Clustering Genes using Gene Expression and Text Literature Data

Chengyong Yang, Erliang Zeng, Tao Li, and Giri Narasimhan[*]

Bioinformatics Research Group (BioRG), School of Computer Science, Florida International University, Miami, Florida, 33199, USA. E-mail: `{cyang01,ezeng001,taoli,giri}@cs.fiu.edu`

ABSTRACT

Clustering of gene expression data is a standard technique used to identify closely related genes. In this paper, we develop a new clustering algorithm, MSC (Multi-Source Clustering), to perform exploratory analysis using two or more diverse sources of data. In particular, we investigate the problem of improving the clustering by integrating information obtained from gene expression data with knowledge extracted from biomedical text literature. In each iteration of algorithm MSC, an EM-type procedure is employed to bootstrap the model obtained from one data source by starting with the cluster assignments obtained in the previous iteration using the other data sources. Upon convergence, the two individual models are used to construct the final cluster assignment. We compare the results of algorithm MSC for two data sources with the results obtained when the clustering is applied on the two sources of data separately. We also compare it with that obtained using the feature level integration method that performs the clustering after simply concatenating the features obtained from the two data sources. We show that the z-scores of the clustering results from MSC are better than that from the other methods. To evaluate our clusters better, function enrichment results are presented using terms from the Gene Ontology database. Finally, by investigating the success of motif detection programs that use the clusters, we show that our approach integrating gene expression data and text data reveals clusters that are biologically more meaningful than those identified using gene expression data alone.

1 INTRODUCTION

DNA microarray technology offers an opportunity to simultaneously measure the expression of all the genes in a given sample, at a given time, and under specific conditions. Recently, large scale microarray experiments performed under a variety of conditions or at various stages during a biological process has resulted in huge amounts of gene expression data, and has presented big challenges for the field of data mining (Ball *et al.*, 2004; Bozdech *et al.*, 2003; Spellman *et al.*, 1998). Challenges include rapidly analyzing and interpreting data on thousands of genes measured under hundreds of different conditions, and assessing the biological significance of the results. Clustering is the exploratory, unsupervised process of partitioning the expression data into groups (or clusters) of genes sharing similar expression patterns. Such co-expressed genes may suggest co-regulation and may be possibly sharing common biological function (Eisen *et al.*, 1998; Sherlock, 2000). Popular clustering methods used to analyze gene expression data include hierarchical clustering (Eisen *et al.*, 1998), K-means (Herwig *et al.*, 1999), and self-organizing maps (Tamayo *et al.*, 1999), among others. However, the quality of clusters can vary greatly, as can the ability to infer biologically meaningful conclusions.

On a different note, the biological and medical literature databases are information warehouses with a store of useful knowledge. They can be used in many ways. For example, they can be used to cross-reference experimental and analytical results with previously known biological facts, theories, and results. On the other hand, they can

---

also be used to identify functional commonalities of genes and to help drive the interpretation and organization of the expression data (Altman and Raychaudhuri, 2001). In fact, text analysis has been applied successfully to many interesting biological problems (Shatkay *et al.*, 2000; Yandell and Majoros, 2002). As shown in several papers, article abstracts can successfully predict gene function (Fleischmann *et al.*, 1999; Raychaudhuri *et al.*, 2002; Tamames *et al.*, 1998) and genes can be clustered into functionally related groups based on the text in the scientific literature databases (Chaussabel and Sher, 2002). Hence, including the literature in the analysis of gene expression data offers an opportunity to incorporate additional functional information about the genes when defining expression clusters. In more general terms, with the availability of multiple information sources, it is a challenging problem to conduct integrated exploratory analyses with the aim of extracting more information than what is possible from only a single source.

Ihmels *et al.* (Ihmels *et al.*, 2002) presented a new algorithm that used additional biological information in the form of sequence data and/or functional annotation to generate an initial gene set. The algorithm then iteratively refined the set of experimental conditions and the set of genes until stopping criteria was met, which signified the discovery of a tightly-connected cluster of genes. In a recent paper (Segal *et al.*, 2003), a generative probabilistic model for combining promoter sequence data and gene expression data was developed to extract biologically meaningful clusters (transcriptional modules) on a genome-wide scale in *S. cerevisiae*. Broadly speaking, there are two existing clustering approaches for combining gene expression data and text literature. These are exemplified by the two software packages called MedMiner (Tanabe *et al.*, 1999) and PubGene (Jenssen *et al.*, 2001). MedMiner first performed clustering on expression data and then interpreted textually while PubGene first performed clustering on textual data and then interpreted numerically. In another recent publication, Raychaudhuri et al., first applied hierarchical clustering to gene expression data, and then used text from abstracts to resolve hierarchical cluster boundaries to identify clusters that are functionally more coherent (Raychaudhuri *et al.*, 2003).

While previous approaches made use of both data types, they tended to ignore the correlation structure between different sources (Wu *et al.*, 1999). Our MSC algorithm implicitly learns the correlation structure among different data sources and provides a semantic scheme to analyze data from heterogeneous data sources.

In this paper, more specifically, we investigate the problem of integrating gene expression data and biological text literature to produce more biologically significant clusters. The problem of learning from multiple information sources has been extensively studied by the machine learning community. In computer vision this problem is referred to as multi-modal learning. In general, there are two approaches to multi-modal learning: feature level integration and semantic integration (Wu *et al.*, 1999). Methods that use feature level integration combine the information at the feature level and then perform the analysis in the joint feature space. The correlation structure between different sources can be discovered via learning (Glenisson *et al.*, 2004). On the other hand, the semantic level integration methods first build individual models based on separate information sources and then combine these models via techniques such as mutual information maximization (Becker, 1996).

In this work, we present a new clustering algorithm: MSC. The algorithm is a generalized version of the standard K-means in the sense that it allows a stochastic exploration of data obtained from multiple sources (Zhong and Ghosh, 2003). The algorithm can be also thought of as a kind of "semantic" integration of data from multiple sources. Semantic integration has several advantages over feature-level integration. First, even though the joint feature space is often more informative than that available from each of the individual sources, naïve feature integration tends to generalize poorly (Wu *et al.*, 1999). Second, the

semantic integration implicitly learns the correlation structure between different sets of features. Our experiments show that our approach performs better than methods that use feature-level integration.

This work also explores the problem of establishing good representations for literature-based information

The rest of the paper is organized as follows. In Section 2, we describe our new approach of integrating by presenting a more general algorithm applicable for multiple data sources. Section 3 introduces the information sources used in this paper and presents a novel keyword-based vector representation of literature. In Section 4, we show the performance of our new clustering approach through a typical example and comparison with other integration approaches. We conclude with some discussions in Section 5.

2 METHODS

In this section we describe our new approach. We first briefly introduce the clustering problem and then present our new clustering algorithm for combining different data types.

2.1 Clustering Fundamentals

The problem of clustering data arises in many disciplines and has a wide range of applications. Intuitively, clustering is the problem of partitioning a finite set of points in a multi-dimensional space into classes (called clusters) so that (i) the points belonging to the same class are *similar* and (ii) the points belonging to different classes are *dissimilar* (Jain and Dubes, 1988). In this paper, our goal is to identify clusters of related gene using the available datasets. The notation used in the paper is listed below in Table 1.

2.3 The MSC algorithm

Here we describe the algorithm in detail. The method extends the model-based K-means cluster algorithm to allow for combined learning of different data types. The algorithm assumes that there are m parameterized models, one for each

Table 1. Notation

n	Number of genes
m	Number of data sources
$M_1, M_2, ..., M_m$	Dimensions for each data source
k	Number of clusters
$\lambda_i^{(l)}$	Model parameters for i-th cluster on data source l
$\lambda_i = (\lambda_i^{(1)}, \lambda_i^{(2)}, ..., \lambda_i^{(m)})$	Model parameters for i-th cluster
$O = \{(o_1^{(1)}, o_1^{(2)}, ..., o_1^{(m)}), ..., (o_n^{(1)}, o_n^{(2)}, ..., o_n^{(m)})\}$	The dataset
$\Lambda = \{(\lambda_1^{(1)}, \lambda_1^{(2)}, ..., \lambda_1^{(m)}), ..., (\lambda_k^{(1)}, \lambda_k^{(2)}, ..., \lambda_k^{(m)})\}$	The cluster model
$Y = \{y_1, ..., y_n\}, y_i \in \{1, ..., k\}$	Cluster assignment vector
P	Vector denoting the weights for each data source.

cluster. The set of parameters for the i-th model is denoted by λ_i. Typically, all the models are assumed to be from the same family, e.g., Gaussian or multinomial distribution. In the sample reassignment step, a data point could be assigned to a cluster using three possible approaches: maximum likelihood (ML), soft, or stochastic (Zhong and Ghosh, 2003).

The MSC algorithm, shown in Table 2, is a variant of the EM method (Dempster *et al.*, 1977; Neal and Hinton, 1998). It stochastically builds the models for each data source by boosting the models using the cluster assignments from the other models. Let $\lambda^{(j)}$ be the set of parameters for the models of data source j. In each iteration, we first randomly select a data source j based on the weight vector P. We then perform the following steps: (i) find the model parameters $(\lambda_1^{(j)}, \lambda_2^{(j)}, ..., \lambda_m^{(j)})$ that maximize the likelihood of the data given the current cluster assignment (shown in step 2b of the algorithm); (ii) assign the data points to the cluster that maximizes the pos-

terior probability (shown in step 2c of the algorithm). Our experiments show that the MSC algorithm implicitly learns the correlation structure among the multiple data sources.

Table 2. The MSC Algorithm

Input: Data samples O, model structure Λ, and weight vector P.

Output: Trained models Λ and a partition of the data samples given by the cluster identity vector Y

1. **Initialization**: initialize the cluster identity vector Y

2. **while** stopping criterion is not met
   a. Randomly pick a data source j according to P
   b. Model re-estimation for source j: for each cluster i, let $O_i^{(j)} = \{o_s^{(j)} \mid y_s = i\}$. The parameters of the model for cluster i, $\lambda_i^{(j)}$, are re-estimated as

$$\lambda_i^{(j)} = \arg\max_{\lambda} \sum_{o^{(j)} \in O_i^{(j)}} \log P(o^{(j)} \mid \lambda^{(j)})$$

   c. Sample re-assignment: for each data sample s, set

$$y_s = \arg\max_i \log P(o_s^{(j)} \mid \lambda_i^{(j)})$$

3. **Return** Λ and Y.

2.4 Cluster Assignment

In order to obtain the final clustering, we develop a new approach to combine the clustering results from each data source. Note that in each iteration, one data source is randomly picked and every data point (i.e., gene) is reassigned to one of the k clusters based on information from that data source and on its previous assignment. After all the iterations are completed, each data point has to be given a final assignment to one of the k different clusters based on some criteria that depends on its cluster assignment for each of the m data sources. Note that the cluster assignments for each of the data sources may be different.

One approach is to assign each data point to the maximum probability cluster, as suggested in (Kasturi and Acharya, 2004). This approach has the underlying assumption that cluster assignments for the m data sources are correlated. However, this need not be true. Another approach is to compute a consensus mean of the cluster assignments obtained from the m data sources. This approach may not always be successful especially when the number, variability and reliability of the data sources are large (Bickel and Tobias, 2004).

We introduce a new method of assigning cluster membership to data points by taking into account the cluster assignment obtained from each data source. The cluster assignment for each point, for each data source, can be thought as a k-dimensional vector in which only one entry (corresponding to the assigned cluster) is equal to 1 and all the others are zero. By combining the results obtained from the m data sources, the cluster assignment for each data point now constitutes a km-dimensional vector. Thus we obtain a $n \times km$ matrix whose entries are as follows:

$$C_{i((j-1)k+s-1)} = p_j \cdot \delta(C_s^{(j)}, i)$$

where p_j is the prior probability of data type j, $C_s^{(j)}$ is a cluster s in data type j and

$$\delta(C_s^{(j)}, i) = \begin{cases} 1 & gene\ i \in C_s^{(j)} \\ 0 & otherwise \end{cases}$$

The above matrix is used to cluster using one of standard clustering algorithms, such as K-means. Clearly, genes with similar cluster assignments across all data sources will be assigned to the same cluster.

3 DATA SOURCES AND REPRESENTATION

Currently, many techniques have been developed to analyze high-throughput numeric data (Ben-Dor *et al.*, 1999; Eisen *et al.*, 1998; Getz *et*

al., 2000; Tamayo *et al.*, 1999). However, it is difficult to incorporate the wealth of information contained in existing domain knowledgebases, most of which is present in free-text form. In order to efficiently extract and use this information in a homogeneous way with other numeric data, textual domain knowledge needs to be transformed into numeric data.

Our goal is to cluster a set of genes in a biologically meaningful manner. Gene expression data is clearly numeric data and can be analyzed using well-established techniques. However, there is an enormous volume of biomedical literature containing useful knowledge that needs to be extracted and utilized to enhance the quality of the analysis (Glenisson *et al.*, 2003; Masys, 2001). Starting from a biomedical literature repository, a document index is computed based on the vector space model which results in a document-term matrix, i.e., a matrix that provides information about which term appears in which document and with what frequency (Baeza-Yates and Ribeiro-Neto, 1999; Raghavan and Wong, 1986). Information on all documents linked to each gene is used to compute a gene-term matrix, as described later in Section 3.2. Note that this gene-term matrix now contains numeric data. This is the representation we use for text-based information. Later we will describe how to pick the terms of interest for this representation.

The goal is to have numeric vectors from each data source for each gene for further analysis. In our analysis, all genes of interest have two representations: term vector space from literature repository and expression vector space from the gene expression data. After obtaining the representations, the issue left is how to combine these two representations. As discussed in Section 1, one natural approach to combine the two representations is feature level integration (i.e., to simply concatenate the feature vectors from the expression and text spaces), which has the effect of combining the corresponding distance matrices (Glenisson *et al.*, 2004). Our clustering approach

provides a semantic scheme to learn from the two representations.

In what follows we will specify the data sources used in this work and discuss the gene-term construction from literature repository.

3.1 Expression and Literature Data Sources

For text information, a literature index for yeast genes was constructed from 31924 yeast-related MEDLINE abstracts, which were downloaded using Entrez/Pubmed search engine based on text matching in an entry's fields (Roberts, 2001). The abstract-gene relation information was constructed from the curated literature references available from the Saccharomyces Genome Database (SGD) [ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/] (Dolinski *et al.*, 2004). Gene expression data set was generated from cultures synchronized in cell cycle by three independent methods and consisted of measurements of 6206 genes over 77 experimental conditions (Spellman *et al.*, 1998). After removing those having no literature references, the remaining 5473 genes were retained for further analysis.

3.2 Gene-Term Matrix Construction

Each document from the data source was represented by a vector in which each component of the vector corresponds to a single term from the entire set of terms, i.e., the vocabulary (Gravano *et al.*, 1999; Raghavan and Wong, 1986). The value of each component was calculated using the term weight indexing as follows:

$$w_{ij} = tf_{ij} \times idf_i = tf_{ij} \log\left(\frac{N}{df_j}\right)$$

where term frequency *tf*$_{ij}$ measures the occurrences of a term *j* in a document *i*, and *idf* is inverse document frequency, which is equal to the logarithm of the ratio of the total number of documents (*N*) divided by the number of documents containing term j in the collection (*df*$_j$).

All MEDLINE abstracts referred to in SGD's literature database were considered as acceptable,

noise-free, domain-specific source of information for the yeast genes being considered (Stephens *et al.*, 2001). A restricted vocabulary is suggested in several recent papers (Chiang and Yu, 2003; Glenisson *et al.*, 2004; Stephens *et al.*, 2001). Often these restricted vocabularies involve terms from the GO database. In this work, we chose to eliminate such constraints and have resorted to generic text mining methods to extract the terms. Our reasoning was as follows. Since GO terms were used in the validation of the clusters, it would be inappropriate to bias the text mining part of the process by allowing only terms that would validate positively.

To integrate with the gene expression data, text data was represented as a gene-term matrix, which is obtained by combining the document-term matrix with the gene-document matrix. The textual profile of a gene *i*, a vector of terms *j*, was obtained by taking the average over the N_i documents containing gene *i*:

$$g_i = \{g_i\}_j = \left\{ \frac{1}{N_i} \sum_{k=1}^{N_i} \frac{w_{ki}}{N_{kg}} \right\}_j$$

Here N_{kg} denotes the number of genes linking to document *k*. We added this factor to consider the distribution of number of genes associated with each document and this factor was not considered by Glenisson et al. (Glenisson *et al.*, 2004). After considering this factor, the weight indexing was first averaged over the number of genes in a specific document before being averaged over all documents linking gene *i*.

4 EXPERIMENTAL RESULTS

Four algorithms were implemented in Java. These included model-based K-means algorithm on the individual data sources (text and gene expression), model-based K-means algorithm on feature-level integration of the multi-source data, and the MSC algorithm.
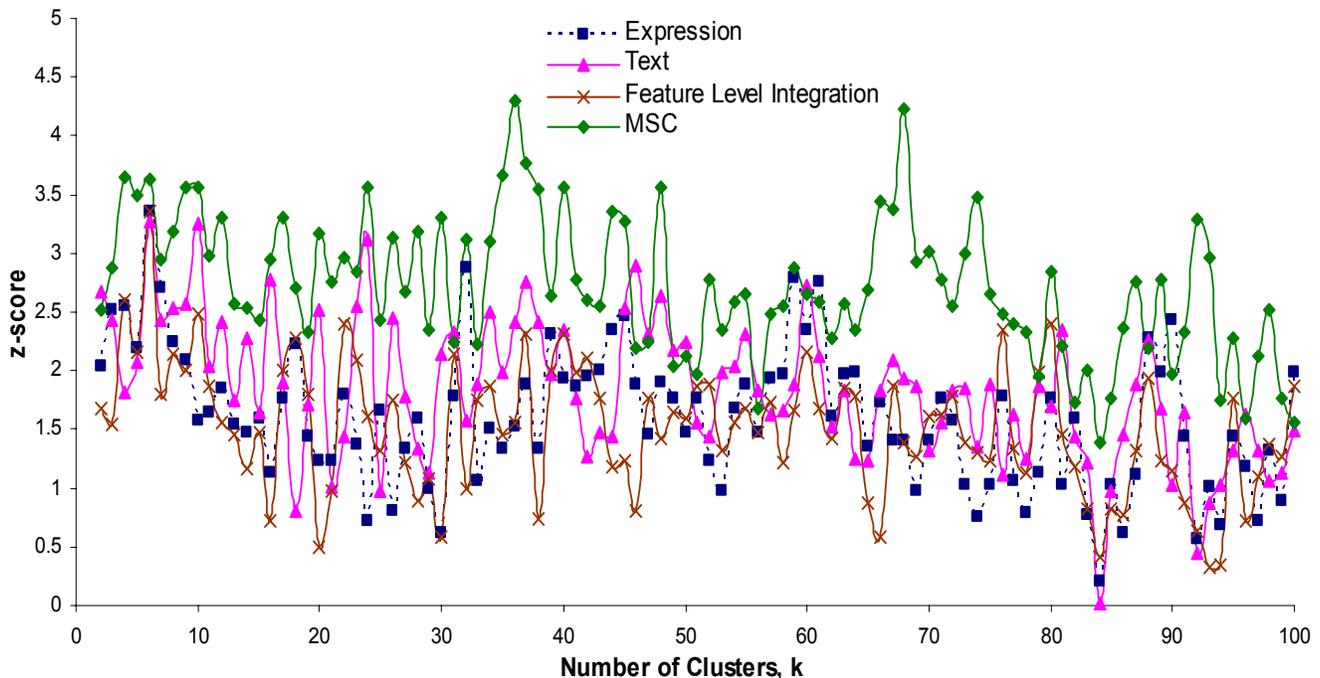
Figure 1. Clustering results from expression, text, expression-text feature level integration, and multi-source clustering. The horizontal axis shows the number of clusters desired, and the vertical axis shows z-scores.

4.1 Evaluating Clusterings

A figure of merit called **z-score** was devised by Gibbons et al. and was used to measure the quality of a clustering (Gibbons and Roth, 2002). The z-score is defined as follows:

$$z\text{-}score = \frac{MI_{real} - MI_{random}}{S_{random}}$$

where MI_{real} is the mutual information between the clustered data and the SGD gene annotation data, MI_{random} is measured for a clustering obtained by randomly assigning genes to clusters, and S_{random} is the standard deviation. Thus, higher z-scores suggest that the clustering results are more significantly related to gene function.

We compared the performance of the four clustering methods: K-means clustering of expression data, K-means clustering of text data, K-means clustering of the feature-level integrated expression and text data, and the MSC algorithm applied to expression and text data. Equal weights were used for the expression and text data in both the two multi-source algorithms, although the weights could be specified using expert knowledge to specify the importance of each data source. The expression data consisted of 5473 genes under 77 experimental conditions and the text data consisted of 5473 genes and 250 terms. The z-scores were plotted against the number of clusters, *k*, for all values of *k* from 2 to 100. The results are shown in Figure 1. Using z-score as a criterion, the results from the multi-source data clustering exhibited the best performance for over 80% of values of *k* and were better than the method that only used text data. Surprisingly, the results from the feature-level integration were worse than the methods that used only a single data source. It suggests that a simple combination of features and distance functions may not be the best approach to improve the quality of clustering. Figure 1 shows that z-scores decay as *k* grows over 70, indicating that clustering with k greater than 70 is not appropriate since a decrease in z-scores implies that clustering results are less significantly related to gene function. In the next two subsections, we explore other ways to evaluate the quality of the resulting clusters.

4.2 Function Enrichment

To assess the classification capability of the clustering algorithms, gene ontology information was used to evaluate whether the clusters have significant enrichment of one or more terms from the gene ontology (GO) database; this was done using *FuncAssociate* (Berriz *et al.*, 2003), a program that takes a list of genes as input and produces a ranked (by P-values) list of the GO attributes for which the input gene list is enriched (Ashburner *et al.*, 2000). Each query gene set is composed of the genes from each cluster in a clustering, and the output gives the terms significantly enriched in each cluster among all genes (in this case, the number of all genes is 5473 which is the total number of genes for clustering). Table 2 shows details of 8 typical clusters with enriched functional groups.

For example, cluster 1 in Table 2 contains 254 genes, 9 of which are annotated with the GO term "nucleosome". Since only 12 genes belong to this category in the whole genome, this is significant, as suggested by its P-value of 10^{-13}. These P-values reflect the statistical significance of the function enrichment by taking into account the ratio of the number of genes within a cluster in comparison to that in the whole genome. Consider the "RNA binding" ontology category in cluster 7, which contains only 18 out of the 382 genes from this category. However, the P-value is 10^{-11}, suggesting a significant enrichment for this category. This is because the 18 RNA binding genes constitute nearly 25% of the genes in cluster 7. As can be seen in the examples in Table 2, there are several functions significantly enriched in a cluster. (Details of all clusters are provided in a supplemental website [http://biorg.cs.fiu.edu/MSC].)

Function enrichment analysis also reveals that within a given cluster, often the enriched functions are closely related. For instance, in cluster 3, which has 277 genes, 101 ontology categories are

Table 3. Function enrichment of clusters generated from Multi-Data clustering.

Cluster	# of Genes in Cluster	Enriched Functional Category(Total genes)	Clustered Genes	$-\log_{10}$ (p-value)
1	254	Nucleosome(12)	9	13
		external encapsulating structure(109)	11	5
		glycoprotein biosynthesis(66)	23	22
		protein amino acid glycosylation(66)	23	22
2	70	Sterol biosynthesis(29)	4	5
		sterol metabolism(34)	4	5
3	277	DNA helicase activity(27)	9	9
		base-excision repair(10)	5	6
		postreplication repair(10)	5	6
		spindle pole body(49)	9	6
		DNA replication and chromosome cycle(238)	45	28
4	39	cytosolic ribosome(167)	36	65
		eukaryotic 48S initiation complex(63)	19	34
5	624	regulation of physiological process(376)	57	13
		regulation of metabolism(376)	57	13
		regulation of biological process/regulation(426)	61	13
		regulation of transcription, DNA-dependent(297)	49	13
		protein modification(398)	59	13
6	179	snoRNA binding(30)	11	13
		rRNA processing(166)	52	56
		rRNA metabolism(235)	52	47
		processing of 20S pre-rRNA(51)	20	24
		small nucleolar ribonucleoprotein complex(32)	14	18
7	74	RNA binding(382)	18	11
		RNA-dependent ATPase activity(25)	7	10
		ATP-dependent RNA helicase activity(25)	7	10
		ribosome assembly(62)	8	9
8	53	cell proliferation(588)	21	9
		acid phosphatase activity(5)	4	8
		cell cycle/cell-division cycle(516)	15	5

enriched. These categories include, but not limited to, DNA replication and chromosome cycle, cell proliferation, helicase activity, mitotic recombination, and so on. All the 101 enriched ontology categories in this cluster are involved in cell proliferation and DNA replication, which is biologically meaningful because when cell occurs, DNA replication and regulation must also occur. Similar

observations can be made about other clusters. For example, the genes in cluster 5 are predominantly involved in regulation, genes in cluster 6 perform RNA processing and RNA metabolism, and so on.

4.3 Transcription Factor Binding Motifs

Next we decided to perform further exploratory analysis of some of the clusters obtained by the MSC algorithm. For the purpose of comparison, we examined the clusters obtained using the MSC algorithm and those obtained using only gene expression data. Upon further examination, we found that cluster 8 (with a total of 53 genes) from the MSC clustering shared 41 of the 43 genes in cluster 14 from the expression clustering. The genes contained in these two clusters are shown in Table 4, with the common ones not shown in bold font. Gene function enrichment tests showed that the significant categories were metabolism, cell growth, cell division, and DNA synthesis (as cluster 8 shown in table 3). Most genes in cluster 8 that were not in cluster 14 also belonged to these same categories, implying that text data can enrich genes with similar functions. Our new approach takes into account both expression and function, giving it increased ability to capture more biologically meaningful features. For instance, YBR093C is an ORF whose product is involved in acid phosphatase activity and YAR071W, YBR092C, and YHR215W in cluster 14 are involved in the same function. YBR093C is present in cluster 8, but not in cluster 14. As shown in table 3, there are only total 5 genes belong to this category in whole genome. With integrating literature data by MSC algorithm, one more gene is enriched.

Further exploratory analysis was performed from the point of view of shared motifs. One underlying assumption in clustering is that genes in a cluster are functionally related, implying that there is a strong possibility that many of them are also co-regulated, and co-regulated genes share transcription factor binding motifs (i.e., regulatory elements) in their upstream sequences. Motif de-

tection is often performed on clusters obtained by clustering gene expression data. Thus clustering schemes can be evaluated by looking for the presence of motifs in gene clusters.

Towards this end, we applied the motif discovery tool, AlignACE (Roth *et al.*, 1998) to find shared motifs in the two clusters. Results revealed a motif, GGCACTCACACGTGGG, located in the upstream sequence of YBR093C, which, according to TRANSFAC (Matys *et al.*, 2003), is known to be the binding site for the transcription factor PHO4 and has been reported previously in the literature (Barbaric *et al.*, 1992; Vogel *et al.*, 1989). Genes that shared this motif are YBR093C, YAR0183, YDR055W, YHR215W, YML034W, and YOR313C. In particular, YBR093C and YHR215W are two of three repressible acid phosphatases (SGD). Thus, clustering obtained by integrating information from the literature databases, as performed by the MSC algorithm was able to better detect motifs.

Table 4. ORFs contained in clusters generated from expression data and multi-source clustering.

Cluster14 from expression data		Cluster 8 from MSC	
YAL022C	YAR018C	YAR018C	YAR071W
YAR071W	YBL043W	YBL043W	YBR038W
YBR038W	YBR054W	YBR054W	YBR092C
YBR092C	YBR202W	**YBR093C**	YBR202W
YDR033W	YDR146C	**YDL117W**	YDR033W
YEL065W	YGL008C	**YDR055W**	YDR146C
YGL021W	**YGL116W**	YEL065W	YGL008C
YGR092W	YGR108W	YGL021W	YGR092W
YGR143W	YHL028W	YGR108W	YGR143W
YHR023W	YHR215W	YHL028W	YHR023W
YIL158W	YJL157C	**YHR152W**	YHR215W
YJR092W	YKR093W	YIL158W	YJL157C
YLR131C	YLR190W	**YJL159W**	YJR092W
YML034W	YML119W	**YKL163W**	**YKL164C**
YMR001C	YMR032W	**YKL185W**	YKR093W
YMR189W	YNL058C	YLR131C	YLR190W
YNL160W	YOL158C	**YLR274W**	YML034W
YOR025W	YOR313C	YML119W	YMR001C
YOR315W	YPL061W	YMR032W	**YMR145C**
YPL242C	YPR019W	YMR189W	YNL058C
YPR119W	YPR149W	**YNL078W**	YNL160W
YPR156C		**YOL070C**	YOL158C
		YOR025W	YOR313C
		YOR315W	YPL061W
		YPL242C	YPR019W
		YPR119W	YPR149W
		YPR156C	

5 DISCUSSION

Clusters obtained from microarray data analysis must correlate to the existing knowledge. Mining on gene expression alone may not be able to reveal the biological information related to the gene expressions. Most biologists focus their research on a small select set of genes, which they know to be functionally related. Consequently, their publications focus on these genes. Therefore publications stored in medical literature databases, such as PubMed, can provide valuable additional information. In this paper, we use text literature as a guide for microarray data analysis. In particular, we want to identify subgroups of genes with commonalities in gene expression and in biological function.

We have developed a new clustering algorithm (MSC) for multi-source data. We applied the MSC algorithm to gene expression and literature data related to the *Saccharomyces* genome. Using the z-score measure, we showed that the MSC algorithm performed significantly better than the feature-level integration approach. Also, the clusters from the MSC algorithm shared regulatory elements which were not found using gene expression data alone. The software is available from the authors upon request.

There are several natural avenues for future research. First, one obvious research direction is to include more sources of biological data for our experiments with the MSC algorithm, such as phylogenetic profiles and DNA sequence information. Second, since genes could be different actors at different conditions, genes may belong to multiple clusters. The MSC algorithm needs to be modified to accommodate this possibility. Third, it would also be interesting to extend the MSC algorithm by incorporating statistical inference techniques to adaptively weight different data sources during the clustering process.

Supplemental Website:
http://biorg.cs.fiu.edu/MSC

ACKNOWLEDGEMENT

REFERENCES

Altman, R.B. and Raychaudhuri, S. (2001). "Whole-genome expression analysis: challenges beyond clustering." *Current Opinion in Structural Biology* **11**(3): 340-347.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. and Consortium, G.O. (2000). "Gene Ontology: tool for the unification of biology." *Nature Genetics* **25**(1): 25-29.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, Addison Wesley Longman Publishing Co. Inc.

Ball, C.A., Brazma, A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., Matese, J.C., Parkinson, H., Quackenbush, J., Ringwald, M., Sansone, S.A., Sherlock, G., Spellman, P., Stoeckert, C., Tateno, Y., Taylor, R., White, J. and Winegarden, N. (2004). "Submission of microarray data to public repositories." *PLoS Biol* **2**(9): E317.

Barbaric, S., Fascher, K.D. and Horz, W. (1992). "Activation of the weakly regulated PHO8 promoter in S. cerevisiae: chromatin transition and binding sites for the positive regulatory protein PHO4." *Nucleic Acids Res* **20**(5): 1031-8.

Becker, S. (1996). "Mutual information maximization: Models of cortical self-organization." *Network: Computation in Neural Systems* **7**(1): 7-31.

Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999). "Clustering gene expression patterns." *J Comput Biol* **6**(3-4): 281-97.

Berriz, G.F., King, O.D., Bryant, B., Sander, C. and Roth, F.P. (2003). "Characterizing gene sets with FuncAssociate." *Bioinformatics* **19**(18): 2502-2504.

Bickel, S. and Tobias, S. (2004). *Multi-View Clustering*. Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04).

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J.C. and DeRisi, J.L. (2003). "The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum." *Plos Biology* **1**(1): 85-100.

Chaussabel, D. and Sher, A. (2002). "Mining microarray expression data by literature profiling." *Genome Biol* **3**(10): RESEARCH0055.

Chiang, J.H. and Yu, H.C. (2003). "MeKE: discovering the functions of gene products from biomedical literature via sentence alignment." *Bioinformatics* **19**(11): 1417-1422.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). "Maximum likelihood from incomplete data via the em algorithm." *Journal of the Royal Statistical Society* **39**: 1-38.

Dolinski, K., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R., Oughtred, R., Theesfeld, C.L., Binkley, G., Lane, C., Schroeder, M., Sethuraman, A., Dong, S., Weng, S., Miyasato, S., Andrada, R., Botstein, D. and Cherry, J.M. (2004). Saccharomyces Genome Database

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). "Cluster analysis and display of genome-wide expression patterns." *Proceedings of the National Academy of Sciences of the United States of America* **95**(25): 14863-14868.

Fleischmann, W., Moller, S., Gateau, A. and Apweiler, R. (1999). "A novel method for automatic functional annotation of proteins." *Bioinformatics* **15**(3): 228-233.

Getz, G., Levine, E. and Domany, E. (2000). "Coupled two-way clustering analysis of gene microarray data." *Proc Natl Acad Sci U S A* **97**(22): 12079-84.

Gibbons, F.D. and Roth, F.P. (2002). "Judging the quality of gene expression-based clustering methods using gene annotation." *Genome Research* **12**(10): 1574-1581.

Glenisson, P., Antal, P., Mathys, J., Moreau, Y. and De Moor, B. (2003). "Evaluation of the vector space representation in text-based gene clustering." *Pac Symp Biocomput*: 391-402.

Glenisson, P., Mathys, J. and De Moor, B. (2004). "Meta-Clustering of Gene Expression Data and Literature-based Information." *SIGKDD Explorations* **5**(2): 101-112.

Gravano, L., Garcia-Molina, H. and Tomasic, A. (1999). "Gloss: Text-source discovery over the internet." *ACM Transactions on Database Systems* **24**(2): 229-264.

Herwig, R., Poustka, A.J., Muller, C., Bull, C., Lehrach, H. and O'Brien, J. (1999). "Large-scale clustering of cDNA-fingerprinting data." *Genome Research* **9**(11): 1093-1105.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002). "Revealing modular organization in the yeast transcriptional network." *Nature Genetics* **31**(4): 370-377.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for clustering data*, Prentice Hall.

Jenssen, T.K., Laegreid, A., Komorowski, J. and Hovig, E. (2001). "A literature network of human genes for high-throughput analysis of gene expression." *Nat Genet* **28**(1): 21-8.

Kasturi, J. and Acharya, R. (2004). Clustering of diverse genomic data using information fusion *Proceedings of the 2004 ACM symposium on Applied computing* Nicosia, Cyprus ACM Press**:** 116-120

Masys, D.R. (2001). "Linking microarray data to the literature." *Nat Genet* **28**(1): 9-10.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003). "TRANSFAC(R): transcriptional regulation, from patterns to profiles." *Nucl. Acids Res.* **31**(1): 374-378.

Neal, R.M. and Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse,

and other variants. *Learning in Graphical Models*. M. Jordan**:** 355-368.

Raghavan, V.V. and Wong, S.K.M. (1986). "A critical analysis of vector space model for information retrieval." *Journal of the American Society for Information Science* **37**(5): 279-87.

Raychaudhuri, S., Chang, J.T., Imam, F. and Altman, R.B. (2003). "The computational analysis of scientific literature to define and recognize gene expression clusters." *Nucleic Acids Research* **31**(15): 4553-4560.

Raychaudhuri, S., Chang, J.T., Sutphin, P.D. and Altman, R.B. (2002). "Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature." *Genome Research* **12**(1): 203-214.

Roberts, R.J. (2001). "PubMed Central: The Gen-Bank of the published literature." *Proc Natl Acad Sci U S A* **98**(2): 381-2.

Roth, F.P., Hughes, J.D., Estep, P.W. and Church, G.M. (1998). "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." *Nature Biotechnology* **16**(10): 939-945.

Segal, E., Yelensky, R. and Koller, D. (2003). "Genome-wide discovery of transcriptional modules from DNA sequence and gene expression." *Bioinformatics* **19 Suppl 1**: 273-82.

Shatkay, H., Edwards, S., Wilbur, W. J. and Boguski, M. (2000). "Genes, themes and microarrays: using information retrieval for large-scale gene analysis." *Proc Int Conf Intell Syst Mol Biol.*

Sherlock, G. (2000). "Analysis of large-scale gene expression data." *Current Opinion in Immunology* **12**(2): 201-205.

Spellman, P.T., Sherlock, G., Futcher, B., Brown, P.O. and Botstein, D. (1998). "Identification of cell cycle regulated genes in yeast by DNA microarray hybridization." *Molecular Biology of the Cell* **9**: 371a-371a.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., 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." *Molecular Biology of the Cell* **9**(12): 3273-3297.

Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R. and Mostafa, J. (2001). *Detecting gene relations from MEDLINE abstracts*. Proc of the sixth Ann Pac Symp Biocomp (PSB 2001).

Tamames, J., Ouzounis, C., Casari, G., Sander, C. and Valencia, A. (1998). "EUCLID: automatic classification of proteins in functional classes by their database annotations." *Bioinformatics* **14**(6): 542-543.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999). "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation." *Proceedings of the National Academy of Sciences of the United States of America* **96**(6): 2907-2912.

Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L. and Weinstein, J.N. (1999). "MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling." *Biotechniques* **27**(6): 1210-4, 1216-7.

Vogel, K., Horz, W. and Hinnen, A. (1989). "The two positively acting regulatory proteins PHO2 and PHO4 physically interact with PHO5 upstream activation regions." *Mol Cell Biol* **9**(5): 2050-7.

Wu, L., Oviatt, S.L. and Cohen, P.R. (1999). "Multimodal integration - a statistical view." *IEEE Transactions on Multimedia* **1**(4): 334-341.

Yandell, M.D. and Majoros, W.H. (2002). "Genomics and natural language processing." *Nature Reviews Genetics* **3**(8): 601-610.

Zhong, S. and Ghosh, J. (2003). "A comparative study of generative models for document clustering." *Proceedings of the workshop on Clustering High Dimensional Data and Its Applications in SIAM Data Mining Conference.*