An Automatic Gene Ontology Software Tool for Bicluster and Cluster Comparisons

Fadhl M. Al-Akwaa and Yasser. M. Kadah

Abstract— **We propose an Automatic Gene Ontology (AGO) software as a flexible, open-source Matlab software tool that allows the user to easily compare the results of the bicluster and cluster methods. This software provides several methods to differentiate and compare the results of candidate algorithms. The results reveal that bicluster/cluster algorithms could be considered as integrated modules to recover the interesting patterns in the microarray datasets. The further application of AGO could to solve the dimensionality reduction of the gene regulatory networks.**

Availability: AGO and help file is available at
http://home.k-space.org/FADL/Downloads/AGO_prgram.zip

I. Introduction

One of the main research areas of bioinformatics is functional genomics, which focuses on the interactions and functions of each gene and its products (mRNA and protein) through the whole genome (the entire genetic sequences encoded in the DNA and responsible for the hereditary information). In order to identify the functions of certain gene, one should be able to capture the gene expressions that describe how the genetic information is converted to a functional gene product through the transcription and translation processes. Functional genomics uses microarray technology to measure the genes expression levels under certain conditions and environmental limitations. In the last few years, microarray technology has become a central tool in biological research. Consequently, its corresponding data analysis methods have become among the most important research disciplines in bioinformatics. The analysis of microarrays data poses a large number of exploratory statistical aspects including clustering and biclustering algorithms, which help identify similar patterns in gene expression data and group genes and conditions into subsets that share biological significance. Recent understanding of cellular processes lead to the expectation of certain subsets of genes to be co-regulated and co-expressed under certain experimental conditions, and to behave almost independently under other conditions [1]. A bicluster is defined as a subset of genes that exhibit compatible expression patterns over a subset not all the conditions [2].

There are many biclustering algorithms, which differ in their approaches, computational complexity and prediction ability [3]. Examples of such methods include Bimax [1,2], OPSM [4], SAMBA [5], MSBE [6], BBC [7], CC [8], Bivisu [9] and ISA [10, 11], which are the most common bicluster Algorithms. BicAT [1] is a common biclustering analysis toolbox in which most of the above algorithms were implemented. BicAT were used in [2] to compare between different biclustering/clustering methods. A synthetic dataset was used to study the effects of noise and the regulatory complexity on the performance of the biclustering methods. Researchers are mostly interested in how different biclustering methods output for real datasets. Using real dataset. many studies [2,6,12] attempt to used Gene Ontology (GO) to investigate whether a set of genes discovered by bi-clustering/clustering methods presents significant enrichment with respect to a specific GO annotation provided by Gene Ontology Consortium[13]. BINGO [14], FUNCAT [15], GeneMerge [16] and FuncAssociate [17] are the most popular tools that analyze GO term enrichment in a given gene set. A comprehensive list can be found in [14]. The limitation of these programs is that the user is required to manually input every bicluster/cluster then use a word processing program to extract the interesting patterns, which is both time consuming and hard to do for the user. In this work, we propose a software package that provides a wide range of comparisons between the bicluster/cluster algorithm, resulting in a simpler and faster utility and allowing such comparisons to be readily accessible to many more users.

II. METHODOLOGY

A. AGO Implementation

AGO has been developed under Matlab 7.2. It runs on a PC with P4 1.8GHz CPU and 2 G Bytes of memory. The software inputs are the bicluster/cluster results as shown in Fig. 1. BicAT toolbox [1] was used to get the bicluster/cluster results of the OPSM, CC, ISA and K-means whereas Bivisu Matlab program [9] was used to get Bivisu bicluster results.

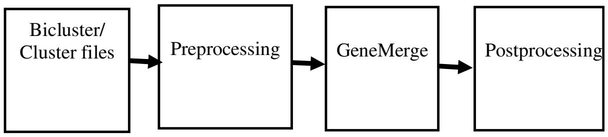

Fig. 1. AGO Block diagram

F. Alakwa is with the University of Science and Technology, Sana'a, Yemen, and the Biomedical Engineering Department, Cairo University (e-mail: f_alakwa@k-space.org).

Y. M. Kadah with the Biomedical Engineering Department, Cairo University and Center for Informatics Science (CIS), Nile University, Egypt (e-mail: ymk@k-space.org).

978-1-4244-2756-7/09/$25.00 ©2009 IEEE

In the preprocessing step AGO creates a file for each biclusters/clusters that contains the Open Read Frame (ORF) within each bicluster/cluster. Then these files are entered to the GeneMerge Perl program [16], which estimates the Gene Ontology term for each bicluster/cluster. GeneMerge provides a statistical test for assessing the enrichment of each GO term in the sample test. The basic question answered by this test is as described by [14] as "when sampling *X* genes (test set) out of *N* genes (reference set, either a graph or an annotation), what is the probability that *x* or more of these genes belong to a functional category *C* shared by *n* of the *N* genes in the reference set?" The hyper-geometric test, in which sampling occurs without replacement, answers this question in the form of *p*-value. Its counterpart with replacement, the binomial test, which provides only an approximate *p*-value, but requires less calculation time.". Finally, the postprocessing step involves asking the user to choose the suitable comparison methods of preference to be used.

TABLE I
Parameter settings for different bi-cluster/cluster methods and post-filtration. The definition of these parameters are defined in their original publications ([10],[8],[4],[12]).

Bi/clustering Algorithm	Parameter settings
ISA	$t_g = 2.0$, $t_c = 2.0$, seeds = 500
CC	$\delta = 0.5$, $\alpha = 1.2$, $M = 100$
OPSM	$l = 100$
BiVisu	$E = 60$, $N_r = 10$, $N_c = 5$, $P_o = 25$
K-means	K=100
PostFiltartion	Maximum overlapped allowed=25% Maximum number of biclusters/clusters=100 Min # of Rows =10 Min # of Columns=5

As the number of generated biclusters varies strongly among the considered methods, a filtration procedure similar to [2,6,12] is used and the algorithm output is used to provide a common baseline for comparison. We filtered out the bi-clusters/clusters with over 25% overlapped elements and output the largest 100 biclusters as shown in Table I.

B. Bicluster/Cluster Algorithms selections

Four biclustering algorithms are considered in this study according to their prediction strength, their promising results, to what they extend in the community and the feedback from their authors to explain some ambiguous issues. Also traditional clustering techniques such as K-means are included to compare with the bi-clustering methods.

C. *Parameter selection*

We used the default parameters as authors recommend in their publications as listed in Table I.

D. *Data set selections*

Saccharomyces cerevisiae is a widely used model organism in science. Also, it is a one of the most studied (along with E. coli). *S. cerevisiae* has obtained this important position because of its established use in industry (e.g. beer, bread and wine fermentation, ethanol production). Additionally, yeasts are comparatively similar in structure to human cells and both are eukaryotic creatures as well, not prokaryotes (bacteria and archaea). Furthermore, Many important proteins in the human body were first discovered by studying their Homology in yeast; these proteins include cell cycle proteins, signaling proteins, and protein-processing enzymes. Finally, the petite mutation in *S. cerevisiae* is considered as one of most particular interest in molecular biology[18]. Table II exhibits the most common yeast time series data sets that have been studied.

TABLE II
Analysis of the common time series *Yeast* data sets

Author	Data Size	% Missing value	% genes with small variance[a]	% genes with low Absolute expression values[b]	% genes whose profiles have low entropy[c]
[20]	2467 x 79	1.93	78	69.3	15
[19]	2993 x 173	3	9.9	33.4	15
[21]	2884 x 17	5.8	10	0.1734	15

[a] Gene profiling experiments have genes that exhibit little variation in the profile and are generally out of interest in the experiment . These genes are commonly removed from the data.So, genes with small profile less than the 10th percentile are removed.

[b] Gene expression profile experiments have data where the absolute values are very low. The quality of this type of data is often bad due to large quantization errors or simply poor spot hybridization. So,gene profiles with low absolute values less than log2(4) are removed.

[c] Remove genes with low entropy expression values less than the 15th percentile

The analysis done in this paper is performed on the gene expression data of *S. cerevisiae* provided by Gasch [19] which contains 2993 genes and 173 conditions because of the variety of experimental conditions and the low percentage of genes with small variance, see table II.

III. RESULTS & DISCUSSION

Table III demonstrates the statistical parameters of the biclustering/clustering results after filtration. They differ in the number of bicluster/cluster outputs, the number of genes and conditions within each bicluster/cluster.

Comparing biclusters results is not straightforward task because each method searches for biclusters/clusters with different structure and mechanism. Consequently, the number of produced biclusters/clusters differs from each method to another.

AGO provides different reasonable methods for comparison. For instance, we can compare the percentage of overrepresented biclusters/clusters in one or more GO annotation. A bicluster/cluster is said to be overrepresented in a functional category if it gives a small *p*-value [12].

Fig. 2 shows the proportion of bi-clusters/clusters for each method in which one or several GO categories are overrepresented at different levels of significance. By comparing Fig. 2 and the Fig. 3 in [2], we found that the percentages of enriched biclusters for the matched algorithms are almost the same. This does validate the results of the proposed comparative tool. Investigating both figures, we observed that OPSM algorithm gave a high portion of functionally enriched biclusters at all significance levels (from 85% to 100%) ; this is mainly because it has a few number of biclustering results(Table III). Next to OPSM, ISA shows relatively high portions of enriched biclusters.

Based on the output from many simulations, we found that most functionally enriched biclusters had low number of annotated genes. AGO gives the user the option to filter the enriched biclusters/clusters. For instance, in Fig. 3 we report the percentage of enriched biclusters/clusters after neglecting the biclusters/clusters which have less than ten genes in each GO category or have study fraction less than 50% (In GeneMerge program at GO:0008150, if the study fraction = (6/12) that means over the twelve test cluster genes there are six genes which share the same function in category GO:0008150).

TABLE III
STATISTICAL COMPARISON OF BICLUSTERS/CLUSTERS IDENTIFIED IN THE YEAST CELL-CYCLE DATASET USING VARIOUS ALGORITHMS

Algorithm	No of Bi/clusters	Bicluster/Cluster Sizes[a]	
		Min	Max
ISA	9	50 x 35	155 x 37
CC	69	11 x 5	2259 x 134
OPSM	2	11 x 15	575 x 6
BiVisu	100	27 x 142	99 x 52
K-means	100	20 x 173	50 x 173

[a] The size of a bicluster is determined by its number of values, i.e. product of numbers of rows and columns

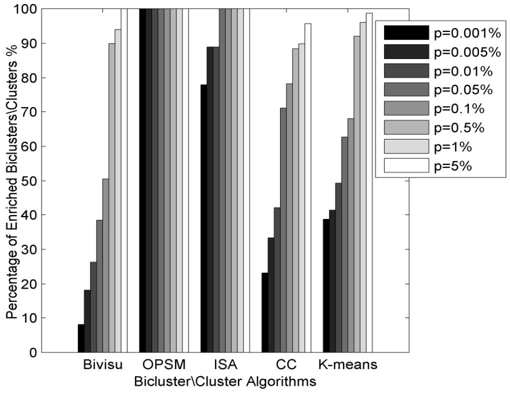

Fig. 2. Percentage of significantly enriched biclusters/clusters by GO Biological Process category (S. cerevisiae) for the five selected biclustering methods and Kmeans at different significance levels p.

Fig. 3 shows that OPSM and ISA have highly enriched biclusters/clusters that have large number of genes per each GO category. On the other hand, Bivisu biclusters are strongly affected by this filtration and they contain a lower number of genes per each category.

AGO also has an option to make the user to compare extract the interesting patterns. Genes whose transcription is responsive to a variety of stresses have been implicated in a general yeast response to stress. Other gene expression responses appear to be specific to particular environmental conditions[19].

The conditions applied in Gasch experiments varied from temperature shocks, hydrogen peroxide, the peroxide-generating drug menadione, the sulfhydryl-oxidizing agent diamide, the disulfide-reducing agent dithiothreitol, hyper- and hypo-osmotic shock, amino acid starvation, nitrogen source depletion, and progression into stationary phase. For more details see the supplementary material. In Table IV AGO was used to summarize the difference between the biclusters/clusters contents based on the conditions of Gasch experiments [19]. The GO term definitions were been summarized in the supplementary material. Although OPSM show high percentage level of enriched biclusters (as shown in Fig. 2 and Fig. 3), its biclusters do not contain any genes within any GO category response to Gasch experiments (Table IV). The explanation of this behavior is still under investigation. Although the low number of ISA biclusters (9 biclusters), it's biclusters show large number of enriched GO category for instance, one of the ISA biclusters contains 11 genes which are sharing the **GO:0006979** which is defined as the response to oxidative stress (the change in state or activity of a cell or an organism in terms of movement, secretion, enzyme production, gene expression, etc. as a result of oxidative stress, a state often resulting from exposure to high levels of reactive oxygen species, e.g. superoxide anions, hydrogen peroxide (H_2O_2), and hydroxyl radicals [22], which are the same applied conditions on Gasch experiments.

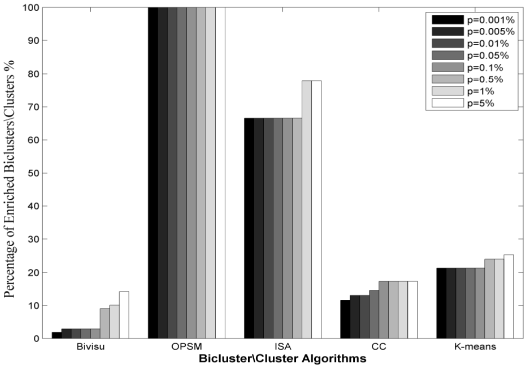

Fig. 3 Percentage of significantly enriched biclusters by GO Biological Process category by setting the allowed minimum number of genes per each GO category to 10 and the study fraction to large than 50%.

The K-means cluster results distinguished a unique GO category which is **GO:0000304** (response to singlet oxygen), while Bivisu shows a unique **GO:0042542** (response to hydrogen peroxide) and CC **GO:0042149** (cellular response to glucose starvation). This indicates that each Algorithm has its unique predictability of extraction. The powerful usage of these bi/cluster algorithms is significantly appeared in **GO:0006995** (cellular response to nitrogen starvation) where these algorithms were able to discover 4 out of 5 annotated genes without any prior biological information or on desk experiments.

TABLE IV
Gene Ontology category per number of annotated genes of the Bicluster/cluster algorithm results for the experimental condition on Gasch Experiments[21].

GO Term[a] / (number of annotated genes)	**K-means**	**CC**	**ISA**	**Bivisu**	**OPSM**
GO:0042493 Response to drug / (**118**)	4	5	7	6	0
GO:0006970 response to osmotic stress / (**83**)	3	5	6	3	0
GO:0006979 response to oxidative stress / (79)	2	7	11	0	0
GO:0046686 response to cadmium ion / (102)	2	3	2	2	0
GO:0043330 response to exogenous dsRNA / (7)	2	3	2	2	0
GO:0046685 response to arsenic / (77)	2	0	2	2	0
GO:0006950 response to stress / (532)	9	11	16	2	0
GO:0009408 response to heat / (24)	3	0	2	2	0
GO:0009409 response to cold / (7)	0	0	2	0	0
GO:0009267 cellular response to starvation / (44)	0	2	0	0	0
GO:0006995 cellular response to nitrogen starvation / (5)	4	4	4	0	0
GO:0042149 cellular response to glucose starvation / (5)	0	2	0	0	0
GO:0009651 response to salt stress / (15)	2	7	0	0	0
GO:0042542 response to hydrogen peroxide /(5)	0	0	0	2	0
GO:0006974 response to DNA damage stimulus / (240)	0	22	0	3	0
GO:0000304 response to singlet oxygen / (4)	2	0	0	0	0

[a] The GO Term definitions were been summarized at the supplementary file.

IV. CONCLUSION

We proposed AGO as an open-source software that runs under Matlab. It has the advantage of being easily extended or modified. AGO users can view and alter the algorithms, as well as add or substitute modules for those who want to make their own comparison strategy. The comparison methodology used in this study confirms that the bicluster and cluster algorithms can be considered as integrated techniques. It was not possible to find one particular algorithm that can extract all the interesting patterns. That is, what algorithm A succeeds to recover in certain data sets, algorithm B may fail, and vice verse. An attempt will be made to embed AGO in the BicAT which will be quit helpful for researchers who are looking for predetermined patterns. The extended work could be achieved by using the highly enriched bicluster/clusters to solve the dimensionality reduction problems of the Gene Regulatory Network.

REFERENCES

[1] S. Barkow, S. Bleuler, A. Prelic, P. Zimmermann, and E. Zitzler, "BicAT: a biclustering analysis toolbox," *Bioinformatics,* vol. 22, pp. 1282-1283, May 15, 2006 2006.

[2] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, "A Systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics,* vol. 22, pp. 1122 - 1129, 2006.

[3] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE/ACM Trans Comput Biol Bioinform,* vol. 1, pp. 24 - 45, 2004.

[4] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, "Discovering local structure in gene expression data: the order-preserving submatrix problem," *Journal of Computational Biology,* vol. 10, pp. 373 - 384, 2003.

[5] A. Tanay, R. Sharan, and R. Shamir, "Discovering statistically significant biclusters in gene expression data," *Bioinformatics,* vol. 18, pp. S136-144, July 1, 2002 2002.

[6] X. Liu and L. Wang, "Computing the maximum similarity bi-clusters of gene expression data," *Bioinformatics,* vol. 23, pp. 50-56, January 1, 2007 2007.

[7] J. Gu and J. Liu, "Bayesian biclustering of gene expression data," *BMC Genomics,* vol. 9, p. S4, 2008.

[8] Y. Cheng and G. M. Church, "Biclustering of expression data," *Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology,* pp. 93 - 103, 2000.

[9] K. O. Cheng, N. F. Law, W. C. Siu, and T. H. Lau, "BiVisu: software tool for bicluster detection and visualization," *Bioinformatics,* vol. 23, pp. 2342 - 2344, 2007.

[10] J. Ihmels, S. Bergmann, and N. Barkai, "Defining transcription modules using large-scale gene expression data," *Bioinformatics,* vol. 20, pp. 1993 - 2003, 2004.

[11] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai, "Revealing modular organization in the yeast transcriptional network," *Nature Genetics,* vol. 31, pp. 370 - 377, 2002.

[12] K.-O. Cheng, N.-F. Law, W.-C. Siu, and A. Liew, "Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization," *BMC Bioinformatics,* vol. 9, p. 210, 2008.

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

[14] S. Maere, K. Heymans, and M. Kuiper, "BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks," *Bioinformatics,* vol. 21, pp. 3448-3449, August 15, 2005 2005.

[15] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhaupt, M. Munsterkotter, and H. W. Mewes, "The FunCat, a functional annotation scheme for

systematic classification of proteins from whole genomes," *Nucl. Acids Res.,* vol. 32, pp. 5539-5545, October 14, 2004 2004.

[16] C. I. Castillo-Davis and D. L. Hartl, "GeneMerge - post-genomic analysis, data mining, and hypothesis testing," *Bioinformatics,* vol. 19, pp. 891 - 892, 2003.

[17] G. F. Berriz, O. D. King, B. Bryant, C. Sander, and F. P. Roth, "Characterizing gene sets with FuncAssociate," *Bioinformatics,* vol. 19, pp. 2502-2504, December 12, 2003 2003.

[18] http://www.nationmaster.com/encyclopedia/Saccharomyces-cerevisiae

[19] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes," *Mol. Biol. Cell,* vol. 11, pp. 4241-4257, December 1, 2000 2000.

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

[21] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature Genetics,* vol. 22, pp. 281-285, 1999.

[22] http://www.geneontology.org/