BicAT_Plus: An Automatic Comparative Tool For Bi/Clustering of Gene Expression Data Obtained Using Microarrays

Fadhl M. Al-Akwaa[1,2], Mohamed H. Ali[4], Yasser M. Kadah[2,3]

[1]Biomedical Eng. Dept., Univ. of Science & Technology, Sana'a, Yemen (E-mail f_alakwa@k-space.org)
[2] Biomedical Engineering Department, Cairo University, Giza , Egypt
[3]Center for Informatics Sciences, Nile University, Egypt
[4]Computer Science School, Nottingham University, Nottingham, United Kingdom

Abstract

In the last few years the gene expression microarray technology has become a central tool in the field of functional genomics in which the expression levels of thousands of genes in a biological sample are determined in a single experiment. Several clustering and biclustering methods have been introduced to analyze the gene expression data by identifying the similar patterns and grouping genes into subsets that share biological significance. However, it is not clear how the different methods compare with each other with respect to the biological relevance of the biclusters and clusters as well as with other characteristics such as robustness and predictability. This research describes the development of an automatic comparative tool called *BicAT plus* that was designed to help researchers in evaluating the results of different bi/clustering methods, compare the results against each others and allow viewing the comparison results via convenient graphical displays. *BicAT plus* incorporates a reasonable biological comparative methodology based on the enrichment of the output bi/clusters with gene ontology functional categories. No exact algorithm can be considered the optimum one. Instead, bi/clustering algorithms can be used as integrated techniques to highlight the most enriched biclusters that help biologists to draw biological prediction about the unknown genes.

1. 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, protein) through the whole genome (the entire genetics sequences encoded in the DNA and responsible for the hereditary information). In order to identify the functions of certain gene, we should be able to capture the gene expressions which describe how the genetic information converted to a functional gene product through the transcription and translation processes. Functional genomics uses microarrays technology to measure the genes expressions levels under certain conditions and environmental limitations. In the last few years, Microarray has become a central tool in biological research, consequently, the corresponding data analysis becomes one of the important work disciplines in bioinformatics. The analysis of microarrays data poses a large number of exploratory statistical aspects including clustering and biclustering algorithms which help to identify similar patterns in gene expression data and group genes and conditions into subsets that share biological significance. There are several bi/clustering methods that have been proposed to achieve this target (see [1] for a survey) , but the question is: which algorithm is better? And do some algorithms have advantages over others. Generally, comparing different bi/clustering algorithms is not straightforward as they differ in strategies, approaches, time complicity, number of parameters and prediction ability. They are strongly influenced by user-selected parameter values. For these reasons, the quality of bi/clustering results is also often considered more important than the required computation time. Although there are some analytical comparative studies to evaluate the traditional clustering algorithms [2-4], for biclustering; no such extensive comparison exist even after initial trails have been taken[5].In the end, biological merit is the main criterion for evaluation and comparison between the various bi/clustering methods. BicAT [6] is a common biclustering analysis toolbox in which most important bi/clustering algorithms like k-means, SOM, HCL, Bimax [5], OPSM [7], X-motif [8],CC[9], and ISA [10] were implemented, see Figure 1. We have developed a comparative tool "Bicat_plus" that includes the biological comparative methodology and to be as an extension to the BicAT program. The Goal of *BicAT_plus* is to enable researchers and biologists to compare between the different bi/clustering methods based on set of biological merits and draw conclusion on the biological meaning of the results. Also *BicAT_plus* help researcher in comparing and evaluating the algorithms results multiple times according to the user selected parameter values as well as the required biological perspective on various datasets. *BicAT_plus* has many features added to BicAT which could be summarized in the following:

a. Adding more algorithms to the BicAT tool in order to have one software package that employs most of the commonly used bi/clustering algorithms. The additional algorithms are MSBE constant biclustering and MSBE additive biclustering.
b. Extending the BicAT to perform functional analysis using the three subontologies or categories of GO (biological process, molecular function and cellular component) and visualizing the enriched GO terms per each bi/cluster in a separate histogram.
c. Evaluating the quality of each bi/clustering algorithm results after applying the GO functional analysis and displaying the percentage of the enriched biclusters at the standard P-values (significance levels) which are: 0.00001,0.00005,0.0001,0.0005,0.001,0.005,0.01 and 0.05 .
d. Comparing between the different bi/clustering algorithms according to the percentage of the functionally enriched bi/clusters at the required significance levels, the selected GO category and with certain filtration criteria for the GO terms.
e. Evaluating and comparing the results of external bi/clustering algorithms (not included in the *BicAT plus* current version).This gives the *BicAT plus* the advantage to be a generic tool that doesn't depend on the employed methods only. For example; it can be used to evaluate the quality of the new algorithms introduced to the field and compare against the existing ones.
f. Displaying the analysis and comparison results using graphical and statistical charts visualizations in multiple modes (2D and 3D).

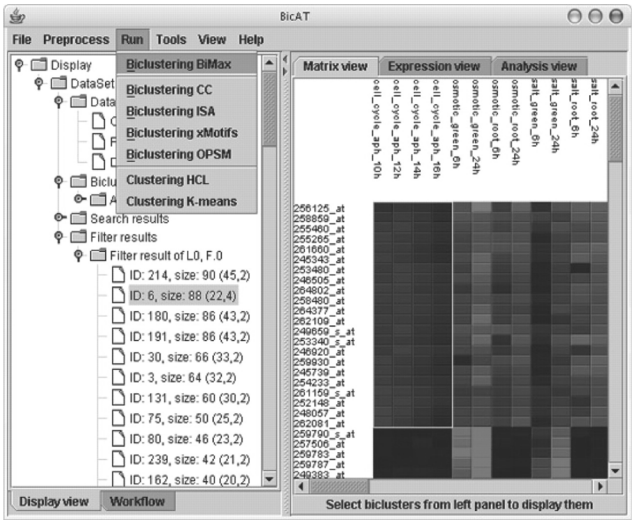

Fig. 1. Bi/clustering algorithms employed by BicAT [6].

2. Methodology

2.1. Software Development and Architecture

Before using the *BicAT plus*, Active Perl version 5.10 and Java Runtime Environment (JRE) version 6 are required to be installed on your machine. *BicAT plus* has been tested and show good performance on a PC machine with the following configurations: CPU: Pentium 4, 1.5 GHZ, RAM: 2.0 GB, Platform: windows XP professional with SP2.
BicAT plus is structured in the hierarchy of packages which are shown in figure 2. The highlighted blocks with dashed boundary are the additional modules developed for the comparative tool while the black ones are the original modules of the BicAT program. We faced many problems during the implementations like 1- lack of documentation of the BicAT tool which influenced the planned time to understand the source code and extend it. 2- All bugs reported about BicAT should be fixed in order to avoid its effect on the comparative tool. Ex: delete node from the navigation tree. 3- Technical problems like calling GeneMerge Perl script from java code. The used solution was to save the Perl commands in a batch file, then call the batch file from the java code using the *Runtime* class provided by SUN. 4- One of the objectives of this research was to enrich the BicAT (written using java) with more biclustering algorithms. But, some of these algorithms are written using C and C++. Thus, to solve such a compatibility problem, we converted the C files to dynamic link library (DLL) file then loaded it

to the system class path library. Another possible solution was to use the Java native interface (JNI) to call the C files.

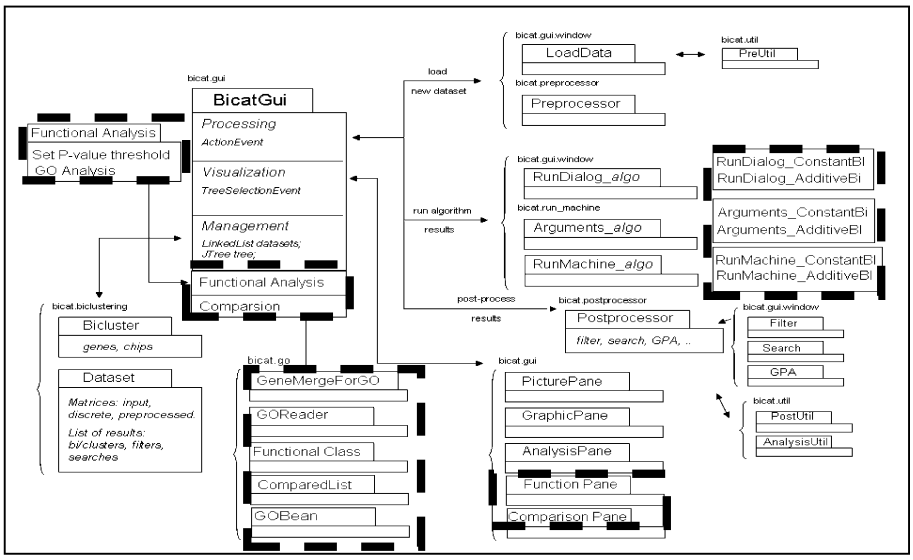

Fig. 2. The general design of the BicAT plus. Dashed block for the comparative tool packages and classes. The black entities are the original packages and interfaces of the BicAT program. Modified from [11].

2.2. GO Overrepresentation Programs

Many programs like: BINGO[12], FUNCAT[13], GeneMerge[14] and FuncAssociate[15] were used to investigate whether the set of genes discovered by bi-clustering/clustering methods present significant enrichment with respect to a specific GO annotation provided by Gene Ontology Consortium[16]. *BicAT Plus* used GeneMerge program as the most popular GO program. 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 follows: 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, provides only an approximate P-value, but requires less calculation time.

2.3. Comparative methodologies based on GO

BicAT plus provides reasonable method for comparing the results of different bi/clustering algorithms by:

2.3.1 identifying the percentage of enriched or overrepresented biclusters with one or more GO term per multiple significance levels (p-values) for each algorithm.

$$Percentage\ of\ enriched\ bicluster\ significance\ level = \frac{Number\ of\ enriched\ biclusters\ at\ this\ level}{total\ number\ of\ biclusters} \quad (1)$$

The definition of significance depends on the user selection of threshold p-values. A bi/cluster is said to be significantly overrepresented (enriched) with a functional category if the p-value of this functional category is lower than the preset threshold P-value [17, 5]. The results are displayed using a histogram for the entire compared algorithms at the different preset significance levels, and the algorithm which gives higher proportion of enriched bi/clusters per all significance levels is considered to be the optimum one as it does group effectively the genes sharing similar functions in the same bi/cluster.

2.3.2 Estimate Algorithms predictability power to recover interested pattern 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. *BicAT plus* make the user to

compare the perdictibility power of bi/clusters algorithms to interested pattern defined by the user see table 2 for an example.

2.4. Comparison Process Steps

The following process diagram shown in Figure 3 summarizes the required steps by the user to compare between the different algorithms using the *BicAT plus*.

1- download *BicAT_plus* from our site http://home.k-space.org/FADL/Downloads/BicAT_plus.zip.
2- Load Gene Expression Data to *BicAT plus* then run the selected five prominent bi/clustering methods with setting parameters as table II
3- Run GO comparison tool in the *BicAT plus* and add the available bi/clustering algorithms to the compared list as shown in Figure[4] .
4. Select the on of the available GO category e.g. biological process, molecular function and cellular components.
5. Select the P-values e.g. 0.00001, 0.0001, 0.01, 0.005, and 0.05.
6. Press compare button.
7. Press comparison menu, Functional enrichment and select 2D or 3D charts see Figure 5.

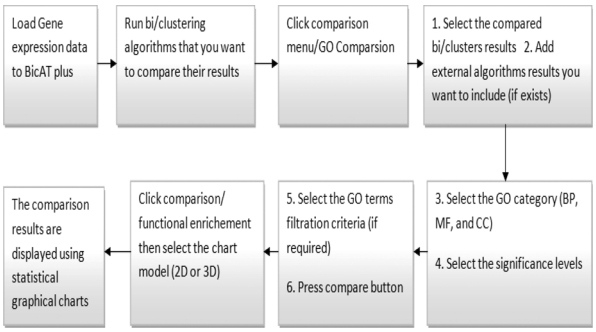

Fig. 3. *BicAT Plus* Comparison process steps.

Table 1. Default Parameter settings of the compared bi/clustering methods.
The definitions of these parameters are listed in their original publications ([10],[9],[18]) respectively.

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	$E = 0.82$, $N_r = 10$, $N_c = 5$, $P_o = 25$
K-means	K=100

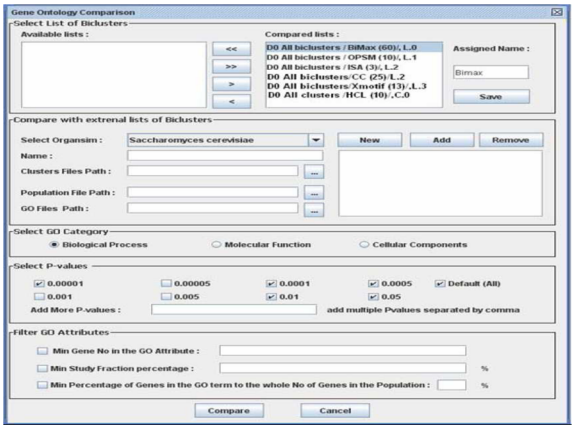

Fig. 4. *BicAT Plus* Comparison Dialog

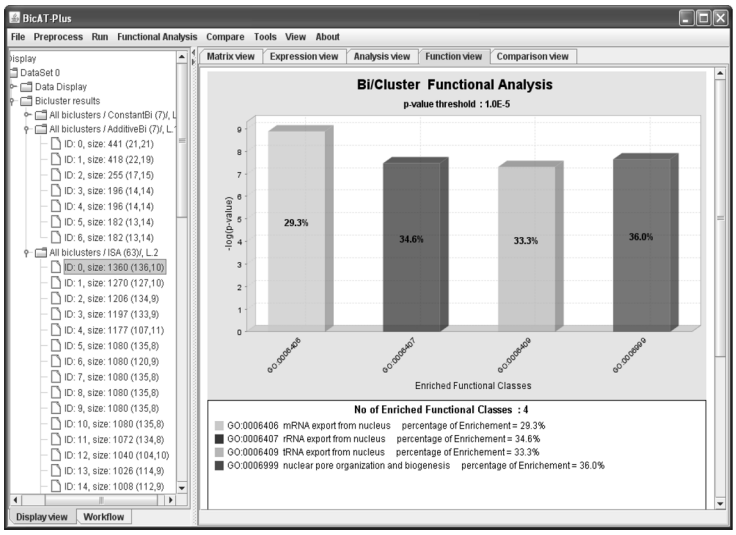

Fig. 5. Functional analysis results of the selected bi/cluster. Each column represents an enriched GO functional class. The height of the column is proportional to the significance of this enrichment

3. Results and Discussion

The above comparison steps is performed on the gene expression data of *S. cerevisiae* provided by Gasch [19]. The dataset contains 2993 genes and 173 conditions of diverse environmental transitions such as temperature shocks, amino acid starvation, and nitrogen source depletion. This dataset is freely available from [20]. For each bi/clustering algorithm we used the default parameters as authors recommend in their publications. See Table I.

3.1 The percentage of enriched function

After applying the above steps on Gasch data, *BicAT plus* produce the histogram shown in figure 1. By comparing Figure 6 and Figure 3 in [5] , we found that the percentage 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 %) . Next to OPSM, ISA and Bimax show relatively high portions of enriched biclusters.

In order to evaluate the ability of the algorithms to group the maximum number of genes whose expression patterns are similar and sharing the same GO category, we use the filtration criteria developed in the comparative tool by neglecting those bi/clusters which have study fraction less than 25%. The study fraction of a GO term is the fraction of genes in the study set (bicluster) with this term.

$$Study\ \ fraction\ \ of\ a\ \ GO\ \ term = \frac{No\ \ of\ \ genes\ sharing\ \ the\ \ GO\ term\ in\ a\ bicluster}{total\ number\ \ of\ \ genes\ in\ this\ bicluster} \times 100 \qquad (2)$$

Figure 7 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 contains a lower number of genes per each category. This filtration will help in identifying the powerful and most reliable algorithms which are able to group maximum numbers of genes sharing same functions in one cluster.

3.2 The predictability power to recover interested pattern

The user could compare bi/clusters algorithms based on which of them could recover defined pattern like which one of them could recover clusters which have response to the conditions applied in Gasch experiments. In Table 2, the difference between the biclusters/clusters contents were summarized. Although OPSM show high percentage level of enriched biclusters (as shown in Figures. 2 and 3), its biclusters do not contain any genes within any GO category response to Gasch experiments. The k-means and Bivisu cluster/bicluster results

distinguished a unique GO category, which is **GO:0000304** (response to singlet oxygen), and **GO:0042542** (response to hydrogen peroxide) 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.

4. Conclusions

We have introduced the *BicAT plus* with reasonable comparative methodology based on the Gene Ontology. To the best of our knowledge such an automatic comparison tool of the various bi/clustering algorithms has not been available in the literature. *BicAT plus* is an open source tool written in java swing and it has a well structured design that can be extended easily to employ more comparative methodologies that help biologists to extract the best results of each algorithm and interpret these results to useful biological meaning. In other words, the algorithms that show good quality of results (per the dataset) can be used to provide a simple means of gaining leads to the functions of many genes for which information is not available currently (unannotated genes).

Using *BicAT plus*, we can identify the highly enriched bi/clusters of the whole compared algorithms. This might be quite helpful in solving the dimensionality reduction problem of the Gene Regulatory Network construction from the gene expression data. This problem originates from the relatively few time points (conditions or samples) with respect to the large number of genes in the microarray dataset.

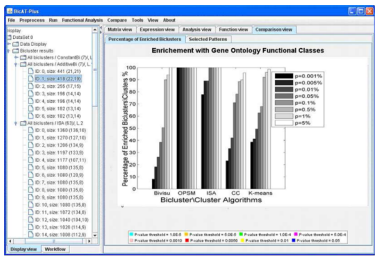

Fig. 6. Percentage of biclusters significantly enriched by GO Biological Process category (*S. cerevisiae*) for the five selected biclustering methods and K-means at different significance levels p.

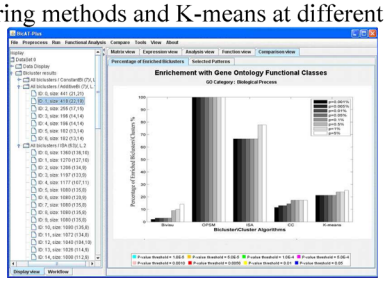

Fig. 7. Percentage of significantly enriched biclusters by GO Biological Process category by setting the allowed minimum number of genes per each GO category to 10 and the study fraction to large than 50%.

Table 2
Gene Ontology category per number of annotated genes of the Bicluster/cluster algorithm results for the experimental condition on Gasch Experiments[19].

GO Term / (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

Acknowledgments

The authors thank S. Barkow and C. I. Castillo-Davis for the BicAT Toolbox and GeneMerge Program respectively. Fadhl Al-Akwaa was supported by The University of Science &Technology, Sana'a - Yemen.

References

[1] 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.

[2] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, "Validating clustering for gene expression data," *Bioinformatics,* vol. 17, pp. 309-318, April 1, 2001.

[3] S. Datta and S. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data," *Bioinformatics,* vol. 19, pp. 459-466, March 1, 2003.

[4] F. Azuaje, "A cluster validity framework for genome expression data," *Bioinformatics,* vol. 18, pp. 319-320, February 1, 2002

[5] 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.

[6] 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

[7] 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.

[8] T. M. Murali and K. S, "Extracting conserved gene expression motifs from gene expression data. ," *Pac. Symp. Biocomput*. vol. 8, pp. 77–88, 2003

[9] 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.

[10] 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.

[11] http://www.tik.ee.ethz.ch/sop/bicat/?page=developersGuide.php

[12] 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, 2005 .

[13] 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, 2004.

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

[15] 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.

[16] 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.

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

[18] 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.

[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.

[20] http://genome-www.stanford.edu/yeast_stress/