A Knowledge-Driven Method to Evaluate Multi-Source Clustering

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

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

ABSTRACT

Traditional exploratory analysis of gene expression data involves the application of clustering algorithms to obtain clusters of related genes. Recent research has focused on improving such analyses using additional biological information. It has been demonstrated that biological literature can complement the information extracted from gene expression data to obtain better gene clusters. The Multi-Source Clustering (MSC) algorithm, which was recently proposed by the authors, performs semantic integration of information obtained from gene expression data and biomedical text literature. To address the challenge of evaluating clustering results, a new knowledge-driven approach is proposed based on information extracted from a database of published binding sites of known transcription factors (TF). Thus a new data source is used as a basis for evaluation. We propose the use of a measure called C-index for an objective, quantitative evaluation. We compare the results of algorithm MSC for the integrated data sources with the results obtained (a) & (b) by clustering applied to the two sources of data separately, and (c) by clustering after using a feature-level integration (i.e., after concatenating the features obtained from the two data sources). We show that the C-index measurements of the clustering results from MSC are better than that from the other three approaches. We also identify TFs whose binding sites are significantly over-represented in promoter regions of clustered genes.

[*] To whom correspondence should be addressed.

1 INTRODUCTION

Clustering genes based on gene expression data is now a routine method to partition genes into groups (or clusters) sharing similar expression patterns (Eisen, Spellman et al. 1998; Sherlock 2000; Sharan, Elkon et al. 2002). Two critical questions have been pursued by researchers: (a) How to improve the clustering by combining information from different biological data sources? (b) How to validate or evaluate the resulting clusters?

The large (and growing) biological literature database has been considered as an important source of additional information for any exploratory analysis of biological data. It was shown to be useful for identifying functional commonalities of genes and to help drive the interpretation and organization of the expression data (Altman and Raychaudhuri 2001). Several algorithms have been proposed to combine gene expression data and text literature data sources to perform clustering (Shatkay, Edwards et al. 2000; Stephens, Palakal et al. 2001; Chiang and Yu 2003; Raychaudhuri, Chang et al. 2003; Glenisson, Mathys et al. 2004; Yang, Zeng et al. 2005). Many other sources of data have also been successfully used to perform exploratory analysis. These sources include annotations form biological databases, protein interactions, transcription factor binding, etc. (Ihmels, Friedlander et al. 2002; Adryan and Schuh 2004; Tanay, Sharan et al. 2004).

In general, there are two existing clustering approaches for combining multiple sources of data: semantic integration and feature-level integration. Methods that use feature-level integration combine the features and then perform the analysis in the joint feature space (Glenisson, Mathys 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 a recent paper, 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* (Segal, Yelensky et al. 2003). The MSC algorithm, which was recently devised by the authors, is an example of the semantic integration method (Yang, Zeng et al. 2005). It implicitly learns the

correlation structure among heterogeneous data sources and provides a semantic scheme to analyze data from them. Using a measure called **z-score** (Gibbons and Roth 2002), it was shown that the MSC clustering outperformed those using single data source only or multiple sources combined at the feature level (Yang, Zeng et al. 2005).

To address the question of validating or evaluating clustering outcomes, researchers have used annotations from the Gene Ontology (GO) database. The Gene Ontology (GO) represents an important knowledge resource to describe the function of genes, and the GO database contains annotations for a large number of genes from a variety of organisms (Ashburner, Ball et al. 2000). The z-score evaluation measure is based on mutual information between cluster membership and GO annotations, and was used to judge the quality of clustering methods (Gibbons and Roth 2002; Yang, Zeng et al. 2005). A different approach based on similarity information extracted from GO annotations has also been proposed (Bolshakova, Azuaje et al. 2005). An important point to note is that if GO annotations are to be used for evaluation purposes, then it should not be used as a data source in the clustering algorithm, since this would bias the evaluation. By the same token, it would be inappropriate to use GO terms and attributes to perform text mining of biological literature databases.

This raises some general questions: what other sources of data can be used for the purpose of evaluating clustering outcomes? And what evaluation measures are appropriate for these data sources? In this paper, we explore new data sources and measures for evaluating clusters. The idea is to use databases containing information about transcription factors (TF), which are involved in gene regulation, and their binding sites (i.e., the regulatory elements, or TFBS).

The remainder of the paper is organized as follows. In Section 2, we briefly review the previously described MSC algorithm, describe our new gene cluster assessment using information about TF binding sites in the promoter regions of the genes, and then introduce the evaluation measures used in our experiments. In Section 3, we show the per-

formance of different clustering approaches through a typical example and present TF enrichment results. We conclude with a discussion in Section 4.

2 METHODS

2.1 The MSC algorithm

Intuitively, clustering is the problem of partitioning a set of points in a multi-dimensional space into clusters such that the points belonging to the same cluster are *similar* while the points belonging to different clusters are *dissimilar* (Jain and Dubes 1988). For our purposes, the goal is to identify clusters of related genes using the available datasets. The MSC algorithm, a variant of the EM method (Dempster, Laird et al. 1977), stochastically builds the models for each data source by boosting the models using the cluster assignments from the other models. In each iteration, we first randomly select a data source based on the weight vector. We then perform the following steps: (i) find the model parameters that maximize the likelihood of the data given the current cluster assignment; (ii) assign the data points to the cluster that maximizes the posterior probability. Our previously reported experimental results show that the MSC algorithm implicitly learns the correlation structure among the multiple data sources (Yang, Zeng et al. 2005).

In order to obtain the final clustering, the cluster assignment for each point, for each data source, can be thought of 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 and the whole data set corresponds to an $n \times km$ matrix, which is used to cluster using one of standard clustering algorithms, such as K-means. Detailed descriptions of the MSC algorithm can be found in (Yang, Zeng et al. 2005).

2.2 Cluster validity assessment

Here we propose a method to use a new source of knowledge — gene regulatory

information — to evaluate the validity of clusters. A key source of information used is TRANSFAC, which is a database containing information on eukaryotic transcription factors and profiles of their genomic binding sites (Wingender, Chen et al. 2000).

Similar to the approach used for other data sources, we propose a binary matrix **M**, such that its *ij*-th entry, m_{ij} = 1 if there is at least one TFBS for TF_j in the promoter region of gene G_i, and 0 otherwise. The matrix provides a basis for defining a distance function between the genes. Using the cosine distance measure (other distance measures could also have been used), we define the distance between gene *i* and gene *j* as:

$$d_{ij} = 1 - \frac{\sum_k m_{ik} m_{jk}}{\sqrt{\sum_k m_{ik} \cdot \sum_k m_{jk}}}$$

The above distance measure can be used to compute the **C-index**, a cluster validity estimator (Hubert and Schultz 1976), which was recently used in a different context (Bolshakova, Azuaje et al. 2005). It is defined as follows:

$$C = \frac{S - S_{min}}{S_{max} - S_{min}}$$

Here *S* is the sum of distances over all pairs of genes from the same cluster (over all clusters). Let *l* be the number of those pairs, then S_{min} is the sum of the *l* smallest distances between all pairs of genes and S_{max} is the sum of the *l* largest distances. It is easy to see that the numerator in the above formula will be small for pairs of genes with a small distance. Hence, a small value of C-index indicates a good clustering.

2.3 Data sources and representation

Briefly, the goal is to build numeric vectors from each data source for each gene for further analysis. In our analysis, all genes of interest have two representations: *Term Vector* based on information from the literature repository, and *Expression Vector* from the gene expression data (obtained from microarray data).

To represent information from text data, the *Document-Term matrix* was constructed from a biomedical literature repository (MEDLINE abstracts) using tf-idf indexing (Baeza-

Yates and Ribeiro-Neto 1999). Then a *Gene-Term matrix* was obtained by combining the D*ocument-Term matrix* with the *Gene-Document matrix* from the SGD database [ftp://genomeftp.stanford.edu/pub/yeast/data_ download/literature_curation/].

Data sources and representations are similar to that in our previous study (Yang, Zeng et al. 2005). One difference from the previous study was in the way the tf-idf indexing was constructed, for which several papers have recommended a restricted vocabulary (Stephens, Palakal et al. 2001; Chiang and Yu 2003; Glenisson, Mathys et al. 2004). In this work, only GO terms were used for the indexing. Specifically, an index for yeast genes was constructed from 31,924 yeast-related MEDLINE abstracts. Gene expression data set was generated from cultures synchronized in cell cycle by three independent methods and consisted of measurements of 720 genes over 77 experimental conditions (Spellman, Sherlock et al. 1998).

3 EXPERIMENTAL RESULTS

3.1 Evaluating Clustering Outcomes

We used the C-index to compare 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 720 genes under 77 experimental conditions and the text data consisted of 720 genes and 213 GO related terms. The C-indices were plotted against the number of clusters, k, for all values of k from 2 to 50. The results are shown in Figure 1. Using C-index as a criterion, the results from the multi-source data clustering exhibited the best performance for about 70% of values of k, implying that for a range of cluster sizes, the MSC algorithm has superior performance. The results from the feature-level integration were comparable to the methods that used only a single data source, suggesting that a simple combination of features and distance functions may not be the best approach to improve the quality of

clustering, and that the semantic level integration does add value to the clustering outcomes.

C-index could also be used to choose the optimal number of clusters. Figure 1 indicates that C-index with *k* equal to 17 is smallest for MSC clusters. In the next subsection, we will explore transcription factor enrichment analysis on those 17 clusters.

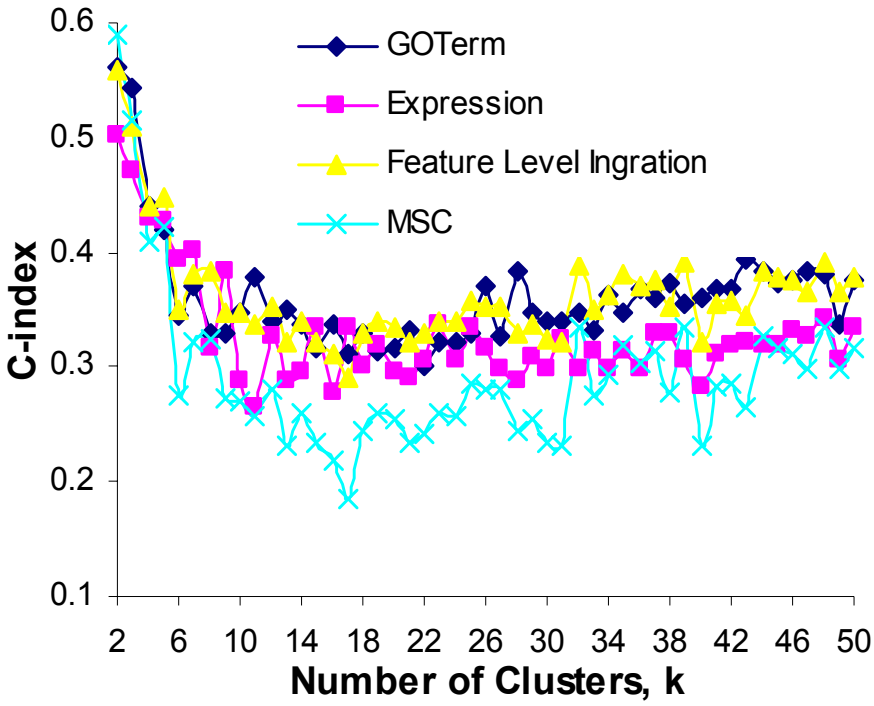

Figure 1. Clustering results from expression, text (using Go Terms), expression-text feature level integration, and multi-source clustering. The horizontal axis shows the number of clusters desired. and the vertical axis shows C-index.

3.2 Transcription Factor Enrichment

To assess the classification capability of the clustering algorithms, known information on binding sites (TFBSs) were used to evaluate whether the clusters have significant enrichment of being regulated by one or more TFs. A software package written in Java takes a list of genes as input and produces a ranked (by P-values) list of the TFs whose TFBSs are significantly over-represented in promoter regions of the genes in the list. Such

significant TFs could be candidates regulating the corresponding set of genes. Each query gene set is composed of the genes from each cluster in a clustering (in this case, 17 clusters from MSC clustering were used). Table 1 shows details of 5 typical clusters with enriched transcription factors.

For example, cluster 1 in Table 2 contains 61 genes, 31 of which share TFBSs regulated by TF SWI4. Since only 115 genes are known to be regulated by this TF, this is considered statistically significant (P-value = 10^{-27}). These P-values take into account the ratio of the number of genes within a cluster in comparison to that in the whole genome. As can be seen in the examples in Table 1, there are several transcription factors significantly enriched in a cluster. (Details of all clusters from our experiments are provided in a supplemental website [http://biorg.cs.fiu.edu/TFF].)

Table 1. TF enrichment of clusters generated from Multi-Source Clustering.

Cluster	# of Genes in Cluster	Enriched TF (Total genes)	Clustered Genes	$-\log_{10}$ (p-value)
1	61	SWI4(115)	31	27
		STB1(88)	23	17
		SWI6(125)	24	13
2	36	HAP1(49)	11	9
		HAP2/3/4(38)	4	7
3	76	MBP1(69)	25	24
		SWI6(125)	35	18
		MAT1(54)	13	7
4	18	MET31(11)	6	11
		CBF1(41)	9	41
		PHO4(46)	7	13
5	58	SWI6(125)	31	20
		MBP1(69)	23	19
		SWI4(115)	19	8

4 DISCUSSION

The repositories of biomedical literature are increasing at a dramatic rate and should play an increasingly important role in exploratory analyses of genes. Vector space models were used to convert textual domain knowledge into numeric data (term vector space). Tailored term vocabulary (GO terms), which reflects the knowledge of this domain, was used to reduce the noise in the information.

A new approach based on knowledge extracted from TRANSFAC (a database of gene regulatory information) is used to assess the quality of clustering. Effectively, a new data source is used for evaluation. The C-index was used to compare results from four different clustering approaches and showed that MSC algorithm (with semantic integration of gene expression and biomedical text data) outperformed three other approaches. Also, the clusters from the MSC algorithm (with 17 clusters) were used to explore significant TFBSs, which could be potentially responsible for regulating the genes in that cluster. The software is available from the authors upon request.

ACKNOWLEDGEMENTS

ELZ was supported by a Florida International University Presidential Graduate Fellowship. The research of GN was supported in part by NIH Grant P01 DA15027-01.

REFERENCES

Adryan, B. and Schuh R. (2004). "Gene-Ontology-based clustering of gene expression data." Bioinformatics **20**(16): 2851-2.

Altman, R. B. and Raychaudhuri S. (2001). "Whole-genome expression analysis: challenges beyond clustering." Curr Opin Struct Biol **11**(3): 340-7.

Ashburner, M., Ball, C. A. et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

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

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

Bolshakova, N., Azuaje, F. et al. (2005). "A knowledge-driven approach to cluster validity assessment." In Press, Bioinformatics.

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. et al. (1977). "Maximum likelihood from incomplete data via the em algorithm." Journal of the Royal Statistical Society **39**: 1-38.

Eisen, M. B., Spellman, P. T. et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-8.

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

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

Hubert, L. and Schultz, J. (1976). "Quadratic assignment as a general data-analysis strategy." British Journal of Mathematical and Statistical Psychologie **29**: 190-241.

Ihmels, J., Friedlander, G. et al. (2002). "Revealing modular organization in the yeast transcriptional network." Nat Genet **31**(4): 370-7.

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

Raychaudhuri, S., Chang, J. T. et al. (2003). "The computational analysis of scientific literature to define and recognize gene expression clusters." Nucleic Acids Res **31**(15): 4553-60.

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

Sharan, R., Elkon, R. et al. (2002). "Cluster analysis and its applications to gene expression data." Ernst Schering Res Found Workshop(38): 83-108.

Shatkay, H., Edwards, S. et al. (2000). "Genes, themes and microarrays: using information retrieval for large-scale gene analysis." Proc Int Conf Intell Syst Mol Biol **8**: 317-28.

Sherlock, G. (2000). "Analysis of large-scale gene expression data." Curr Opin Immunol **12**(2): 201-205.

Spellman, P. T., Sherlock, G. et al. (1998). "Identification of cell cycle regulated genes in yeast by DNA microarray hybridization." Mol Biol Cell **9**: 371a-371a.

Stephens, M., Palakal, M. et al. (2001). "Detecting gene relations from Medline abstracts." Pac Symp Biocomput: 483-95.

Tanay, A., Sharan, R. et al. (2004). "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data." Proc Natl Acad Sci U S A **101**(9): 2981-6.

Wingender, E., Chen, X. et al. (2000). "TRANSFAC: an integrated system for gene expression regulation." Nucleic Acids Res **28**(1): 316-9.

Yang, C., Zeng, E. et al. (2005). "Clustering Genes using Gene Expression and Text Literature Data. " To Appear, Proc. Of Computational Systems Bioinformatics CSB2005.