

Mining Gene Expression Database for Primary Human Disease Tissues

Andrew Campen^{#1}, Yuni Xia^{#2}, Dan Rigsby^{#3}, Ying Guo^{#4},
Xingdong Feng^{*5}, Eric W. Su^{*6}, Mathew Palakal^{#7}, Shuyu Li^{†*8}

[#]*Department of Computer Science, Indiana University Purdue University - Indianapolis
Indianapolis, Indiana, USA*

¹andrew.m.campen@gmail.com

²yxia@cs.iupui.edu

⁷mpalakal@iupui.edu

^{*}*Integrative Biology, Eli Lilly and Company,*

Indianapolis, Indiana, USA

[†]LI_SHUYU_DAN@lilly.com

Abstract— Studies of gene expression in primary human disease tissue often span several years in order to achieve reasonably large sample sizes and to collect patient clinical information making this data particularly valuable. Due to the lack of a central repository, this data has only been available through disparate and non-publicly accessible sources following publication. We developed Disease-to-Gene Expression Mapper (D-GEM) as a publically accessible database and data mining toolbox for microarray data of human primary disease tissue. A statistical pipeline has also been implemented to identify genes over-expressed in disease tissue samples in comparison with normal control samples, or genes whose expression values are associated with clinical parameters such as patient survival rate. One potential application of this data is the identification of pathway specific cancer prognosis markers. By applying a novel, gene signatures for cancer prognosis in the context of known biological pathways in cancer development were identified and confirmed.

I. INTRODUCTION

Gene expression patterns could provide valuable insights into the molecular mechanism of human diseases. It often involves a broad collaboration between universities and major hospitals to carry out a microarray gene expression profiling experiment using human disease tissue samples. These studies usually span several years in order to acquire large sample size to achieve statistical significance. Furthermore, substantial efforts are required to collect patient clinical information accompanying the gene expression data. Therefore, the data generated from these studies represent a valuable asset to the scientific community. However, a database is lacking to centralize microarray gene expression data of other human disease tissues. Because of the great value that these gene expression data and the associated clinical data provide to the community, it is imperative to implement such a database as well as related tools for data access and analysis. Here we report an ongoing effort of developing the DGEM (Disease-to-Gene Expression Mapper) database [1], a statistical analysis pipeline and a web interface to access the raw data and data analysis results.

II. DGEM DATABASE

Our goal is to compile and analyse all public disease microarray data. We identified potential datasets by literature searching, focusing on those that are generated using the primary human tissue samples. As of May 2007, our collection includes 41 human diseases. There are a total of 1,420 human samples and 24,311,392 gene expression measurements. In our data mining pipeline, we examined the association between gene expression and cancer patient survival.

For each probe set on a microarray, a pair-wise t-test was used to evaluate if a gene is over expressed in disease tissues compared with the corresponding normal tissue. The p values were adjusted for multiple testing using the false discovery rate. The association between gene expression and cancer patient survival was assessed following the method of Cox Proportional Hazards regression [5]. Data analysis was implemented using Java and the R statistical computing package. The DGEM database is designed around three components: the gene expression data, the patient sample data and the data mining results.

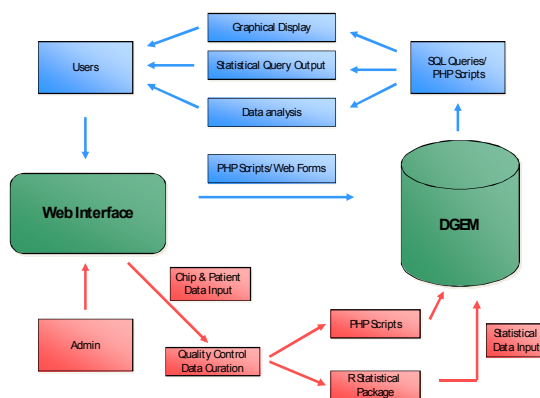


Fig. 1 the D-GEM system workflow.

Figure 1 illustrates the flow of information that occurs when DGEM is accessed by either Users or administrators through the web interface. The red arrows indicate work flow for administrative functions and blue arrows indicate the step in returning results to users.

III. IDENTIFYING PATHWAY SPECIFIC CANCER PROGNOSIS MARKERS USING MICROARRAY DATABASE

One application of the microarray data is the identification of pathway specific cancer prognosis markers. In this study, we propose a novel approach to identify gene signatures for cancer prognosis in the context of well defined pathways and apply it to microarray data sets representing several different types of cancer. Initially, gene expression data was filter to only represent gene documented to be involved in a specific pathways. Unsupervised hierarchical clustering analysis was then applied to divide patient samples into separated groups based on the filtered gene expression. Patient survival rates in different groups were compared using Kaplan-Meier survival analysis. If a specific pathway plays a critical role in tumor progression and metastasis, patients with distinct gene expression patterns in that pathway may have very different clinical outcome.

A. Hierarchical Clustering of Gene Expression and Sample data

Each pathway specific data set was analyzed by two-way hierarchical average-linkage clustering. Hierarchical clustering uses agglomerative methods which begin with a set of n separated objects that are joined into successively smaller numbers of groups. All hierarchical clustering methods result in a two dimensional tree diagram known as a dendrogram which illustrates the fusions made at each stage of the cluster analysis. Here the distance between two clusters is defined as the average of distance between objects in one cluster and objects in another cluster. Mathematically, the average linkage function between two clusters X and Y is computed as:
$$D(X, Y) = \frac{1}{N_X \times N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} d(x_i, y_j)$$
 where $d(x, y)$ is the distance between objects $x \in X$ and $y \in Y$, and N_X and N_Y are the number of objects in clusters X and Y respectively. At each stage of hierarchical clustering, the clusters X and Y are merged if $D(X, Y)$ is the minimum. This analysis used the uncentered correlation value between expression values as the clustering distance metric.

The hierarchical analysis was also two-way, meaning that both the genes as well as the patient samples were clustered independently. The output of this clustering was both dendrograms for genes and patient samples and a heat map which graphically illustrated the variation of the expression values through a color scale. The dendrogram for patient samples was examined first to see if adequate separation into two groups were achieved then to place each sample into one of two separate clusters.

B. Patient Survival Analysis

Each of the data sets was selected for containing additional clinical information about each patient sample. Included in this information was patient survival state, given as alive or dead, and follow-up time. Following, the cluster analysis and separated of the patient samples into two groups, the associated clinical information was gathered for each group then used in Kaplan-Meier survival analysis. Kaplan-Meier analysis aims to estimate the survival function given life-time data [2]. This estimate can be illustrated graphically by plotting the survival estimate as a series of horizontal step over a given time series. This plot is referred to as a Kaplan-Meier curve. Each of the resulting clusters from the previous analysis was plotted in this manner then the Kaplan-Meier curves were compared using a log rank test [3]. Curves that displayed a significantly different survival estimated resulted in a small p-value were compared using the log rank test while curves that were more similar resulted in a higher p-value. The log rank test p-value was used as the measure of significance as to how well that particular pathway performed as a prognosis marked for that data set.

IV. DEMONSTRATION

DGEM microarray data repository is available at <http://dgem.cs.iupui.edu>. Queries can be performed at either the disease level for retrieving a list of genes up-regulated in a disease, or at the gene level for identifying diseases that the gene of interest is over-expressed.

The main query page includes a drop-down menu that displays the disease names, as shown in Figure 2. A disease-centric search generates a list of genes that are up-regulated in user specified disease. We use the adjusted P value or false discovery rate 0.2 as a default threshold to specify significant aberrant gene expression in the disease tissues.



Figure 2: DGEM Search

Figure 3 shows the output table for search by the disease. Included in the output table are columns of Gene ID, gene symbol, gene description, average expression values in the normal and disease samples, the adjusted P values, fold change in the disease group vs. the normal group.

Probeset	Gene Symbol	Description	Entrez Gene ID	Expression in disease samples	Expression in normal samples	Fold Change	P Value	Adjusted P Value (FDR)	KEGG Pathway
215306_at	LHCGR	Luteinizing hormone/choriogonadotropin receptor	3973	619.23195	188.30650	3.28843	0.0000011	0.0238956	3973
203461_at	CHD2	chromodomain helicase DNA binding protein 2	1106	351.85658	129.86568	2.70939	0.0000246	0.0913154	1106
219746_at	DPF3	D4, zinc and double PHD fingers, family 2	8110	599.06446	315.34968	1.89968	0.0000078	0.0436984	8110
203824_at	TUBG2	tubulin, gamma 2	27175	629.71478	1043.34396	-1.65685	0.0000075	0.0598279	27175
218801_at	UGCGL2	UDP-glucose ceramide glucosyltransferase-like 2	55757	332.64638	614.32428	-1.84678	0.0000066	0.0737734	55757
213666_at	SEPT6	septin 6	23157	466.49811	885.59496	-1.89839	0.0000260	0.0627942	23157

Figure 3: An example of search by disease

Many of these columns can be further linked to more detailed information. The fold changes are linked to a plot graphically showing expression values of individual samples in the normal and the disease groups that are distinguished by two colors. Figure 4 shows an example of a box plot which compares the expression values in the normal and disease groups. Entrez Gene IDs are used to link the output of a search to Entrez Gene page at NCBI and the KEGG pathway database.

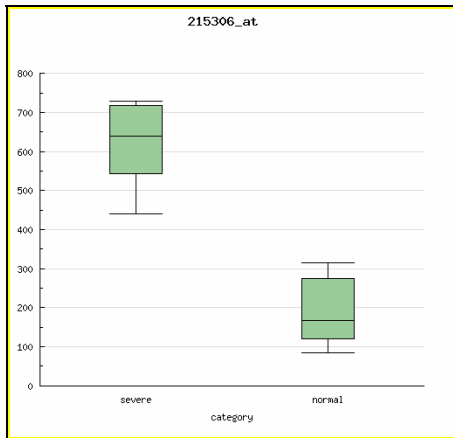


Figure 4: Box plots of gene expression values of individual human samples

A gene centric search generates an output table including the analysis results for all of the disease datasets in the DGEM database. The output provides information such as average expression values in the normal and disease samples, the adjusted P values, fold change in the disease group vs. the normal group, and links to the OMIM page at NCBI for each disease. Figure 5 shows an example of the table of search by gene.

Integration of gene expression data from human samples with that from specific animal studies provides a unique approach to identify clinical biomarkers for drug response. For example, comparison of gene expression profiles of the cultured cells or animals treated with a drug with that of untreated controls will identify candidate biomarkers for drug response. In order to determine if these markers are applicable in clinical trials, one can investigate the expression pattern of these potential markers in human disease tissues. Up-regulation of these potential biomarkers would suggest they could serve as markers not only for drug response in human but also for patient stratification.

Disease	Tissue	Expression in disease samples	Expression in normal samples	Fold Change	P Value	Adjusted P Value (FDR)	OMIM
Acquired Immunodeficiency Syndrome (AIDS)	PMC	133.22956	11.58342	11.50174	0.0008960	0.3193954	609423
Acquired Immunodeficiency Syndrome (AIDS)	PMC	65.10462	11.58342	5.62050	0.1415487	1.0040801	609423
Acquired Immunodeficiency Syndrome (AIDS)	PMC	260.10857	137.48667	1.89188	0.2550046	0.7084675	609423
Diabetes	skeletal muscle	186.07509	191.15100	-1.02728	0.8890342	0.9637953	125853
Amotrophic Lateral Sclerosis (Lou Gehrig's Disease)	spinal cord gray matter	205.73287	273.92435	-1.33146	0.4121281	0.7073916	105400
Acquired Immunodeficiency Syndrome (AIDS)	PMC	95.06932	137.48667	-1.44617	0.3037871	0.9302829	609423
Acquired Immunodeficiency Syndrome (AIDS)	PMC	172.47542	251.73497	-1.45954	0.4767566	0.6317468	609423
Pulmonary Hypertension	PBMC	30.00668	55.24801	-1.84119	0.1061171	0.273213	178600
Acquired Immunodeficiency Syndrome (AIDS)	PMC	115.57467	251.73497	-2.17812	0.0622942	0.3140704	609423
Amotrophic Lateral Sclerosis (Lou Gehrig's Disease)	spinal cord gray matter	79.83336	273.92435	-3.43120	0.0072219	0.6899843	105400
diffused Large B Cell Lymphoma	blood			N/A	N/A	0.7849116	605027

Figure 5: An example of search by gene

V. USE DGEM FOR IDENTIFYING PATHWAY SPECIFIC GENE EXPRESSION PREDICTORS FOR CLINICAL OUTCOME

Six cancer data sets were analyzed as demonstrated for the 20 molecular pathways that were assembled. As described previously, through the unsupervised hierarchical clustering methods, cancer patients were separated into distinct groups based on gene expression patterns in one of the cancer pathways. The survival probabilities were then compared between the patient groups to determine if differential gene expressions in a specific pathway were associated with differential patient survival, which would suggest that the pathway may play a critical role in tumor progression. Our results suggest expression of genes in cell cycle, specifically G1-S, apoptosis, angiogenesis, and p53 pathways could be used as breast cancer prognosis markers.

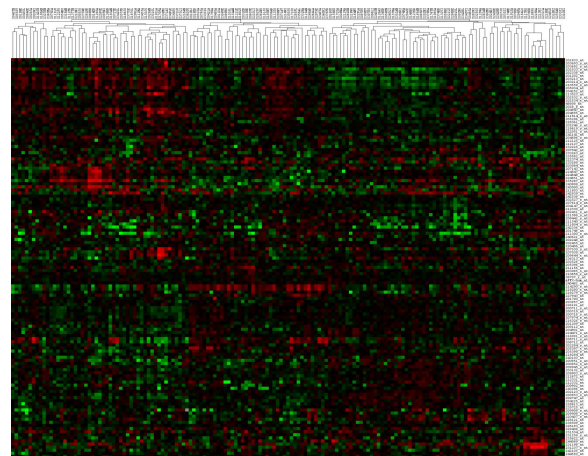


Figure 6: Hierarchical clustering heatmap of breast cancers in G1-S pathway.

As a positive control experiment to verify the conceptual basis of this study, we tested genes well known to be involved in breast cancer pathways to determine how well they function as molecular markers for prognosis and diagnosis of breast cancer. A total of 264 genes, derived from literature as well as previous microarray studies, were tested. It is expected that breast cancer data sets would demonstrate significantly

different survival curves from their respective patient groupings. This is indeed that case

Figure 6 illustrates a heat map of the breast cancer pathway gene expressions in 159 samples of one dataset. The column dendrogram indicated these 159 patients were clustered into two groups with completely opposite expression patterns as illustrated by the opposing green and red values. The two groups also exhibited a dramatically different clinical outcome as revealed by the Kaplan-Meier analysis curves in Figure 7 which resulted in a small p-value when the curves were compared using a log rank test. This single pathway example serves as a successful proof-of-concept test verifying the initial hypothesis.

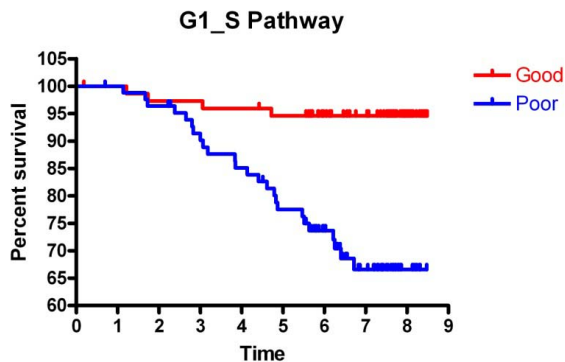


Figure 7: Kaplan-Meier survival analysis of breast cancer patient clusters.

VI. USE DGEM FOR IDENTIFYING PATHWAY SPECIFIC GENE EXPRESSION PREDICTORS FOR CLINICAL OUTCOME

As a public gateway to gene expression profiling data of human disease samples, DGEM provides scientists and physicians a valuable tool to study disease mechanisms and to identify novel gene targets for drug discovery. We will continue to acquire microarray datasets of human diseases and at the same time, enhance our data mining process.

REFERENCES

- [1] Y. Xia, A. Campen, D. Rigsby, Y. Guo, X. Feng, E. Su, M. Palakal, S. Li, DGEM: Mining Gene Expression Database for Primary Human Disease Tissues, *Molecular Diagnosis & Therapy*, Issue 3, 2007
- [2] Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, 53, 457-481.
- [3] Bland, J. M., & Altman, D. G. (2004). The logrank test. *Bmj*, 328(7447), 1073.