33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Gene Expression Analysis with Integrated Fuzzy C-means and Pathway Analysis

Mingrui Zhang, Beya Adamu, Chi-Cheng Lin and Ping Yang

***Abstract*—A workflow for associating fuzzy clusters to biological pathways has been implemented as a Java-based software tool. Its software implementation is comprised of a correlation-based fuzzy c-means algorithm and an enrichment test on Kyoto Encyclopedia of Genes and Genomes' pathways. We applied this workflow to gene expression in classification of lung cancer cell types and achieved satisfactory results. The software could aid in the validation of results of fuzzy clustering algorithms and the exploration of un-annotated associations between genes and gene ontology categories.**

I. Introduction

THE study of human genome involves a tremendous amount of data, and clustering methods are often used to explore the patterns within. Clustering algorithms are classified into crisp or fuzzy methods. Researchers have shown that most of the commonly used crisp algorithms were unable to identify genes whose expression is similar to multiple, distinct gene groups. Though crisp algorithms failed to mask the relationships between genes that are co-regulated with different groups of genes, fuzzy clustering method was able to identify those relationships[1].

In clustering microarray data, a fuzzy clustering algorithm assigns a gene with degrees of memberships to multiple clusters. The membership is a value between 0 and 1 with one indicating a complete association to a cluster [2]. During clustering, the algorithm minimizes an objective function. A recent study has shown that the fuzzy c-means (FCM) algorithm with a correlation-based objective function outperforms the algorithm with Euclidean distance metrics [3].

Biologists use different databases to store and for sharing biological information. In addition to millions of scientific publications, these databases provide descriptive genomics, gene annotations, and simulation of biological data. Constantly updated and verified, most of them have Web interfaces to facilitate user's research [4, 5]. Among the information are the gene ontology (GO) and pathway databases. The GO describes the molecular functions of gene products and their roles in multi-step biological processes. The database is created using existing genomic databases and published literatures. It is built on a set of controlled vocabulary which describes gene and gene product attributes in organisms. The GO categories have often been adopted in validating the results of clustering in several studies [6, 7]. The pathway is another way to capture the roles genes play in biochemical reactions that help sustain life. A pathway is a sequence of enzymatic reactions by which one biological material is converted to another [8]. Identifying genes in biological pathways is an important instrument for early disease detection and diagnosis.

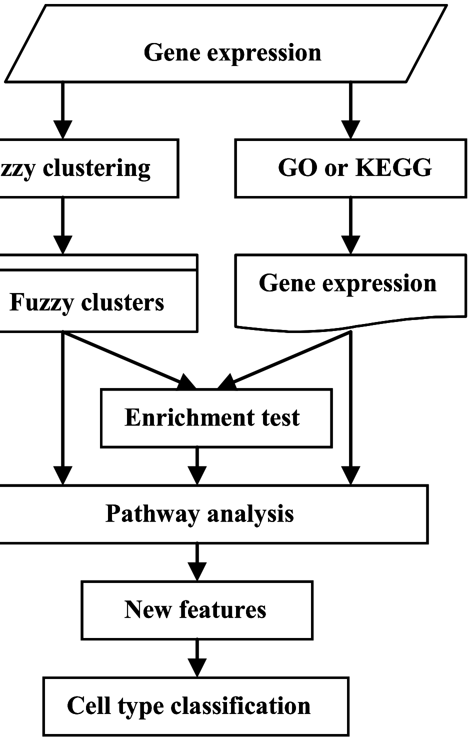


Fig. 1. Classification of cell types by integrating FCM analysis and Pathway annotations.

With a fuzzy clustering method, genes are grouped together according to a mathematic or statistical metric. Given the fact that a gene may play roles on multiple pathways, the numbers of genes attributed to cancer prognosis would be abundant. However, studies have suggested that underlining pathways could be limited and essentially identical. We propose a workflow to integrate fuzzy clustering and pathway analysis, and apply the workflow to classify lung cancer cell types. The rest of the paper is organized as follows. We introduce a workflow for integrating fuzzy clustering and pathway analysis in Section II. Then, we discuss a software implementation of the

Manuscript received March 26, 2011.

M. Zhang is with the Computer Science Department, Winona State University, Winona, MN 55987, USA (phone: 507-457-2980; fax: 507-457-2464; e-mail: MZhang@winona.edu).

B. Adamu and C. Lin are with the Computer Science Department, Winona State University, Winona, MN 55987, USA.

P. Yang is with Department of Health Science Research, Mayo Clinic College of Medicine, MN 55905, (e-mail: Yang.Ping@mayo.edu).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

workflow in Section III, and apply it to classify cell types of lung cancer in Section IV. We conclude in Section V.

II. Integration of FCM and KEGG Pathways

The workflow shown in Fig. 1 illustrates a means to associate fuzzy clusters of genes to predefined functional categories, such as GO terms or pathways. An FCM algorithm is applied to gene expression data to produce cluster centroids and to assign each gene to cluster centroids with memberships. In relating a fuzzy cluster to biological pathways, we test the enrichment of pathways by the cluster of genes. Clusters of genes with fuzzy memberships, *p*-values of enrichment tests and gene annotations are inputs to the visualization software developed in-house. After genes are identified, they are used in computing new features in classifications of cell types of lung cancer.

Table I. Tabular form of fuzzy memberships

Gene symbol	Cluster 1	Cluster 2	Cluster 3
PAPSS2	0.089	0.28	0.631
SULT1A2	0.044	0.38	0.576
GPD1	0.69	0.134	0.176

A. Fuzzy C-means Algorithm

Studies have shown that fuzzy clustering methods are capable of identifying genes whose expression is similar to multiple, distinct gene groups [6]. When it is used, a fuzzy clustering algorithm assigns genes to a given number of clusters such that each gene may belong to more than one cluster with different degrees of memberships [2]. Outputs from fuzzy clustering include matrices of memberships (Table I) and cluster centroids. For example, gene PAPSS2 belongs to Cluster 2 by membership 0.28 and to Cluster 3 by 0.631. A fuzzy membership is a value between 0 and 1 with one indicating a complete association to a cluster. A gene has a total membership value of 1.0 across the clusters.

Table II. Pathway enrichment of fuzzy clusters

C_k	Functional Category	n_k	m_k	*p*-Value
6	Cell cycle process	206	30	8.61E-08
	Cell cycle		36	8.38E-10
	Cellular metabolic process		139	2.39E-17
28	Defense response	120	27	1.13E-12
	Immune response		43	4.52E-25
	Response to external stimulus		26	9.38E-11
31	Immune response	77	16	2.35E-06

B. Enrichment Test

Genes are classified by GO terms or pathway annotations. In GO, annotation terms are organized into biological process, molecular function and the cellular component categories. In Kyoto Encyclopedia of Genes and Genomes (KEGG), they are mapped to pathways. Enrichment test is done assuming a hypergeometric probability distribution of genes within a cluster. Hypergeometric probability distribution can be used to compute a probability that an observed enrichment of a functional category comes from randomly selected genes [9]. The significance (*p*-value) of enrichment is defined as:

$$p\text{-value} = \left(\sum_{i=m_k}^{M_k} \binom{M_k}{i} \binom{N_k - M_k}{n_k - i} \right) \Big/ \binom{N_k}{n_k} \qquad (1)$$

Here N_k and n_k are the numbers of genes in fuzzy cluster k before and after membership threshold, respectively; and M_k and m_k are the numbers of genes in the cluster assigned to a functional category before and after membership threshold, respectively.

Examples of p-values computed using GO biological processes are given in Table II. It is worth of noting that a cluster may enrich multiple pathways, e.g., Cluster 6 significantly enriches cell cycle, cell cycle, and cellular metabolic process pathways.

III. Software Implementation

Graph-based methods are often used in visualizing cluster structures and their relationship. They include heatmaps, neighborhood graphs and scatter plots [6, 10, 11]. Though results of hierarchical clustering are best presented as trees, i.e., dendrograms, results of centroid-based clustering is commonly projected into 2-D space like scatter plots [12]. With scatter plots, one may visualize the clustered data in different colors and with annotations to represent the cluster membership.

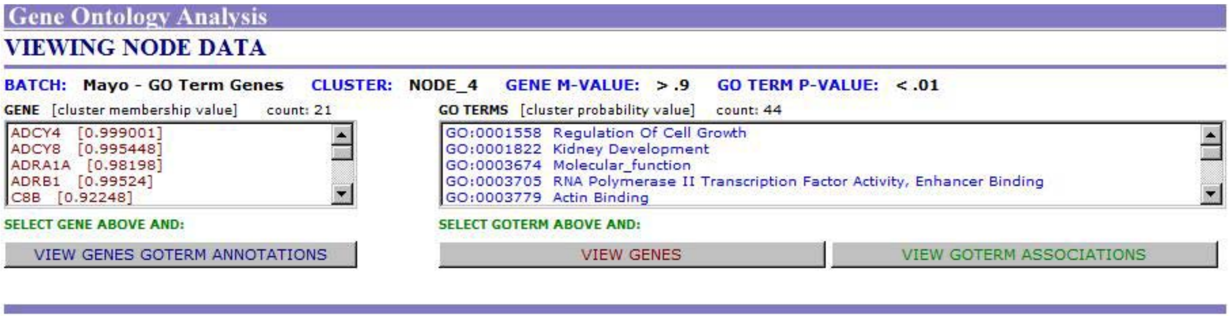
The workflow is implemented as a Java applet with Web-based user interfaces. Java Servlets and Apache Derby are used to build, store data to and query biological data from local and remote servers. To speed up the data analysis, the GO databases are downloaded from the GO consortium [4] and stored locally. The software accesses the remote KEGG database [13] to acquire its pathway networks.

A. Remote Access to Sources of Annotations

Our software implementation takes in three datasets: gene clusters with fuzzy memberships, gene annotations that reveal the relationships between genes and functional categories (e.g. pathways, biological processes), and the *p*-values of enrichment tests. These datasets are modeled with three entity sets, *Cluster Membership Values*, *Pathway Information*, and *Cluster p-value* in our database.

The software consists of three main java classes: *file loader*, *data validator*, and *cluster processor*. The *file loader* helps user upload gene clusters with memberships, *p*-values, and classifications of genes in a chosen functional category. The *validator* verifies the uploaded fuzzy clusters, pathways or GO annotations acquired from remote databases. Once all files are uploaded successfully, the application issues a confirmation batch id. This batch id can be used to reference uploaded clusters for future use.

Data validator uses an XML parser and a remote reader object (RRO). The XML parser is implemented to process the System Biology Markup Language (SBML), an XML-based, machine readable markup language for representing models of biological processes [14]. The RRO initiates an ftp



Gene Ontology Analysis

VIEWING NODE DATA

BATCH: Mayo - GO Term Genes CLUSTER: NODE_4 GENE M-VALUE: > .9 GO TERM P-VALUE: < .01

GENE [cluster membership value] count: 21

```
ADCY4  [0.999001]
ADCY8  [0.995448]
ADRA1A  [0.98198]
ADRB1  [0.99524]
C8B  [0.92248]
```

GO TERMS [cluster probability value] count: 44

```
GO:0001558  Regulation Of Cell Growth
GO:0001822  Kidney Development
GO:0003674  Molecular_function
GO:0003705  RNA Polymerase II Transcription Factor Activity, Enhancer Binding
GO:0003779  Actin Binding
```

SELECT GENE ABOVE AND:

SELECT GOTERM ABOVE AND:

Fig. 2. Pathways enrichment by fuzzy clustering and gene-pathway relationship revelation

connection with a remote bioinformatics database server, retrieves information from the server, and stores the information locally. With the RRO, software may routinely check with the database servers for updates. A widely accessed bioinformatics database is KEGG [15], whose pathway database provides a collection of metabolic pathway annotations. The pathway network is a way of describing biological interactions and the interacting compounds. Its database is accessed using RRO.

B. *Visualizing Fuzzy Clusters*

To work with a fuzzy clustering algorithm, the *cluster processor* is implemented to help user to select fuzzy cluster(s) and to visualize them. Its Web interface lists multiple fuzzy clusters for user to choose from and to set the threshold values for fuzzy memberships and *p*-values. Once a fuzzy cluster is selected, all genes with memberships greater than and gene clusters with *p*-values less than user entered thresholds are processed (Fig. 2). The Web interface lists the genes in a chosen cluster, and all pathways being enriched by the cluster.

Each fuzzy cluster is processed by two java objects, *TablularForm* and *NetworkDrawer*. The former is used to manage fuzzy clusters of genes with enriched pathways. Each gene in it is associated with a set of functional categories according to KEGG's or GO's annotations. The *NetworkDrawer* class draws the pathway networks and annotates fuzzy clusters to them. Information needed for redrawing a pathway network is obtained from the KEGG database server. They are described in the KEGG Markup Language (KGML), a data exchange format used by KEGG's pathway graph objects. KGML makes automatic redraw of KEGG pathways possible and provides interfaces for modeling protein networks. A typical KGML file contains specifications of graph objects, where entry elements are treated as nodes and relations and reaction elements as edges. When a network is redrawn, the fuzzy cluster processor passes a list of genes to the applet and displays them on pathways.

C. *Software Validation*

Before the workflow is applied to RNA expression, we compared our software to a pathway network analysis tool MetaCore (http://www.genego.com). We randomly picked several fuzzy clusters to examine by setting the thresholds to 0.8 for fuzzy membership and 0.01 for the *p*-value. For instance, one of the genes being grouped into the cluster is PAPSS2 with the membership > 0.9. Two pathways are enriched by the cluster of gene PAPSS2, namely the Purine Metabolism pathway and the Sulfur Metabolism pathway. The pathway analysis on MetaCore confirmed our results. Although it is inclusive, our software implementation provides a means to relate fuzzy clusters to biological pathway networks. User may choose a specific pathway to explore and redraw the pathway. In addition, it allows a user to drag nodes on pathways to arrange a better view of a pathway (Fig. 3). The initial version of the software is available at http://cs.winona.edu/fuzzygenes.

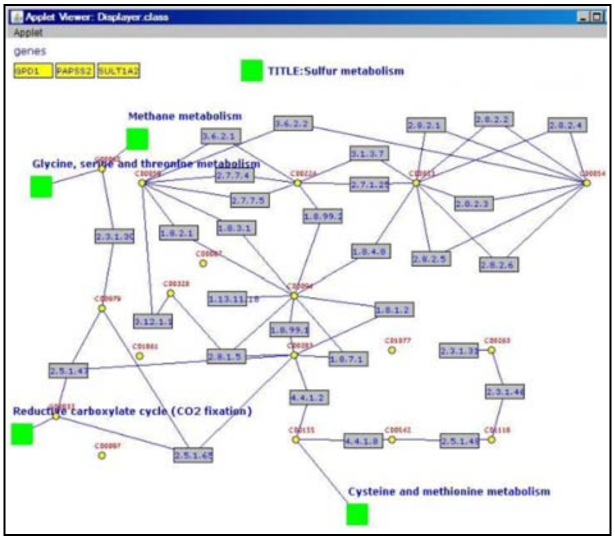


Fig. 3. The redrawn of pathway network enriched by a fuzzy cluster.

IV. Analysis of Lung Cancer RNA Expression

Lung cancer is a heterogeneous disease resulting from the acquisition of multiple somatic mutations. Its heterogeneity calls for a fuzzy clustering method.

A. *Gene Expression*

The RNA was extracted from a total of 190 never-smoker lung cancer patients, from both normal and cancer tissues. Eighty of them are adenocarcinoma, and thirty five are carcinoid typical. For Mayo Clinic discovery set, patients with lung cancer who were classified as never smokers were identified and recruited between January, 1997, and September, 2008. Never smokers were defined as individuals who had smoked less than 100 cigarettes during their lifetime. A detailed explanation of the recruitment process has been reported previously in [16].

Illumina Human WG DASL beadchip (Illumina, Inc, San Diego, CA, USA) was used for gene expression profiling. The expression data consisted of transcript levels for 24526 microarray probes, representing 18626 unique genes. Samples that passed quality control were merged and normalized together by use of the R faster cyclic loess function (Fastlo) [17].

B. Classifications of Lung Cancer Cell Types

A correlation-based FCM algorithm is chosen; a validity measure *fWCSS* was computed to determine appropriate number of clusters [7]. A close examination of *fWCSS* index revealed 40 potential clusters within the data set, and the dataset is thus segmented into 40 clusters. Fuzzy clusters of genes are input to the software implementation of the workflow. Pathway information was obtained from the KEGG database. There are 1,923 relationships found between the genes and the pathways. Each fuzzy cluster is processed if it significantly enriches a pathway (*p*-value<0.005). Within each cluster, only genes with fuzzy membership □ *>0.2* are considered.

After all fuzzy clusters processed, genes are ranked by number of pathways they enrich. The top 50 genes are selected. Gene is excluded from further analysis if more than 80% of pathways it enriches are already covered by genes ahead in the list. We were able to identify 27 genes out of five fuzzy clusters enriching total of 44 pathways. A new feature is computed for each cluster *C* and sample tissue *j* as

$$feature_{jC} = \left(\sum_i \mu_{iC} \times p_{ij}\right) / \left(\sum_i \mu_{iC}\right) \qquad (2)$$

Where, □ □C is membership of i^{th} gene to C^{th} cluster, and $\boldsymbol{p}_{ij}$ is gene expression of sample tissue *j*. By projecting labeled samples to feature space with new features, we were able to identify boundaries between adenocarcino, carcinoid typical and the rest of cell types. Overall, we correctly classified 93% of adenocarcinoma and 83% of carcinoid typical cell types.

V. Discussion

In this paper, we present a workflow to associate fuzzy clusters with pathways, and discuss a software implementation for unveiling the associations between clusters or genes and pathways. To demonstrate the use of the workflow, a software tool has been implemented and applied to classify cell types of lung cancer samples. Clustering methods are intended for exploration purposes. One may apply the methods to clustering an un-annotated gene and associate it to a functional category. This software provides a means to do so. The same workflow could be applied to GO categories with minor modification to the software. The significance of functional enrichment is calculated by comparing genes in a cluster to its GO or pathway annotations. If a gene is un-annotated, we would include this gene with annotated genes in a clustering process and restrict the number of un-annotated genes to be small. To work with a large number of un-annotated genes, one should break them into subsets and work with one subset at a time. Otherwise, statistical power of clustering results would be affected and so are the *p*-values. As a result, one may never discover any associations between the genes under study and the existing functional categories.

References

[1] A. Gasch and M. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biology,* vol. 3, pp. research0059.1 - research0059.22, 2002.

[2] J. C. Bezdek, *Pattern recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1988.

[3] M. Zhang, T. Therneau, M. A. McKenzie, P. Li, and P. Yang, "A Fuzzy C-Means Algorithm Using a Correlation Metrics and Gene Ontology," in *The 19th International Conference on Pattern Recognition*, Tampa, Florida, USA, 2008.

[4] G. O. Consortium, "Gene ontology: tool for the unification of biology," *Nature Genetics,* vol. 25, pp. 25-29, 2000.

[5] I. H. G. S. Consortium, "Initial sequencing and analysis of the human genome," *Nature,* vol. 409, 2001.

[6] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *PNAS,* vol. 95, pp. 14863-14868, December 8, 1998.

[7] M. Zhang, W. Zhang, H. Sicotte, and P. Yang, "A New Validity Measure for a Correlation-Based Fuzzy C-means Clustering Algorithm," in *Proceeding the 31st Annual International Conference of IEEE Engineering in Medicine and Biology Society*, Minneapolis, 2009.

[8] Q. Brian, P. Meeyoung, and H. Jun, "Biological pathways as features for microarray data classification," presented at the Proceeding of the 2nd international workshop on Data and text mining in bioinformatics, Napa Valley, California, USA, 2008.

[9] D. Dembele and P. Kastner, "Fuzzy C-means method for clustering microarray data," *Bioinformatics,* vol. 19, pp. 973-980, May 22, 2003.

[10] R. Peter, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.,* vol. 20, pp. 53-65, 1987.

[11] H. Jeffrey, B. Michael, and O. Vadim, "A Tour through the Visualization Zoo," *Queue,* vol. 8, pp. 20-30.

[12] P. Greet, S. Anja, and J. R. Peter, "Displaying a clustering with CLUSPLOT," *Comput. Stat. Data Anal.,* vol. 30, pp. 381-392, 1999.

[13] "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature,* vol. 455, pp. 1061-1068, 2008.

[14] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, S. F. and the rest of the SBML Forum, "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics,* vol. 19, pp. 524-531, March 1, 2003 2003.

[15] *Kyoto Encyclopedia of Genes and Genomes*, http://genome.jp/kegg.

[16] P. Yang, Z. Sun, M. J. Krowka, et al., "Alphal-antitrypsin deficiency carriers, tobacco smoke, chronic obstructive pulmonary disease, and lung cancer risk," *Arch Intern Med,* vol. 168, pp. 1097-1103, 2008.

[17] K. V. Ballman, D. E. Grill, A. L. Oberg, and T. M. Therneau, "Faster cyclic loess: normalizing RNA arrays via linear models," *Bioinformatics,* vol. 20, pp. 2778-86, 2004.