A Hybrid Knowledge-Driver Approach to Clustering Gene Expression Data

Spiridon C. Denaxas[1], Christos Tjortjis[1]

[1] School of Informatics, University of Manchester, PO Box 88, Manchester, M60 1QD, UK
S.Denaxas@postgrad.manchester.ac.uk tjortjis@manchester.ac.uk

Abstract. Microarray technology has enabled scientists to monitor and process the expression of thousands of genes in parallel, within a single experiment. However, the efficient interpretation and validation of the analysis results, based on current medical and biological knowledge, remains a challenge. Most gene expression analysis approaches do not incorporate existing background knowledge in the process, thus necessitating laborious manual interpretation. In this paper we propose a novel hybrid knowledge-driven approach for analyzing gene expression data which integrates currently available biological and medical knowledge within the actual clustering process. Existing published scientific information is correlated to create, validate and biologically interpret the resulting clusters. Some preliminary experimental results are supplied using a sample yeast genome data set.

1 Introduction

DNA microarray technology has made it possible to simultaneously monitor the expression levels of thousands of genes in parallel during important biological functions and across large collections of samples, providing insight into gene functionality and their regulatory mechanisms. Once the expression levels of the genes have been determined, it is often an important task to identify and group together genes with similar expression patterns (coexpressed genes). The analysis of gene expression data with the ultimate goal of identifying genes that share similar expression patterns and grouping them together is known as *cluster analysis* [1].

During clustering, data algorithms and statistical techniques are deployed in order to partition the gene expression data set in a manner that genes which share a similar behavior pattern under a given set of specific conditions are members of the same cluster. Ideally, the majority of genes in the data set should be placed in distinct non-overlapping and biologically meaningful clusters. Such groups of genes are much more tractable to study by domain experts than raw expression data.

One of the primary goals of clustering is to attribute functions to unidentified genes and locate novel functions based on co-expression [2], [3], [4]. Relatedness in biological function often implies similarity in expression behavior and vise-versa. The membership of a gene with unidentified functionality in a functional coherent cluster of

coexpressed genes implies that it also shares the same functionality as the rest of the genes belonging in the cluster. Additionally, the examination of genes outside of the formed clusters may yield to the discovery of novel functionalities previously unidentified [5].

2 Related Work

Numerous clustering algorithms exist for the analysis and examination of gene expression data. Algorithms can be classified as *first-generation* and *second-generation* [6]. A brief overview of each group is provided in the next sections.

2.1 First generation clustering algorithms

First-generation clustering algorithms are existing traditional data clustering algorithms which are applied to gene expression data sets. First generation algorithms include direct visual inspection [7], *K*-means clustering [8], Self Organizing Maps (SOM) [9] and hierarchical clustering [4]. Despite the fact that they have been developed outside the biological community, their application on gene expression data may yield biologically meaningful results.

However, very often these algorithms are associated with one or more user-defined parameters which may render their use very difficult. For example, both *K*-means and SOM require the total number of clusters to be predefined which may be very difficult if not impossible to be predicted. Additionally, even the slightest alteration in these parameters will yield major changes in the resulting clusters thus making extensive tuning a mandatory element of the process.

Finally, first-generation algorithms usually suffer from processing and memory constraints when applied to very large data sets. Recently, a number of algorithms were proposed to tackle these limitations and to include more domain-specific parameters. These algorithms are known as *second-generation* clustering algorithms.

2.2 Second generation clustering algorithms

Second-generation clustering algorithms have been proposed to address the shortcomings and limitations of first-generation algorithms. Additionally, they often include domain-specific knowledge in the clustering process which tends to lead to more precise and biologically meaningful results.

Examples of second-generation algorithms include the self-organizing tree algorithm (SOTA) [10], quality-based [3] and adaptive quality-based clustering [11], model-based clustering [12], simulated annealing [13], the cluster affinity search technique (CAST) [14], CLICK [15] and DHC [16].

QT_Clust for example, was designed having cluster analysis of gene expression data in mind. As a quality-based algorithm, it produces clusters that have a quality guaran-

tee which ensures that all members of the cluster have similar expression patterns with all other members of the specific cluster. The quality guarantee is essentially a predefined by the user threshold which represents the maximal distance between any two given points within the cluster.

2.3 Result Interpretation

One of the most significant tasks in the process of clustering gene expression data is the actual interpretation of the results [1], [17], [6]. The interpretation of co-expressed genes and coherent patterns mainly depends in associating existing domain knowledge with the current data set, which in itself presents several significant challenges.

The main limitation of many gene expression analytic approaches is the fact that they do not successfully incorporate domain knowledge about the genes into the actual process, compromising the quality of the results obtained.

Once the clustering algorithm has terminated, the challenge is to validate and interpret the resulting clusters, define their boundaries and chose the optimal ones so that genes are divided forming non-overlapping biologically meaningful clusters. Typically, the cluster boundaries are manually defined and evaluated so that the selected clusters contain functionally relevant genes.

Several alternative solutions have been proposed for automatically defining the boundaries of the generated clusters based on several statistical criteria and parameters [12]. These however do not ensure that the final set of clusters selected contains biologically meaningful group of genes and omit any existing relevant biological knowledge in the process.

It has been argued that the effective integration of external information elements, such as functional information of the genes and upstream oligonucleotide sequence information, will drive the process of organizing and analyzing gene expression data in a more efficient and precise manner [18]. Published scientific text contains a distilled version of the most significant biological and medical discoveries and is a potent source of functional information for analytical algorithms. It is critical to include relevant and comprehensive background literature to appropriately analyze such data sets and eventually understand them [19].

A number of solutions embracing this notion have been developed and take a literature-based approach to clustering gene expression data. These include the meta-clustering of gene-expression data combined with existing literature [20], using gene annotation to judge cluster quality [21], profiling gene groups with based text information and applying text-mining techniques for organizing and integrating large amounts of available information [22].

3 Framework Definition

We propose a novel system framework which integrates the information located and obtained by the large number of medical and biological database systems available

with the actual gene clustering process. Existing information about a genes molecular function, the biological process in which it takes part and the cellular component in which it resides is retrieved and encapsulated within the actual clustering process. Including the vast amounts of available literature within the analysis of gene expression data offers the opportunity to incorporate functional and other types of information about the genes when creating, validating and interpreting the resulting gene clusters.

Instead of assigning each gene randomly into a cluster, genes are initially grouped together according to function or biological process. Their respective gene expression vectors are then processed and genes are moved among the existing clusters until the algorithm terminates. The functional categorization of genes prior to the actual clustering process effectively minimizes the number of iterations the algorithm is executed since functional similarity between genes often implies similar expression patterns and vise versa. In contrast with a conventional approach, genes are moved less times among the existing clusters since their initial categorization is not done at random but based on existing published knowledge instead.

3.1 The underlying methodology

The proposed methodology comprises two main steps which form an iterative process. First, the system processes the given expression data set and identifies all relevant entities and genes that are involved in the current experiment. Their respective biological knowledge is then retrieved from well-known available databases, is indexed and stored locally. More specifically, the system utilizes three distinct types of databases: primary sequence databases (i.e. GenBank, DDJB), secondary sequence databases (i.e. UNIgene, TIGR) and genomic databases (i.e. Ensembl, GDB).Information about the molecular function, biological process and cellular component in which each gene composing the experiment is retrieved in the form of widely accepted GO keywords [23].

This ensures that relevant existing knowledge on the genes composing the given data set is taken into consideration. Alternatively, in the event that a certain gene has not been previously investigated and thus lacks any relevant background literature, *homologue* associations can be identified and their respective literature retrieved. Additionally, homologue associations act as secondary references during the assessment of the cluster functional coherence.

The retrieved information is utilized to calculate the functional similarity between the genes in the data set and initially place them accordingly. The expression data is initially sorted based on the genes functional category or biological process as dictated by the relevant GO annotations retrieved. In the event that a gene has no relevant information associated with it and no homologues can be identified, it is placed in a cluster randomly chosen.

The initial k number of clusters is calculated by scanning the functional distribution of the genes in the microarray and selecting the most overrepresented functional categories available from the resulting set. Genes that have an unknown biological process

or category as well as outliers, which represent statistically underrepresented genes are randomly inserted into one of the existing clusters. A functional category is defined as overrepresented when the number of genes composing it is larger than the average number of genes forming all categories in the experiment.

Alternatively, k can be set to an initial value and then adjusted during the algorithm iterations by assessing the cluster functional similarity after each iteration, and comparing it to a predefined 'quality' threshold. Should the average functional coherence of the resulting clusters be below the threshold after a successful iteration, the value of k is shifted accordingly. This is discussed in more detail in the conclusions section of this paper.

A number of existing distance and semimetric distance metrics, such as Euclidean distance, Spearman correlation and Manhattan distance, are defined and available for use [24]. In this case, we use the centered Pearson correlation similarity coefficient r, taking values between -1 and +1. The main reason for choosing it is the ability to detect exact opposite expression vectors ($r = -1$) and expression vectors that are completely uncorrelated and independent between them ($r = 0$).

3.2 Gene Ontology terms

Existing information about a gene is retrieved from the available sources in the form of GO annotation terms. The GO set consists of a widely accepted and standardized gene annotation vocabulary used by scientists in order to express and define in a clear and concise manner certain attributes about a specific gene [23],[25].

Each ontology is structured in a manner that specific terms are considered children of more broad terms. Additionally, in order to appropriately model biological data, the structure developed also supports many-to-many relationships such as potential node within the ontology can have many parents and many children, all connected with relationships between them. The selected terms are then organized into directed acyclic graphs, forming a complete network of interconnected terms describing specific genes properties.

An $n \times n$ term distance matrix is created which contains the absolute minimum distance between any two GO terms contained within the database. This is essential in order to calculate the functional similarity between two possible genes in the experiment. This is achieved by sequentially parsing every possible path of the GO ontology using nested queries and inner joins and eventually creating a separate table with the distance information. Two GO terms are considered identical when they have a distance of zero.

3.3 Gene functional similarity

In order to calculate the functional similarity between the genes composing the microarray experiment, a functional distance matrix is created containing the distance between any two genes in the microarray experiment. Using the previously con-

structed term matrix, the functional distance between the specific genes composing the experiment is extracted and stored locally. Since GO terms essentially specify a genes biological function or goal, their relevant term graph distances reflect their actual biological similarity and are used as a metric to weight the relationships between them. This is mainly achieved due to the GO ontologies hierarchical structure and multiple parent-child relationships.

The example below illustrates the relationship between the GO annotation terms assigned to three separate yeast genes, more specifically APN1, CDC2 and LCD1 as extracted by the SGD [23]. Genes with a similar biological goal have a relatively small distance value between them: APN1 (*DNA repair*) and CDC2 (*mismatch repair*) have a distance of 1 ; genes that do not share common or similar biological goals have a larger distance value of 8: LCD1 (*establishment of protein localization*) and CDC2 (*mismatch repair*).

Utilizing the created functional similarity matrix, genes are initially grouped according to their 'annotation term distance'. This is achieved by randomly selecting one gene from each of the overrepresented functional categories and using it as a starting node for the initial groups. Using information extracted from the functional similarity matrix, genes are placed in each group according to their relevant distance. Genes which have an unknown molecular function (GO:0005554) or take part in an unknown biological process (GO:0000004) are randomly placed into one of the existing clusters.

Table 1. A merged version of both matrices created: the generalized GO term distance matrix and the functional similarity matrix between the APN1, CDC2 and LCD1 yeast genomes

	GO:0006281 (APN1)	**GO:0006298 (CDC2)**	**GO:0045184 (LCD1)**
GO:0006281 (APN1)	*0*	*1*	*7*
GO:0006298 (CDC2)	*1*	*0*	*8*
GO:0045184 (LCD1)	*7*	*8*	*0*

3.4 Assessing cluster functional coherence

The functional coherence of each cluster can be directly deducted by the distance between the GO annotation terms of the gene which compose it. As mentioned above, should two genes has identical or similar molecular functions or biological goals, the minimum absolute distance between their respective annotation terms will be relatively small. In order to assess the functional coherence of each cluster, the arithmetic mean of the functional dispersion within the cluster (1).

$$C = \frac{\sum_{i=1}^{N} \chi i \psi i}{N^2} \quad (1)$$

Functional coherence C is defined as the sum of the respective absolute path distances between all genes composing the cluster divided by the total number of paths contained in the specific cluster. Functional coherent clusters have a small coherence value while highly skewed clusters have higher C values.

The above described approach operates under the hypothesis that the relevant functionality of the genes / ORF's composing the experiment has already been determined and mapped. During a typical microarray experiment however, a substantial number of genes have unknown or undetermined functionality in which case they are clustered solely on their relevant expression vectors. However, by assessing the functional coherency of a cluster, one can define the functional boundaries of a cluster with greater precision, giving a potential scientist greater insight into the exact function of an unidentified gene within it.

Alternatively, the functional coherence of each cluster could be calculated by utilizing the Neighbor Divergence Per Gene (NDPG) [26], [27], [28]. NDPG uses available scientific literature (corpus) in order to compute an information theoretical score which indicates how functionally coherent the group of genes under investigation is.

4 Experimental Validation

An initial sample domain was assembled in order to validate the proposed methodology and generate some preliminary experimental results. A subset consisting of 98 genes extracted from the original *Saccharomyces* data set used by Eisen et. al. [4], [29] was created. The [29] data set contains expression data during the yeast sporulation on 80 individual experiment conditions.

The genes within the data set were initially sorted and grouped together according to their broad biological goal. All existing knowledge was extracted from the SGD in the form of GO annotation terms and stored locally. The relevant paths between the existing genes were calculated and used to group the genes together into an initial group. Additionally, a small percentage of genes were intentionally mislabeled as participating in an *unknown biological process* in order to verify the algorithms ability to identify homologues.

The data set contained four overrepresented categories of annotated biological processes: DNA replication (GO:0006260), cell cycle (GO:0007049), protein biosynthesis (GO:0006412) and aerobic respiration (GO:0009060).

Based on the identified overrepresented biological goals extracted from the data set the value of k was manually set to 5 and genes were grouped accordingly. Genes marked as unknown where randomly placed into one of the existing groups. K-means clustering was then deployed using the predefined groups and Pearson Correlation

Coefficient as a distance metric. The algorithm terminated after 4 successful iterations.

From the five resulting clusters, clusters 1, 2 and 5 display the largest functional coherence based on the genes biological goal. DNA replication is the dominating category in cluster 1, protein biosynthesis in cluster 2 and finally cell cycle in cluster 5. The average coherence of the resulting clusters is 65.03%. Table 2 summarizes the characteristics of the resulting clusters. The calculated coherence metric can additionally be used in order to prioritize the resulting clusters for further examination.

Table 2. The resulting clusters and their characteristics

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
# of genes	*26*	*18*	*17*	*27*	*10*
% of genes	*27%*	*18%*	*17%*	*28%*	*10%*
GO:0006260	*18*	*0*	*1*	*4*	*2*
GO:0007049	*4*	*3*	*8*	*3*	*7*
GO:0006412	*2*	*15*	*2*	*5*	*0*
GO:0009060	*2*	*0*	*6*	*15*	*1*
C Score	*2.7928*	*1.388*	*3.7785*	*4.032*	*2.52*

5 Future directions and conclusions

In this paper we briefly examined the current methodologies and trends associated with the analysis of gene expression data obtained by performing microarray experiments. One of the most significant steps in analyzing gene expression data is clustering which involves grouping together genes into distinct, non-overlapping and biologically meaningful clusters. Both traditional and second generation algorithms were briefly discussed.

Including the vast amounts of available biological information into the actual data clustering and gene expression analysis operations still remains one of the most significant challenges. We propose a novel hybrid knowledge driven approach to clustering gene expression profiles which utilizes the relevant GO annotation terms associ-

ated with them. Genes are grouped together prior to clustering according to their molecular function, biological process or subcellular location. Additionally, the algorithm can also be deployed in order to assess the resulting clusters functional coherence. Finally, some preliminary experiments results obtained by a sample domain are also supplied and discussed.

One future direction currently developed is the design of a greedy functional quality algorithm for clustering gene expression data. Gene ontology annotation is used in order to judge the functional quality of the resulting clusters after each iteration. Using a predefined quality metric, the algorithm will be able to produce clusters that have a certain quality guarantee: the maximal functional distance between any two genes composing the cluster should not exceed a certain predefined threshold. This approach will eventually produce clusters with tightly related expression patterns and functional properties while discarding genes that do not appear to be coexpressed with any of the formed categories. Despite the fact that during microarray experiments the exact function of a substantial number of genes composing the dataset is undetermined, the above mentioned approach will explicitly narrow down the functional boundaries of a cluster, thus providing biologists with a more clear insight of the potential functionality of such ambiguous genes.

Finally, a promising idea for future extension, enhancing the accuracy of the functional coherence metric between the genes composing a cluster is the use of weighted word vectors. All annotation terms in the path from a specific gene to the topmost element of the hierarchy are tokenized and converted into a weighted word vector where each dimension represents the occurrence of a specific term. To quantify the functional similarity of any two genes, the cosine between their weighted word vectors is used. This will essentially offer a more precise view on the functional similarity between two individual genes.

References

[1] Quackenbush, J., Computational analysis of microarray data, *Nature Genetics*, Vol. 2, 2001.

[2] Amir, B., Zohar, Y., Clustering gene expression patterns, *Journal of Comp. Biology*, 1999.

[3] Heyer, J.L., Kruglyak, S., Yooseph, S., Exploring Expression Data: Identification and Analysis of Coexpressed genes, *Genome Research*, 9:1106-1115, 1999.

[4] Eisen, M., Spellman, P., Brown., P., Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci*., 14863-14868, 1998.

[5] Raychaudahri, S., Chang, J.T., Imam, F., Altman, R.B., The computational analysis of scientific literature to define and recognize gene expression clusters, *Nucleic Acids Research*, Vol. 31., No. 15., 4553-4560, 2003.

[6] Moreau, Y., De Smet, F., Thijs, G., Marchal, K., De Moor, B., Functional Bioinformatics of microarray data: from expression to regulation, *Proceedings of the IEEE*, Vol. 90., No. 11, 2002.

[7] Cho R. J., Cambell, M.J., Winzeler, E.A., Steinmetz, A., Conway, A. et al., A genome wide transcriptional analysis of the mytiotic cell cycle, *Mol. Cell*., vol. 2., 65-73, 1998.

[8] Tou, J.T., Gonzalez, R.C., Pattern classification by distance functions, Pattern Recognition Principles, Reading, Adison-Wesley, 1979.

[9] Tamayo, P., Slonim, D., Mesirov, Q., et al., Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci*., vol. 69, 2907-2912, 1999.

[10] Herrero, J., Valencia, A., Dopazo, J., A hierarchical unsupervised growing neural network for clustering gene expression patterns, *Bioinformatics*, vol. 17., 126-136, 2001.

[11] De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B., Moreau, Y., Adaptive quality based clustering of gene expression profiles, *Bioinformatics*, Vol. 18, No. 5, 735-746, 2002.

[12] Ghosh, D., Chinnaiyan A.M., Mixture Modelling of gene expression data from microarray experiments, *Bioinformatics*, Vol. 18, 275-286, 2002.

[13] Lukashin A.V., Fuchs, R., Analysis of temporal gene expression profiles: Clustering by simulated annealing and determing the optimal number of clusters, *Bioinformatics*, Vol. 17, 405-414, 2001.

[14] Ben-Dor, A., Shamir, R., Zakhini, Z., Clustering gene expression patterns, *J. Comput. Biology*, Vol. 6, 281-197, 1999.

[15] Sharan, R., Shamir, R., CLICK: A clustering algorithms with application to gene expression data, *Proc. Int. Conf. Intell. Syst*. Mol. Biol., 2000.

[16] Jiang, D., Pei, J., Zhang, A., DHC: A density-based hierarchical clustering method for time-series gene expression data, *Proc. 3rd IEEE symposium on Bioinformatics*, 2003.

[17] Jiang, D., Pei, J., Zhang, A., Towards interactive exploration of Gene Expression Patterns, ACM *SIGKDD Explorations Newsletter*, vol. 5, no. 2, 2003.

[18] Altman, R.B., Raychaudhuri, S., Whole genome expression analysis: challenges beyond clustering, Cur. *Opin. Struct. Biol*., 340-347, 2001.

[19] Raychaudhuri, S., Chang, T.J, Imam, F., Altman, R.B., The computational analysis of scientific literature to define and recognize gene expression clusters, *Nucleic Acids Research*, Vol. 31, No. 15, 2003.

[20] Glenisson, P., Mathys, J., De Moor B., Meta-Clustering of Gene expression data and literature-based information, *SIGKDD Explorations Newsletter*, vol. 5, no. 2, 2003.

[21] Gibbons, F., Roth, F., Juding the quality of gene expression-based clustering methods using gene annotation, 12:1574-1581, *Genome Research*, 2001.

[22] Glenisson, P., Coessens, B., Van Vooren, S., Moreau, B.., TXTGate: profiling gene groups with text-based information, *Genome Biology*, 5:R43, 2004.

[23] Dwight, S., Harris, M., Dolinski, K., et. al., Saccharomyces Genome Database (SDG) provides secondary gene annotation using the Gene Ontology (GO), *Nucleic Acids Research*, Vol. 30, No. 1, 2002.

[24] Stekel, D., Microarray bioinformatics, *Cambridge University Press*, (2003).

[25] The Gene Ontology Consortium, Gene Ontology: tool for the unification of biology, *Nature Genetics, Vol. 25*, (2000).

[26] Raychaudhuri, S, Altman, R.B., ShuEtze, H.S., Inclusion of textual documentation in the analysis of multidimensional data sets: application to gene expression data, *Machine Learning*, 2003.

[27] Raychaudhuri, S, Altman, R.B., A literature-based method for assessing the functional coherence of a gene group, *Bioinformatics*, 396-401, 2003.

[28] Raychaudhuri, S, Altman, R.B., ShuEtze, H.S., Text analysis of scientific literature can automatically determine if a group of genes share a common biological function, *Genome Res*., 2003.

[29] Chu, S., DeRisi, J., Eisen, M., Mullholland, J., Botstein, D., Brown, P.O., Herskowitz, I., The transcriptional program of sporulation in Budding Yeast, *Science*, Vol. 282, 1998.