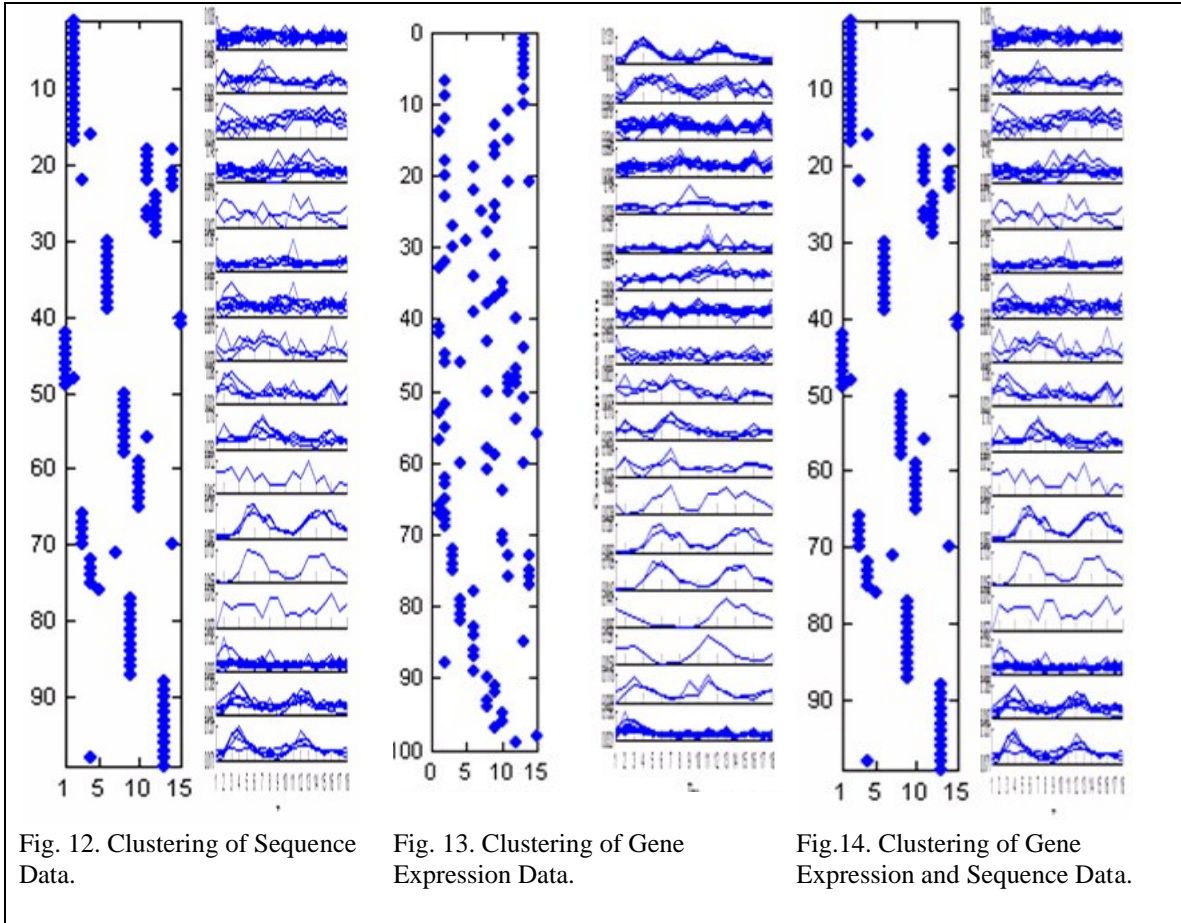
1. Select a gene g : $Xg = (x_g^{r1}, \dots, x_{gr}^{Nr1}, \dots, x_{gr}^{kl}, \dots, x_{grk}^{Nrk})$.
2. Select a dataset r randomly using P .
3. Calculate the distance d_i from the g to each cluster center c_i using the distance function D_r (only using the columns of X_g corresponding to category r , namely $(x_g^{r1}, \dots, x_{gr}^{Nr})$): $d_i = D_r(g, c_i)$, $i = 1, 2, \dots, N$.
4. Identify the cluster l closest to g .
5. Update the weights (for all coordinates corresponding to all categories) for cluster l and its immediate neighbors using the Kohonen learning rule.
6. Update the learning rates for all of the categories.

Iterate until convergence.

Algorithm 2. Fusion based Clustering

We now present the usage of the tool and some sample results obtained by running the tool on gene expression data collected from Spellman et al. [24]. The data set

consists of gene expression data for yeast cell cycle data. The data set was chosen since it was well studied in the literature and several interesting observations have been made based on cluster analysis. The usage of the tool involves the selection of datasets to be clustered, and the weights to be assigned to each data set. The tool outputs the result in a visualizable format depicting the resulting clusters. Figure 12 and 13 presents the results obtained by running this tool by considering only: 1. Gene Expression data 2. DNA sequence data (motif frequency data) respectively. Figure 14 presents the clustering results obtained by using the fusion based clustering algorithm on both the gene expression data and sequence data. The cluster obtained through the fusion based clustering resulted in “tighter clustering” i.e. with lesser intra-cluster distance and larger inter-cluster distance than when compared with clustering on individual datasets. Further analysis of one of the clusters obtained using Integrated clustering showed that 4 out of 5 genes in cluster 8, namely, CTF4, POL30, HYS2 and POL32 were mentioned in the original paper to be DNA Syn related genes. These genes also share a common transcription factor MCBa. These results suggest that fusion based clustering results in better identification of genes with similar function. Furthermore, one can hypothesize that genes that have a similar function might share a common expression profile and also a common motif.



IV. CONCLUSION

In this paper, we present an online platform for performing fusion based visualization and analysis of clinical and genomic datasets. The goal is to demonstrate the significance of information fusion in biomedical research. 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.

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] Falkman, G., Information visualisation in clinical Odontology: multidimensional analysis and interactive data exploration in *Artificial Intelligence in Medicine 22*, pp.133-158, 2001
- [3] Shahar et al. Interactive visualization and exploration of time-oriented clinical data using a distributed temporal-abstraction architecture. *Proceedings of the AMIA Annual Symposium*, 2003
- [4] Granzow,M., Berrar,D., Dubitzky,W., Schuster,A., Azuaje,F. and Eils,R. (2001) Tumor identification by gene expression profiles: a comparison of five different clustering methods. *ACM-SIGBIO Newsllett.*, **21**, 16–22.
- [5] Irene M Mullins, Mir S Siadaty, Jason Lyman, Ken Scully, Carleton T Garrett et al. Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Comput Biol Med.* 2005 Dec 20.
- [6] I. Holmes, W.J. Bruno, "Finding Regulatory Elements Using Joint Likelihoods for Sequence and Expression Profile Data," *ISMB*, 2000, pp.202-210.
- [7] Pennsylvania Cancer Alliance for Bioinformatics Consortium (PCABC), <http://www.pcabc.upmc.edu/main.cfm> .
- [8] D. Singh et al. "Gene Expression correlates of clinical prostate cancer behavior," *Cancer Cell*, March 2002, pp.1(2):203-9.
- [9] 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.
- [10] R. Agrawal, A. Gupta, S. Sarawagi. "Modeling Multidimensional Databases", *Proc of 13th Intl Conf on Data Engineering, ICDE 1995*.
- [11] Y. Tao, C. Friedman, Y.A. Lussier, "Visualizing Information across Multidimensional Post-Genomic Structured and Textual Databases," *Bioinformatics*, Dec 2004.
- [12] Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., Chinnaiyan, A. M., 2001. Delineation of prognostic biomarkers in prostate cancer. *Nature* 412, 822-6.
- [13] Greenacre M.J. "Correspondence Analysis in practice", 1993, Academic Press London.
- [14] Kohonen, T. (1995) *Self-Organizing Maps*. Springer Series in Information Sciences, Springer, Berlin.
- [15] Kasturi, J., Acharya, R. and Ramanathan, M. (2003) An information theoretic approach for analyzing temporal patterns of gene expression. *Bioinformatics*, **19**, 449–458.
- [16] R. Calinski and J. Harabasz, *A dendrite method for cluster analysis*, *Commun. Statistics*, 1974, vol 3, pages 1-27
- [17] J. C. Dunn, *Well separated clusters and optimal fuzzy partitions*, *Journal of Cybernetics*, 1974, vol 4, pages 95-104
- [18] D. L. Davies and D. W. Bouldin, A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979, vol 1, No. 2, pages 224-227
- [19] J. A. Hartigan, Statistical theory in clustering, *Journal of Classification*, 1985, Vol 2, pages 63-76
- [20] W. M. Rand, *Objective criteria for the evaluation of clustering methods*, *Journal of the American Statistical Association*, 1971, volume 66, pages 846-850
- [21] Marina Meila, *Comparing Clusterings*, Department of Statistics, University of Washington, October 2002
- [22] S. Dongen, *Performance criteria for graph clustering and Markov cluster experiments*, Centrum voor Wiskunde en Informatica, 2000
- [23] Ding Zhou, Jia Li and Hongyuan Zha, *A new Mallows distance based metric for comparing clusterings*, *Proc. International Conference on Machine Learning (ICML) 2005*, Bonn, Germany
- [24] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
- [25] MIPS Comprehensive Yeast Genome Database. <http://mips.gsf.de/genre/proj/yeast>