

Integration of Cluster Ensemble and EM based Text Mining for Microarray Gene Cluster Identification and Annotation*

Xiaohua Hu, Xiaodan Zhang, Xiaohua Zhou, Daniel Wu

Abstract—Generating high quality gene clusters and identifying the underlying biological mechanism of the gene clusters are the important goals of clustering gene expression analysis. To get high quality cluster results, most of the current approaches rely on choosing the best cluster algorithm whose design biases and assumptions meet the underlying distribution of the data set. There are two issues for this approach: (1) usually the underlying data distribution of the gene expression data sets is unknown, and (2) there are so many clustering algorithms available and it is very challenging to choose the proper one. To provide a textual summary of the gene clusters, the most explored approach is the extractive approach that essentially builds upon techniques borrowed from the information retrieval, in which the objective is to provide terms to be used for query expansion, and not to act as a stand alone summary for the entire document sets. Another drawback is that the clustering quality and cluster interpretation are treated as two isolated research problems and are studied separately. But cluster quality and cluster interpretation are closely related and must be addressed in a coherent and unified way. It is essential to have relatively high quality clusters first, in order to get a correct, informative biological explanation of the gene cluster, otherwise, the biological explanation will be incorrect or misleading, no matter how good or robust the text summarization technique is. Based on this consideration, we design and develop a unified system **GE-Miner** (Gene Expression Miner) to address these challenging issues in a principled and general manner by integrating cluster ensemble, text clustering and multi document summarization and provide an environment for comprehensive gene expression data analysis. We present a novel cluster ensemble approach to generate high quality gene cluster. In our text summarization module, given a gene cluster, our language modeling based EM algorithm can automatically identify subtopics and extract most probable terms for each topic. Then, the extracted top k topical terms from each subtopic are combined to form the biological explanation of each gene cluster. Experimental results demonstrate that our system can obtain high quality clusters and provide informative key terms for the gene clusters.

I. INTRODUCTION

Huge amounts of gene expression data have been generated as a result of the Human Genomic project, which creates a need and challenge for data mining. Clustering algorithms are used as essential tools to analyze gene expression data sets and provide valuable insight on various aspects of the genetic machinery such as identifying

the functionality of genes, finding out what genes are co-regulated, distinguishing the important genes between abnormal tissue and normal tissues, etc [5, 34]. Generating high quality gene clusters and identifying the underlying biological mechanism of the gene cluster are the ultimate goal of clustering gene expression analysis. Some efforts and progress have been made towards this goal [1, 3, 11, 13, 19, 23, 27, 31].

But there are some drawbacks of these approaches. To get high quality cluster results, these approaches rely on choosing the best cluster algorithm whose design biases and assumptions meet the underlying distribution of the data set. Otherwise, the results will be poor if the assumptions are violated in a data set. There are two issues for this approach: (1) usually the underlying data distribution of the gene expression data sets is unknown, and (2) there are so many clustering algorithms available. It is a challenging and daunting task for genomic researchers to choose the best one for a particular gene expression data set because results of different clustering algorithms are inconsistent. K-Means, Self-Organizing Map (SOM), Hierarchical approaches, Fuzzy C-Means, etc, are very different in some cases [18]. This is because clustering methods have their own biases and function criterion. It is well known that no single clustering algorithm that performs best across various data sets.

On the other hand, clustering indeed reveals potentially meaningful relationships among genes, but cannot explain the underlying biological mechanisms. To provide a textual summary of the gene clusters, the most explored approach is currently the extractive approach that essentially builds upon techniques borrowed from the information retrieval field such as term re-weighting and relevance feedback [24; 26], in which the objective is to provide terms to be used for query expansion, and not to act as a stand alone summary for the entire document sets. Another drawback is that the clustering quality and cluster interpretation are treated as two isolated research problems and are studied separately. But cluster quality and cluster interpretation are closely related and must be addressed in a coherent and unified way. It is essential to have relatively high quality clusters first, in order to get a correct, informative biological explanation of the gene cluster. Otherwise, the biological explanation will be incorrect or misleading, no matter how good or robust the text summarization technique is. Based on this consideration, this paper explores the first step toward dealing with these issues. We design and develop a unified system **GE-Miner** (Gene Expression Miner) to address these challenging issues in a principled and general manner by integrating cluster ensemble and text summarization and provide an environment for comprehensive gene expression data

* Xiaohua Hu, Xiaodan Zhang, Xiaohua Zhou, Daniel Wu, Illhoi Yoo, Xuheng Xu are with College of Information Science and Technology, Drexel University, Philadelphia, PA 19104, USA (Email: thu@cis.drexel.edu)

This work is supported in part by NSF Career grant (NSF IIS 0448023), NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196), and PA Dept of Health Grant (No. 239667).

analysis. The task of establishing a unifying framework for comprehensive gene expression analysis is accomplished in three steps:

- 1) *Cluster Ensemble*: building a cluster ensemble method to combine the clustering results from various clustering algorithms in order to obtain high quality and robust results,
- 2) *Data Integration Server*: developing an extendable data integration server to gather related textual resource from various databases of the genes,
- 3) *Textual Summarization*: integrating biomedical literature mining in gene expression analysis to provide informative biological explanation of the gene clusters.

The rest of the paper is organized as follows. In Section 2, we give a brief overview of the architecture of **GE-Miner** and then we present the cluster ensemble method in Section 3. We discuss the data integration server in Section 4 and text summarization in Section 5. The experimental results are presented in Section 6. We conclude with our future plan and discussion in Section 7.

II. OVERVIEW OF GE-MINER ARCHITECTURE

Obtaining high quality clustering results is very challenging because of the inconsistency of the results of different clustering algorithms and the information contained in microarray data is limited by the number of arrays, their quality, noise and experimental errors. Another significant limitation of the current clustering approach is that most of these algorithms provide no biological interpretation of the cluster results, the users need to discover and interpret the biological similarities that may underlie the expression pattern by cross-referencing the experimental results in related literature or functional annotations in various genomic databases. Since gene cluster may include dozens or even hundreds of different genes, it is beyond the limits of biological researchers to detect and organize these data along multiple lines of conceptual similarity by inspection them manually. Thus, it is essential to develop a system capable of gathering biological information and extracting and summarizing relevant information in a well-organized and coherent manner for the gene cluster.

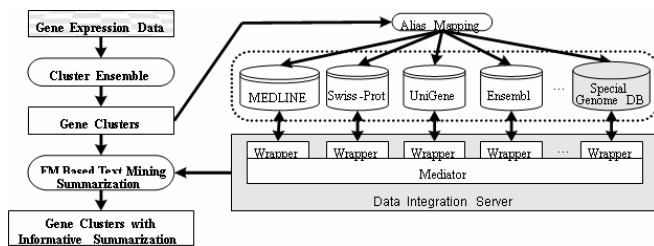


Fig. 1. The Architecture of Gene Expression Miner (GE-Miner)

We develop a comprehensive gene expression mining system **GE-Miner** (Gene Expression Miner) as shown in Figure 1 geared precisely for this task, helping a biologist in cross-referencing experimental and analytical results obtained from microarray experiments and provide concise and meaningful biological explanation of the gene clusters. **GE-Miner** aims to summarize the biomedical knowledge for

genes on a genome-wide scale, generates relevant summaries from relevant biological literatures and summarizes biological information about a group of genes in a concise and coherent manner. It has an open-architecture and can easily add a wrapper if new data sources become available, without affecting the rest of the system. The components of GE-Miner are described in details in the following sections.

III. CLUSTER ENSEMBLE

The purpose of cluster ensemble is to build a robust clustering portfolio that can perform as good as if not better than the single best clustering algorithm across a wide-range of data sets. Different clustering algorithm may take a different approach. For example, K-means is to group the data set so that the total Mean Square Error to the centers of each cluster is minimized while graph-based partitioning clustering is to partition the graph into K parts based on the minimum edge weight cuts. Thus a cluster ensemble can be used to generate many cluster results using various clustering algorithms and then integrate them using a consensus function to yield stable results.

Classification ensemble approaches such as bagging and boosting have been proved very popular and effective in supervised learning to improve the learning accuracy [10; 12]. Even though many clustering algorithms have been developed [18], not much work is done in cluster ensemble in data mining and machine learning literature compared with classification ensemble method. Zeng et al. [32] proposed an adaptive meta-clustering approach for combining different clustering results based on hierarchical clustering strategy. Strethl et al. [30] proposed a hypergraph-partitioned approach to combine different clustering results. Each cluster in an individual clustering algorithm is treated as a hyperedge. This crisp hypergraph lost much useful information, and it is not suitable for ambiguous and noisy environment such as microarray experiments.

In this section we discuss our novel cluster ensemble approach to combine the clustering results from various clustering algorithms as shown in Figure 2. We present a two-phase clustering combination strategy. At the first step, various clustering algorithms are run against the same data sets to generate clustering results. At the second step, these clustering results are combined by an auto-associative additive system based on the distance matrix of graph clustering. The diagram in Figure 2 summarizes our approach.

In our approach, a distance matrix is first constructed based on the cluster results from each individual clustering algorithm; these similarity matrices are combined to form a master distance matrix. Then a similarity graph is constructed from the master distance matrix and a graph-based partitioning algorithm is applied to the graph for the final clustering results. Graph-based clustering uses various kinds of geometric structure or graphs for analyzing data. Different graphs reflect various local structure or inherent visual characteristic in the data set

Clustering divides the graph into connected components by identifying and deleting inconsistent edges, and each subgraph consisting of connected components refers to a cluster.

A. Clustering Ensemble Algorithm

Algorithm 1: Cluster Ensemble Based on Similarity-Graph (CESG)

Input: (i) the data set $X=\{x_1, x_2, x_3, \dots, x_n\}$, (ii) edge threshold value δ , (iii) a set of different clustering algorithms $C^{(q)}$

Output: the final clustering result $C^{(opt)}$

Method:

Step 1: Run the clustering algorithm $C^{(q)}$ one at a time on the same data set

Step 2: Construct a distance matrix ($DM^{(q)}$) for the clustering results for each clustering algorithm. ($DM^{(q)}$ _{ij} represents the similarity of two data x_i and x_j points under cluster algorithm $C^{(q)}$)

Step 3: Combine the distance matrixes by adding them into one master distance matrix (MDM)

Step 4: Construct a weighted graph based on the distance matrix. (There is an edge between data point x_i and x_j if the distance value MDM_{ij} of x_i and x_j is greater than some threshold value δ , MDM_{ij} is also the weight of the edge link x_i and x_j)

Step 5: Cluster the graph into a set of clusters according to the evaluation score

In Step 2, there are so many ways to construct the distance matrix based on cluster results from individual clustering algorithm. We adapted a solution based on [32].

B. Final Stage Cluster-based distance matrix $DM(q)$ for the clustering result $C(q)$.

$DM^{(q)}$ is a pair-wise distance matrix defined between two data points according to the clustering result. The matrix size is $n \times n$. Since its size is independent of the clustering approach, it provides a way to align the different clusters onto the same space even for some situations where the numbers of clusters are different for different clustering algorithms.

We assume that probability density function of s_j is given by $p(x_i|s_j)$, the posterior probability of cluster s_j given x_i can be expressed as:

$$P(s_j | x_i) = \frac{p(x_i | s_j) \times P(s_j)}{\sum_{k=1}^m p(x_i | s_k) \times P(s_k)}, \text{ where}$$

$$P(x_i | s_j) = \frac{\exp[-\frac{1}{2}(x_i - \mu_j)^T \sum_j (x_i - \mu_j)]}{(2\pi)^{m/2} \|\sum_j\|^{1/2}},$$

m is the number of clusters. \sum_j is a matrix of co-variances among attributes in cluster j , μ_j is the mean vector of the data points in the cluster s_j .

For each data point x_i , we calculate the corresponding probability vector $PX_i = \{P(s_1/x_i), P(s_2/x_i), \dots, P(s_m/x_i)\}$, where $\sum_{j=1, \dots, m} P(s_j/x_i) = 1$, the probability vectors form a probability

space of dimension of m , with each dimension corresponding to one cluster. The probability space contains information from both the input data and the cluster results. So we believe the similarity of any two points PX_i and PX_m in the probability space is a good measurement to reflect the distance of the corresponding points x_i and x_m in the original space.

Then for any two points, x_i and x_m , in the data set, their distance is defined as the distance between PX_i and PX_m , namely, $DM^{(q)}(x_i, x_m)$. Many different distance measures such as Euclidean distance, Mahalanobis distance or correlation distance can be used to calculate $DL(PX_i, PX_m)$. We define the distance of two points (x_i and x_j) in the data set under algorithm $C^{(q)}$ as follows:

$$DM_{ij}^{(q)} = 1 - \frac{\sum PX_i PX_j - \frac{\sum PX_i \sum PX_j}{N}}{\sqrt{\left(\sum PX_i^2 - \frac{(\sum PX_i)^2}{N}\right) \times \left(\sum PX_j^2 - \frac{(\sum PX_j)^2}{N}\right)}}$$

In step 5, a graph-based clustering algorithm is applied to the weighted graph for the final clustering result. Many graph-based partitioning algorithms can be used for this purpose. We select METIS [20] for the graph partitioning because of its scalability and efficiency.

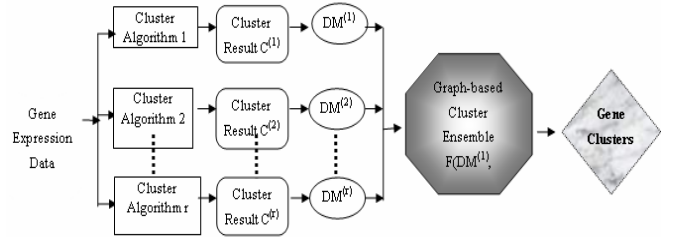


Fig. 2. Data Flow of Cluster Ensemble

C. Clustering Result Evaluation

To evaluate the quality of cluster is a non-trivial and often ill-posed task. Generally speaking, there are internal criteria and external criteria. Internal criteria formulate quality as a function of the given data and/or similarities. External criteria on the other hand impose quality by additional, external information not given to the clusters, such as category labels. This is sometimes more appropriate since groupings are ultimately evaluated externally by humans. For example, when objects have already been categorized by an external source, i.e., when class labels are available, we can use information theoretical measure to quantify the match between the categorization and the clustering. In our cluster ensemble, external criteria fit very well with our architecture. We use the Minkowski Score [4] as our cluster quality indicator. Below is our formula for the clustering quality evaluation.

A clustering solution for a set of n elements can be represented by an $n \times n$ matrix C where $C_{ij} = 1$ iff x_i and x_j are in the same cluster according to the solution and $C_{ij} = 0$ otherwise. A measure of Minkowski Score (MS) between the clustering results $C^{(h)}$ from a particular clustering algorithm CA_h with a reference clustering T (or alternatively, the true

clusters if the cluster information in the data set is known in advance) is defined as

$MS(T, C^{(h)}) = \|T - C^{(h)}\| / \|T\|$, where $\|T\| = \sqrt{\sum_i \sum_j T_{ij}}$. The Minkowski score is the normalized distance between the two matrices. Hence a perfect solution will obtain a score zero, and the smaller the score, the better solution.

We abbreviate the set of cluster groupings from r different clustering algorithms as $\Psi = \{C^{(q)} | q \in \{1, \dots, r\}\}$. The average MS score of combined clustering result C with the Ψ is defined as

$$MS^{(ANMI)}(C, \Psi) = \frac{1}{r} \sum_{q=1}^r MS(C, C^{(q)})$$

D. Yeast gene data set

There are 6221 genes in the data sets but not every gene is classified into a certain function family. In our experiment we considered the genes in a function family as one cluster and created 6 data sets (cluster 2, 3, 4, 5, 6, 7). Table I shows 6 function families of yeast gene and how we construct the six data sets (C2, C3, C4, C5, C6, and C7) for our cluster ensemble comparison. For example, ‘‘C3’’ means the cluster set has 3 clusters (ATP synthesis, mitosis, and vacuolar protein targeting here)

TABLE I
SOME OF YEAST GENE FUNCTION FAMILY

Function Families	# of genes	Cluster Sets		
ATP synthesis	19	C3		
mitosis	19			
vacuolar protein targeting	19			
silencing	20	C5		
fatty acid metabolism	20			
meiosis	21			
phospholipid metabolism	21			
TCA cycle	22			
protein processing	27	C4	C6	C7
DNA repair	29			
protein folding	30			
nuclear protein targeting	31			
signaling	31			
major facilitator superfamily	32			
mRNA splicing	34	C2		
chromatin structure	42			
DNA replication	42			

TABLE II
CLUSTERING RESULTS OF YEAST GENE DATA SETS

Cluster set #	K-means	SOM	Fuzzy C-means	Cluster Ensemble
C2	0.902	0.995	0.993	0.986
C3	0.890	0.931	0.941	0.728
C4	1.180	1.194	1.170	1.071
C5	1.207	1.241	1.229	1.059
C6	1.288	1.355	1.280	1.192
C7	1.326	1.301	1.284	1.196

Table II shows the clustering results including cluster ensemble in Minkowski scores (MS) for each cluster set. As clearly indicated by the MS values of the clusters, the cluster

ensemble method made significant improvement of quality of the clustering results over the individual clustering algorithm on all the six gene data sets. For example, the best individual clustering algorithm for C3 is K-means ($MS=0.890$), while the cluster ensemble has $MS=0.728$. For C5, the best individual clustering algorithm is SOM ($MS=1.241$) and the cluster ensemble reduced them to $MS=1.059$.

IV. DATA INTEGRATION SERVER: COMPILATION OF THE KNOWLEDGE/INFORMATION OF THE GENE CLUSTERS

There are a variety of biological databases that can be mined to find possible functional relationships between genes in a cluster. For example, biomedical literature databases such as PubMed, which are a rich source of information, can be used to discover and analyze significant biological information on a genome-wide scale. We collect and compile several sources for textual annotations of the gene clusters. First, we retrieve the gene descriptions from UniGene and the corresponding genome databases of some species through Ensemble (for example, mouse, human, etc). Second, proteins that are the products of the given genes are also of interest. Hence, it is important to know the proteins made by genes. We use SWISS-PROT, a curate protein sequence database. It serves as an extended textual resource for the genes. But this information is often insufficient and bibliographic information must be consulted by following the links to select a PubMed abstract provided in some sequence databases. Since only a small fraction of these pointers provide direct information about gene function, further references are usually collected by querying PubMed¹ directly with gene names and their synonyms. Given the complexity and growth of biomedical literature, we need a system such as our GE-Miner capable of filtering the literature database and extracting and summarizing relevant information in a well-organized and coherent manner. In GE-Miner, we developed many built-in wrappers for data retrieval for various data sources. For each gene in gene expression experiment, a list of ‘‘aliases’’ is retrieved from LocusLink². The GE-Miner can support any source for alias specification through a customized wrapper. Each query retrieves documents for a single gene that are associated with one or more of the aliases for the gene. The key terms in the gene expression stop word list are removed from the abstracts. For more details about the data integration server, please refer to [38].

Gene Expression Stop Word List. According to [6], in the enriched gene expression data sets, each gene will have words and expressions in common including: (1) standard English words, such as ‘‘the’’, ‘‘experiment’’, (2) words with general biological meaning, such as ‘‘gene’’, ‘‘high-throughput technology’’, ‘‘hybridization’’, ‘‘base sequence’’, (3) specific words and phrases such as cell cycle, glucose, kinase and DNA replication. It is the last set which may be considered specific to the group of genes. The textual information is used to extract those words, which have

¹ <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>

² <http://www.ncbi.nlm.nih.gov/LocusLink>

significant frequency and specificity for each group. We need to create a gene expression stop list to remove those common English words and general biological terms from the text associated with the gene clusters. Following the method proposed in [19], we create the list by query PubMed with a large random collection of genes so that an equal number of genes were chosen from different classes of organisms like Eukaryotes, Prokaryotes and Chordates etc. Then we extract the key phrase [28] for this list and rank them in descending order of their total frequency. The distribution of the list followed Zipf's law which is commonly seen in word-frequency distribution [35]. We consider the first region with high frequency of Zip's curve. The remaining terms from this region constituted the key phrase stop word list.

V. LANGUAGE MODELING BASED EM INFORMATIVE TEXTUAL SUMMARIZATION OF GENE CLUSTERS

A common approach to information retrieval from the biomedical literature is the use of "keywords" representing the essential concepts contained within a text. A variety of approaches to provide a biological explanation of gene clusters have been developed. TextQuest [17] is geared to summarize documents retrieved in response to a keyword(s) based search on PubMed. It does not retain the association between the genes (keywords) and the retrieved documents. MedMiner [31] can provide summarized literature information on genes but it is limited to finding relations between two genes only and also, it returns a few hundreds sentences. Shatkay et al. [27] suggested a system, which attempts to find functional relations among genes on genome-wide scale, but this requires user to specify a representative document for each gene, which describe the gene very well. Looking for the representative document may need a lot of time, effort and knowledge on part of user. Also as genes have multiple biological functions, it is very rare to find a document that covers all aspect of gene across various biological domains. GEISHA [6] is based on a comparison of the frequency of abstracts linked to different gene clusters and containing a given term. Interpretation by the end user of the biological meaning of the terms is facilitated by embedding them in the corresponding significant sentences and abstracts and by establishing relations with other, equally significant terms.

Standard text summarization algorithms [19; 23] are geared to summarizing all the documents retrieved based on keyword searches. In multi-keyword searches, the association between a document and the keywords(s) is not used. However, in our context, interest lies in capturing significant biological properties that are most relevant to the cluster as a whole.

Our approach takes each gene cluster as a major topic. Our approach takes each gene cluster as a major topic. Each major topic is summarized as several subtopics. Each subtopic contains certain number of documents and terms. A language model based algorithm with semantic support is applied to extract topical functional terms from each cluster. These

functionality terms can serve as meaningful functionality explanation of each gene cluster.

A. Text Classification and Summarization

In this section, a semantic language modeling approach is developed to extract functional annotation of each gene cluster. Language modeling is first introduced by Ponte and Croft in [36] applied in text retrieval and later improved by Zhai [37] using relevance feed back through a generative approach. The relative simplicity and effectiveness of the language modeling approach, together with the fact that it leverages statistical methods that have been developed in speech recognition and other areas, make it an attractive framework [36] for not only text retrieval but also theme detection across collections by Zhai [38].

However, as all these works are based on unigram language model without domain ontology support. Especially in biomedical domain, as for protein functionality, with unigram language model, it's very difficult to differentiate functionality terms from other terms. To overcome this problem, in this paper, we fit UMLS domain ontology in Zhai's relevance feed back model. In this way, we can detect most common functionality terms shared by each cluster, while filtering out most of non-related terms.

B. A generative model

A generative model proposed in [38] is used to generate topical terms for each gene cluster and remove the background noise—terms that are generated according to the whole collection. Assume the set of documents containing term w is generated by a mixture model (i.e., interpolating the topical model with the background collection model θ_{B_c}):

$$p(w) = (1 - \lambda_{B_c}) \sum_{j=1}^n \pi_{d_j} p(w | \theta_j) + \lambda_{B_c} p(w | \theta_{B_c})$$

where λ_{B_c} is a coefficient accounting for the background noise. Then we obtain a translation model for term w using an Expectation Maximization (EM) [39] algorithm with the following update formulas:

$$p(y_{d,w} = j) = \frac{\pi_{d,j}^{(n)} p^{(n)}(w | \theta_j)}{\sum_{j'} \pi_{d,j'}^{(n)} p^{(n)}(w | \theta_{j'})}$$

$$p(y_{d,w} = B_c) = \frac{\lambda_{B_c} p(w | \theta_{B_c})}{\lambda_{B_c} p(w | \theta_{B_c}) + (1 - \lambda_{B_c}) \sum_{j'} \pi_{d,j'}^{(n)} p^{(n)}(w | \theta_{j'})}$$

$$\pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w, d) p(z_{d,w} = i, j)}{\sum_{j'} \sum_{w \in V} c(w, d) p(z_{d,w} = i, j')}$$

$$p^{(n+1)}(w | \theta_j) = \frac{\sum_{d \in C} c(w', d) (1 - p(y_{d,w'} = B_c)) p(y_{d,w'} = j)}{\sum_{w' \in V} \sum_{d \in C} c(w', d) (1 - p(y_{d,w'} = B_c)) p(y_{d,w'} = j)}$$

where $c(w, D_k)$ is the frequency count of term w in document d , and θ_j is the j^{th} topic model.

In this case, a term is generated according to a mixture language model. For a generative process, the term is either generated according to some topical model or collection model. It's assumed that most terms are generated according to background model. Experimentally, we find setting λ_{B_c} to 0.9 make the best biomedical sense for gene cluster annotation. which is in charge of generating most common term across the documents. One of drawbacks of background model is that it may generate some common functionality terms. To solve this problem, in stead of using the complex collection, we randomly download ten percent of PubMed documents of 2005 serving as background collection.

C. Experimental Results

The input data set for gene cluster summarization is the relevant document sets of each gene cluster. In Data Integration Server) the input document sets of each gene cluster are generated in the following procedurals.

1. For each gene, the synonyms are searched. If found, the synonym(s) are added to gene list
2. Relevant documents are fetched from PubMed. The search keyword has this format; Gene name AND (Model organism names). For example, for the yeast genes used in our experiment described in Section 6, the number of genes (also synonym) and the related documents are shown in Table III.
3. In practice, we first use cross-validation method to judge the number of subtopics of each gene cluster, and then apply the generative language mixture model presented in section B above to generate topical terms for each subtopic. Finally, we apply UMLs ontology to filter out those terms that are not functionality terms. It should be noted that NLP technique and UMLs ontology are applied to extract medical terms from each abstract. Thus, the representation of the document is basically different traditional unigram model. In UMLs ontology, each term has a concept ID and each concept ID has a series of semantic types. Thus terms without functionality semantic types are filtered out in the last step.

TABLE III
GENE CLUSTER DATASETS

Gene cluster #	# of genes in the cluster (including synonyms)	# of relevant PubMed documents	# of subtopics for each gene cluster
1	19 (25)	157	3
2	19 (35)	820	4
3	19 (69)	417	5
4	20 (30)	1043	9
5	20 (34)	641	5

6	21 (35)	644	3
7	21 (31)	374	4
8	22 (30)	321	2
9	42 (67)	1180	6
10	42 (75)	2769	11

We conduct some experiment study on yeast gene data set (<http://rana.lbl.gov/EisenData.htm>). The reason we choose the Yeast DNA microarray is because the validity of our methods is best assessed by comparison of the results with existing summaries of biological information. The Saccharomyces Genome Databases [8] and the Yeast Proteome Database [9], as well as the functional analysis given by Spellman et al., [29], are critical for objective evaluation of our results. There are 6221 genes in the data sets but not every gene is associated with certain function family. Yeast gene has a lot of function families. Table IV shows the function family name of each yeast gene cluster of Table III. Table V shows the summarization results of 9 gene clusters. As there are 11 subtopics for gene cluster 10, we would not list the results for it. We found we catch most of related terms to each gene cluster according to its function families, which indicates the results are robust and promising.

TABLE IV
10 YEAST GENE FUNCTION FAMILY

Gene cluster #	Function Families
1	ATP synthesis
2	mitosis
3	vacuolar protein targeting
4	silencing
5	fatty acid metabolism
6	meiosis
7	phospholipid metabolism
8	TCA cycle
9	chromatin structure
10	DNA replication

TABLE IV
GENE CLUSTER AND ITS SUBTOPIC SUMMARIZATION

Gene cluster #	Sub topics #	Top 10 key terms
1	1	mutant; Mutation; cell assembly; Biologic Transport; mitochondrial protein import; Oxidative Phosphorylation; F1-ATPase; N-terminal binding; defects; Growth;
	2	F1-ATPase; mutant; gamma protein; Oxidative Phosphorylation; Disruption; cell assembly; Localization; N-terminal binding; translation; Growth;
	3	mutant; Mutation; Growth; Disruption; cell assembly; defects; ATPase activity; C-terminal binding; inhibitors; septation;
2	1	degradation; E3; Cell Cycle; ubiquitin; proteolysis; protein ubiquitination; regulation; cyclin; Progression; mutant;
	2	mutant; Mitoses; Cell Cycle; dynein; protein activation; response to acid; Progression; Localization; defects; Cytokineses;
	3	mutant; Mutation; regulation; protein

		activation;Transcription;DELETION;Growth;Cytokines;response to acid;N signaling pathway;
	4	Mutation; mutant; Growth; regulation; Cell Cycle; Antibody Affinity; degradation; Localization; gene complementation; uptake;
3	1	ubiquitin; Mutation; mutant; Biologic Transport; Localization; Ventricular Extrasystole; protein ubiquitination; protein protein interaction; defects; uptake;
	2	mutant; Growth; defects; phosphatase; protein protein interaction; drug accumulation; Cell Cycle; ligand; Mutation; protein activation;
	3	SNARE;protein protein interaction; mutant; Mutation; Localization; ligand; Conformational change;vacuolar assembly; C-terminal binding; gene complementation;
	4	Biologic Transport; mutant; protein protein interaction; defects; Mutation; protease; trafficking; Transcription; ligand; drug accumulation;
	5	mutant; defects; Mutation; secretory pathway; gene complementation; Biologic Transport; Immunity; Endocytosis; Recombination; ligand;
4	1	mutant; Chromosome Segregation; Cytokines; chromatin silencing; DELETION; Growth; cell growth; Gene Expression; S Phase; phosphatase;
	2	Mutation; mutant; Transcription; S Phase; Sterile; defects; Gene Silencing; PL; mating; M Phase;
	3	protein protein interaction; Transcription; sporulation; protein activation; Apoptosis; mutant; ligand; Growth; Cell Survival; response to acid;
	4	mutant; Localization; Diastasis; sporulation; Recombination; protein protein interaction; Dynamic; defects; conjugation; Cell Division;
	5	inhibitors; cleavage; deacetylase activity; histone deacetylation; mutant; Metabolic Process; DNA Repair; double-strand break repair via nonhomologous end-joining; Genome Stability; Apoptosis;
	6	response to acid; mutant; Aging; Growth; Cell Aging; drug accumulation; Cell Survival; PL; DELETION; protein activation;
	7	DNA Damage; response to acid; Gene Expression; protein modification; protein activation; protein ubiquitination; Transcription; S Phase; inhibitors; DNA Repair;
	8	Apoptosis; Genetic Nondisjunction; DNA Damage; Cell Survival; S Phase; Cell Aging; Embryogenesis; Localization; defects; protein protein interaction;
	9	Recombination; mutant; DELETION; defects; RFB binding; Transcription; replication; telomere maintenance; PL; N-terminal binding;
5	1	mutant; cell elongation; Mutation; regulation; Disruption; single-stranded telomeric DNA binding; PL; defects; N-terminal binding; telomeric DNA binding;
	2	inhibitors; Localization; Metabolic Process; conjugation; Apoptosis; Step; Agent; drug accumulation; response to acid; protein modification;
	3	mutant; DELETION; Transcription; Retardation, Mental; drug accumulation; Disruption; Order; gene induction; Step; Point Mutation;
	4	protein activation; enzyme activity; antibody; Step; Metabolic Process; Order; Detergents; acyltransferase activity; Antibody-Dependent Enhancement; Hypertrophies;
	5	protein activation; enzyme activity; antibody; Step; Metabolic Process; Order; Detergents; acyltransferase activity; Antibody-Dependent Enhancement; Hypertrophies;
6	1	Recombination; mutant; Transcription; Mutation; double-strand break repair; sporulation; protein

		activation; Growth; defects; Mitoses;
	2	mutant; Mutation; response to acid; DELETION; protein activation; Growth; defects; DNA Damage; homologous recombination; Virulence;
	3	protein activation; MEK; ligand; Growth; protein kinase cascade; Mutation; Gene Expression; Cell Cycle; defects; EGF;
7	1	regulation; Growth; mutant; response to acid; Gene Expression; Transcription; protein activation; Base; defects; phospholipase B;
	2	mutant; defects; gene complementation; Growth; Mutation; protein protein interaction; Protein Binding; PL; Distribution; Disruption;
	3	mutant; regulation; Growth; serine exchange enzyme; PL; DGPP phosphatase; response to acid; uptake; drug accumulation; phosphatase;
	4	protein protein interaction; mutant; phospholipase B; Disruption; regulation; response to acid; Transcription; Growth; protein activation; Mutagenesis;
8	1	protein protein interaction; mutant; phospholipase B; Disruption; regulation; response to acid; Transcription; Growth; protein activation; Mutagenesis;
	2	mutant; DELETION; regulation; degradation; Mutation; ubiquitin; Cell Survival; Morphogenesis; carbon catabolite repression; derepression;
9	1	mutant; Mutation; Transcription; DELETION; protein protein interaction; Cell Survival; DNA Damage; Growth; gene induction; histone acetylation;
	2	protein protein interaction; histone acetylation; Transcription Activation; Chromatin Remodeling; Step; inhibitors; PL; gene induction; Gene Activation; histone acetyltransferase activity;
	3	mutant; Mutation; protein protein interaction; ligand; Transcription; PL; Suppression; DELETION; protein heterodimerization activity; Embryogenesis;
	4	Mutation; TyI element transposition; Cell Cycle; Localization; Mutagenesis; replication; PL; Recombination; mutant; Organization;
	5	regulation; mutant; histone acetylation; histone deacetylase activity; PL; protein amino acid deacetylation; protein protein interaction; Virulence; Gene Silencing; Gene Expression;
	6	protein protein interaction; Transcription; protein activation; Hypoxia; Transcription Activation; protein ubiquitination; Mutation; N-terminal binding; protein modification; DNA binding;

VI. CONCLUSION

In this paper we present a novel system GE-Miner for comprehensive gene expression analysis. Our system integrate cluster ensemble, text summarization and the experiment results on yeast gene expression data indicate that the GE-Miner can generate better quality and robustness clusters and provide informative term summary for the gene clusters. Moreover, we provide a language modeling based EM multi document summarization method with domain ontology support. In practice, the algorithm can automatically identify sub topics and assign most probable terms to each subtopic, which are combined to form the biological explanation of each gene cluster. Clustering ensemble is a new and very promising research area. There are a lot of open problems for future research. We plan to expand our ensemble approach to integrate feature selection for

clustering very high dimensional data set and add some inference mechanism to automatically infer valid information from the clustering results. Text summarization for gene literature has attracted a lot of attention recently and one of the challenging issue is how to get relevant literature from the huge and diversified literature because a gene tends to have many alias name and there is no standard name convention. In our future research, we plan to apply data mining techniques to automatically find gene synonyms to enhance the precise rate of the retrieved literature in order to get more relevant textual information. We hope to report our findings in the near future

REFERENCES

- [1] F. Azuaje and N. Bolshakova., *Clustering Genome Expression Data: Design and Evaluation Principles*, in *Understanding and Using Microarray Analysis Techniques: A Practical Guide*, Berrar D, Dubitzky W and Granzow M, editors, Springer Verlag, 2002.
- [2] A. Bellaachia, D. Portnoy, Y. Chen and A. G. Elkahoulou, *E-CAST: A Data Mining Algorithm For Gene Expression Data*, The 2nd Workshop on Data Mining in Bioinformatics (BIOKDD 2002), 49-54, 2002.
- [3] A. Ben-Dor, Z. Yakhini, *Clustering Gene Expression Patterns*, Proc. ACM RECOMB, 33-43
- [4] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981
- [5] D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach To Microarray Data Analysis*, Kluwer Academic Publishers, 2002.
- [6] C. Blaschke, J.C. Oliveros and A. Valencia. *Mining Functional Information Associated With Expression Arrays*. *Funct Integr Genomics*, 1(4): 256-268, 2001.
- [7] D. Boley, M. Gini, R. Gross, E. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore, *Partitioning-Based Clustering For Web Document Categorization*, *Decision Support Systems*, 27:329-341, 1999
- [8] J.M. Cherry et al., *SGD: Saccharomyces Genome Database*. *Nucleic Acids Res* 26:73-79
- [9] M.C. Costanzo et al., *The Yeast Proteome Database (YPD) And Caenorhabditis Elegans Proteome Database (Wordpd): Comprehensive Resources For The Organization And Comparison Of Model Organism Protein Information*, *Nucleic Acids Res* 28:73-76
- [10] T.G. Dietterich, *Ensemble Methods In Machine Learning*. In J. Kittler and F. Roli, editors, *Multiple Classifier Systmes*, page 1-15, LNCS Vol 1857, Springer, 2001
- [11] P. Glenisson, P. Antal, J. Mathys, Y. Moreau and B.D. Moor., *Evaluation Of The Vector Space Representation In Text-Based Gene Clustering*. Pacific Symposium on Biocomputing. 391-402, 2003.
- [12] X. Hu, *Using Rough Sets Theory And Database Operations To Construct A Good Ensemble Of Classifiers For Data Mining Applications*. IEEE ICDM 2001: 233-240
- [13] X. Hu and I. Yoo, *Scalable Learning Method To Extract Biological Information from Huge Online Biomedical Literature*, Chapter 23 in *Computational Web Intelligence: Intelligent Technology for Web Applications*, in Y. Zhang et al (Eds), World Scientific Publisher, in print , 2004a
- [14] X. Hu X., J. Han and N. Cercone, *Discovering of Cyber Communities from the WWW, to appear in Proc. of COMPSAC 2003* , Dallas, TX, Nov. 3-6, 2003
- [15] Hu X., *Integration of Cluster Ensemble and Text Summarization for Gene Expression Analysis*, in Proceedings of the IEEE 2004 Symposium on Bioinformatics and Bioengineering, 251-259, May 19-21, 2004, Taiwan (IEEE BIBE 2004)
- [16] L. Hubert and J. Schultz, *Quadratic Assignment As A General Data-Analysis Strategy*, *British Journal of Mathematical and Statistical Psychology*, 29, pp. 190-241, 1976
- [17] I. Iliopoulos, A.J. Enright and C.A. Ouzounis, *TextQuest: Document Clustering of MEDLINE Abstract For Concept Discovery In Molecular Biology*. Pacific Symposium of Biocomputing 2001, 384-395, 2001.
- [18] K. Jain and M. N. Murty and P. J. Flynn, *Data Clustering: A Review*, *ACM Computing Surveys*, Vol. 31, No. 3, September 1999.
- [19] P. Kankar, S. Adak, A. Sarkar, K. Murari and G. Sharma, *Medmesh Summarizer: Text Mining For Gene Clusters*. In Proceedings of the Second SIAM International Conference on Data Mining, 2002.
- [20] G. Kayypis and V. Kumar. *Multilevel k-way Partitioning Scheme for Irregular Graphs*. *Journal of Parallel and Distributed Computing*
- [21] J. Kleinberg, *Authoritative sources in a hyperlinked environment*. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. Extended version in *Journal of the ACM* 46(1999).
- [22] T. Kohonen, *Self-Organizing Maps*. Springer, New York, 3rd edition, 2000
- [23] D.R. Masys, J.B. Welsh, J.L. Fink, M. Gribskov, I. Klacansky and J. Corbell, *Use Of Keywords Hierarchies To Interpret Gene Expression Patterns*. *Bioinformatics*, 17(4): 319-326, 2001.
- [24] E. Mittendorf, *Using The Co-Occurrence Of Words For Retrieval Weighting*, *Information Retrieval*, 3, 243-251, 2000
- [25] [25] D. Pelleg and A. Moore, X-means: Extending K-means with efficient Estimation of the number of cluster, in *Proceedings of the 17th International Conference on Machine Learning*, 2000
- [26] [26] G. Salton and C. Buckley, *Term-Weighting Approaches In Automatic Information Text Retrieval*. *Information Process Manage*, 24(5): 513-523, 1988
- [27] H. Shatkey, S. Edwards, W.J. Wilbur and M. Boguski. *Genes, Themes And Microarrays: Using Information Retrieval For Large-Scale Gene Analysis*. 8th Int. Conf. on Intelligent Systems Mol. Bio. (ISMB 2000), 8: 317-328, La Jolla, August 2000.
- [28] M. Song, I-Y Song, X. Hu, *KPSpotter: A Flexible Information Gain-based Keyphrase Extraction System*, accepted in the ACM CIKM 5th Workshop on Web Information and Data Management (WIDM 03), New Orleans, LA, Nov. 3-8, 2003
- [29] P.T. Spellman et al, *Comprehensive Identification Of Cell Cycle-Regulation Genes Of The Yeast Saccharomyces Cerevisiae By Micorarray Hybridizatio*. *Molecular Biology of the Cell*:9:3273-3297
- [30] A. Strehl and J. Ghosh, *Cluster Ensembles - A Knowledge Reuse Framework For Combining Multiple Partitions*, *Journal on Machine Learning Research (JMLR)*, 3:583-617, December 2002.
- [31] L. Tanabe, U. Scherf, L.H. Smith, J.K. Lee, L. Hunter and J.N Weinstein. *MedMiner: An Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling*. *Biotechniques*, 27(6): 1210-1217, 1999.
- [32] Zeng, J. Tang, J. Garcia-Frias, and G.R. Gao, *An Adaptive Meta-Clustering Approach: Combining The Information From Different Clustering Results*, CSB2002 IEEE Computer Society Bioinformatics Conference Proceedings August 14-16 Stanford University, page 276-287.
- [33] H Zha, *Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering*, in ACM SIGIR 2002, Aug 11-15, Finland
- [34] Y. Zhao and G. Karypis, *Clustering in Life Science*, In "Functional Genomics", Arkady Khodursky and Michael Brownstein (editors)., 2003
- [35] G.K. Zipf, *Psycho-Biology of Languages*. Houghton-Mifflin, 1935.
- [36] J. Ponte and W. B. Croft. A language modeling approach to information retrieval. In 21st ACM SIGIR Conference on Research and Development in Inormation Retrieval (SIGIR'98), pages 275-281, 1998.
- [37] Chengxiang Zhai and John Lafferty, Model-based feedback in the language modeling approach to information retrieval, Tenth International Conference on Information and Knowledge Management (CIKM 2001), 2001
- [38] ChengXiang Zhai, Atulya Velivelli, Bei Yu, A cross-collection mixture model for comparative text mining, Proceedings of ACM KDD 2004 (KDD'04), pages 743-748, 2004.
- [39] Dempster, A. P., Laird, N. M., & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39 (1), 1-38. 1977
- [40] Hu X., *Gene-Miner: Integration of Cluster Ensemble and Text Mining for Comprehensive Gene Expression Analysis*, accepted in the International Journal of Bioinformatics Research and Application