African Journal of Biotechnology Vol. 11(7), pp. 1744-1753, 24 January, 2012
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB10.1523
ISSN 1684–5315 © 2012 Academic Journals

Full Length Research Paper

Analysis and visualization of gene expression data using biclustering: A comparative study

Fadhl M. Al-Akwaa

Biomedical Engineering Department, University of Science and Technology, Sana'a, Yemen. E-mail: fadlwork@gmail.com. Tel: 00967-733226746, 00967-777012076.

Accepted 13 December, 2011

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 described the development of an automatic comparative tool called BicAT-plus that was designed to help researchers in evaluating the results of different biclustering methods. It also compared the results against each other and allowed a comparison of results via convenient graphical displays. BicAT-plus incorporates a reasonable biological comparative methodology based on the enrichment of the output biclusters with gene ontology functional categories. No exact algorithm can be considered the optimum one. Instead, biclustering algorithms can be used as integrated techniques to highlight the most enriched biclusters that help biologists to draw biological prediction about the unknown genes.

Key words: Bioinformatics, functional genomics, gene expression analysis, microarrays data, comparison, clustering, biclustering, functional analysis, gene ontology.

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 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 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 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 has become one of the important work disciplines in bioinformatics. The analysis of microarray 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 biologically significance. There are several biclustering methods that have been proposed to achieve this target; Madeira and Oliveira (2004) compared the most common algorithms, but the questions still asked are: which algorithm is better and what are the disadvantages of each algorithm?

Recently, Kevin et al. (2006) proposed a semantic web algorithm to recommend the best algorithm based on user inputs like: is the dataset contain outliers, is it allowed to get overlapped clusters and the time to retrieve the biclusters. Generally, comparing different biclustering algorithms is not straightforward as they differ in strategies,

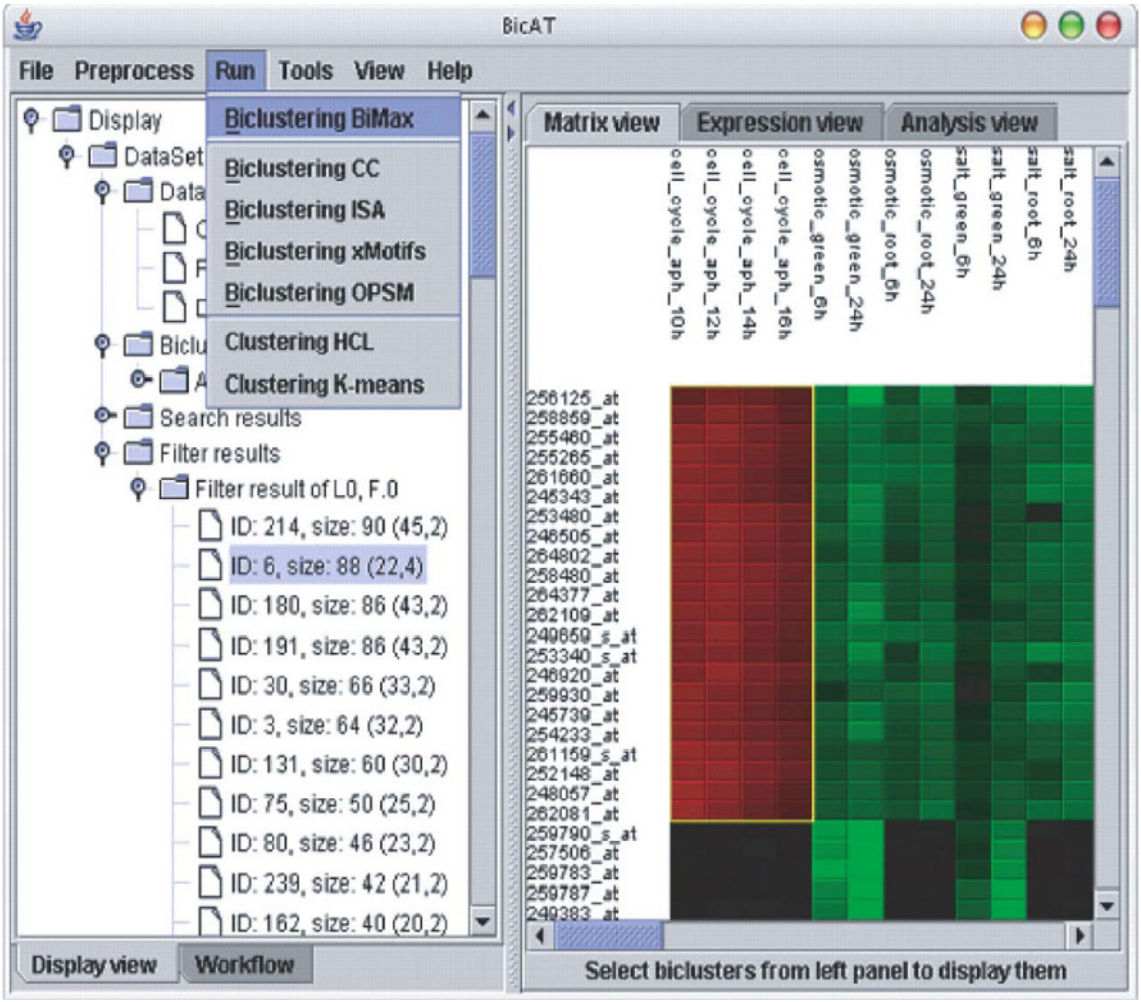

Figure 1. Biclustering algorithms employed by BicAT (Barkow et al., 2006).

approaches, time complicity, number of parameters and prediction ability. Also they are strongly influenced by user selected parameter values. For these reasons, the quality of biclustering results is often considered more important than the required computation time. Although, there are some analytical comparative studies to evaluate the traditional clustering algorithms (Yeung et al., 2001; Datta and Datta, 2003; Azuaje, 2002), for biclustering, no such extensive comparison exist even after initial trails have been taken (Prelic et al., 2006). In the end, biological merit is the main criterion for evaluation and comparison between the various biclustering methods.

BicAT (Barkow et al., 2006) is a common biclustering analysis toolbox in which most important biclustering algorithms like k-means, HCL (Szeto et al., 2003), Bimax (Prelic et al., 2006), OPSM (Ben-Dor et al., 2003), X-motif (Murali et al., 2003), CC (Cheng and Church, 2000) and ISA (Ihmels et al., 2002, 2004) were implemented (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 biclustering methods based on set of biological merits, and draw conclusion on the biological meaning of the results. Also BicAT-plus help researchers 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 as follows:

Adding more algorithms to the BicAT tool in order to have one software package that employs most of the commonly used biclustering algorithms. The additional algorithms are MSBE constant biclustering and MSBE additive biclustering (Liu and Wang, 2007).

Extending the BicAT to perform functional analysis

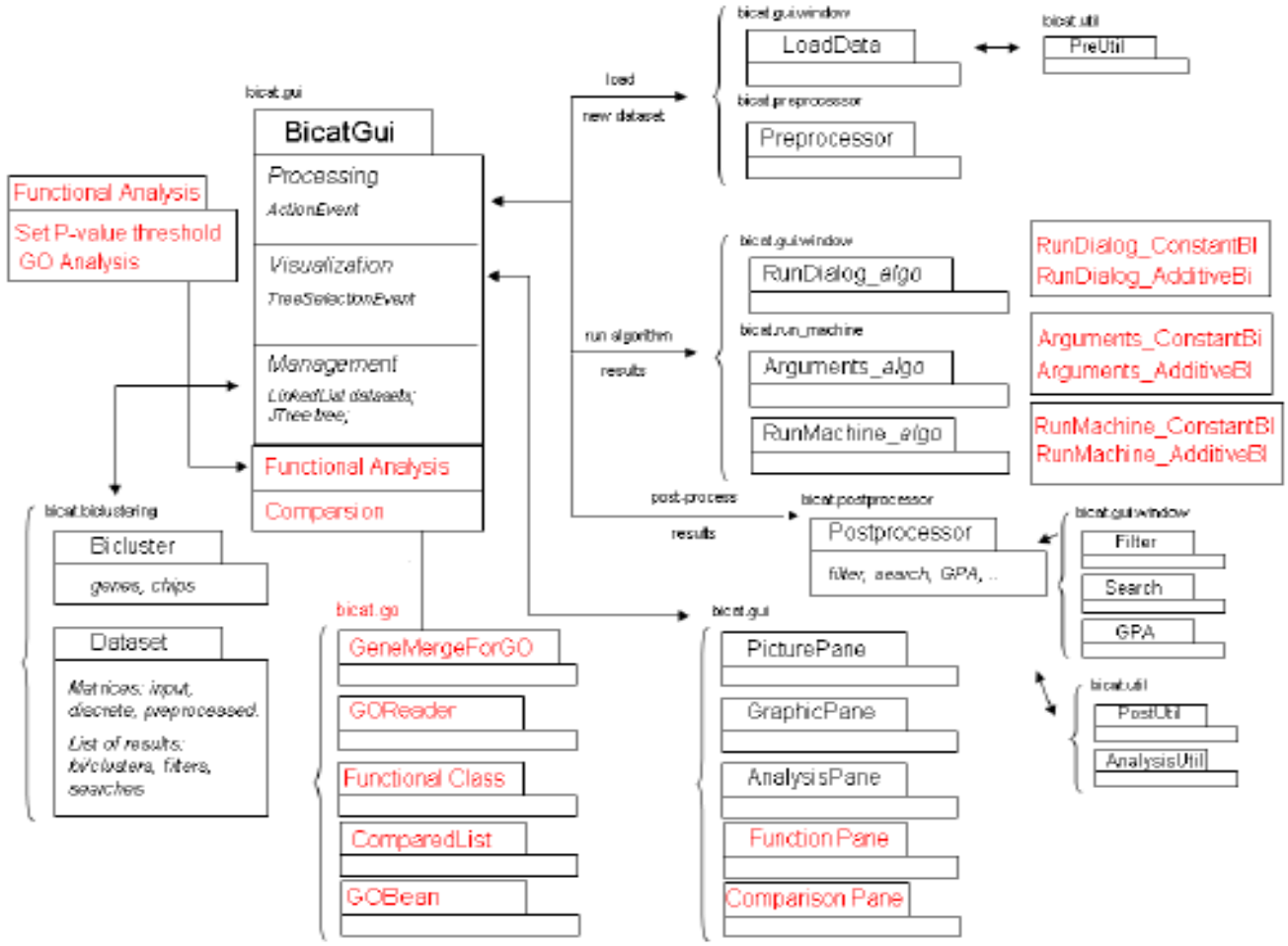

Figure 2. The general design of the BicAT-plus. Red color indicates the comparative tool packages and classes. The black entities are the original packages and interfaces of the BicAT program (modified from Barkow et al., 2006).

using the three sub-ontologies or categories of gene ontology (GO) (biological process, molecular function and cellular component) and visualizing the enriched GO terms per each bicluster in a separate histogram.

Evaluating the quality of each biclustering 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.

Comparing between the different biclustering algorithms according to the percentage of the functionally enriched biclusters at the required significance levels, the selected GO category and with certain filtration criteria for the GO terms.

Evaluating and comparing the results of external biclustering algorithms (not included in the BicAT-plus current version). This gives the BicAT-plus the advantage to be a generic tool that does not 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.

Displaying the analysis and comparison results using graphical and statistical charts visualizations in multiple modes (2D and 3D).

MATERIALS AND METHODS

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 as shown in Figure 2. The highlighted blocks with red color 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:

Lack of documentation of the BicAT tool, which influenced the planned time to understand the source code and extend it.

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.

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.

One of the objectives of this research was to enrich the BicAT (written using java) with more biclustering algorithms. However, some of these algorithms were written using C and C++. Thus, to solve such a compatibility problem, we converted the C files to dynamic link library (DLL) file, and 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.

GO overrepresentation programs

Many programs like BINGO (Maere et al., 2005), FUNCAT (Ruepp et al., 2004), GeneMerge (Castillo-Davis and Hartl, 2003) and FuncAssociate (Berriz et al., 2003) were used to investigate whether the set of genes discovered by biclustering/clustering methods present significant enrichment with respect to a specific GO annotation provided by Gene Ontology Consortium (Ashburner et al., 2000). BicAT-Plus uses 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 described by Steven et al. (2005): "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 hypergeometric test in which sampling occurs without replacement answers this question in the form of P-value. It counters replacement, the binomial test, which provides only an approximate P-value, but requires less calculation time."

Comparative methodologies based on GO

BicAT-plus provides reasonable method for comparing the results of different biclustering algorithms by:

(1) Identifying the percentage of enriched or overrepresented biclusters: This percentage is calculated for each algorithm with one or more GO terms per multiple significance levels (p-values) for each algorithm using the equation:

$$\%c_{\text{enriched}} = \frac{c_{\text{enriched}}}{n_c} * 100$$

Where, $\%c_{enriched}$ is the percentage of enriched bicluster per significance level; $c_{enriched}$ is the number of enriched biclusters at this significance level and n_c is the total number of biclusters. The definition of significance depends on the user selection of threshold p-values. A bicluster 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 (Prelic et al., 2006; Liu and Wang, 2007). 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 biclusters per all significance levels is considered to be the optimum one as it does group effectively the genes sharing similar functions in the same bicluster.

(2) Estimating algorithms' predictability power to recover user 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 (Gasch et al., 2000). BicAT-plus make the user to compare the predictability power of biclusters algorithms to interested pattern defined by the user (Table 2 for an example).

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:

Download BicAT-plus from our site (http://home.k-space.org/FADL/Downloads/BicAT_plus.zip).

Load gene expression data to BicAT-plus, then run the selected five prominent biclustering methods with setting parameters as shown in Table 1.

Run GO comparison tool in the BicAT-plus and add the available biclustering algorithms to the compared list as shown in Figure 4.

Select one of the available GO category example biological Process molecular function and cellular components.

Select the P-values example 0.00001, 0.0001, 0.01, 0.005 and 0.05.

Press compare button.

Press comparison menu, functional enrichment and select 2D or 3D charts (Figure 5).

RESULTS AND DISCUSSION

The above comparison steps are performed on the gene expression data of S*accharomyces cerevisiae* provided by Gasch et al. (2000). 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 Stanford University website (http://genome-www.stanford.edu\/yeast_stress). For each biclustering algorithm, we used the default parameters as recommended by some authors in their corresponding publications (Table 1).

The percentage of enriched function

After applying the above steps on Gasch data (Gasch et al., 2000), BicAT-plus produced the histogram shown in Figure 6. By comparing Figure 6 and Figure 3 and Figure 7 in the previous work (Prelic et al., 2006; Liu and Wang, 2007), respectively we found that the percentages of enriched biclusters for the matched algorithms are almost the same. This validated the results of the proposed comparative tool. Investigating both figures, we observed that OPSM algorithm gave a high portion of functionally

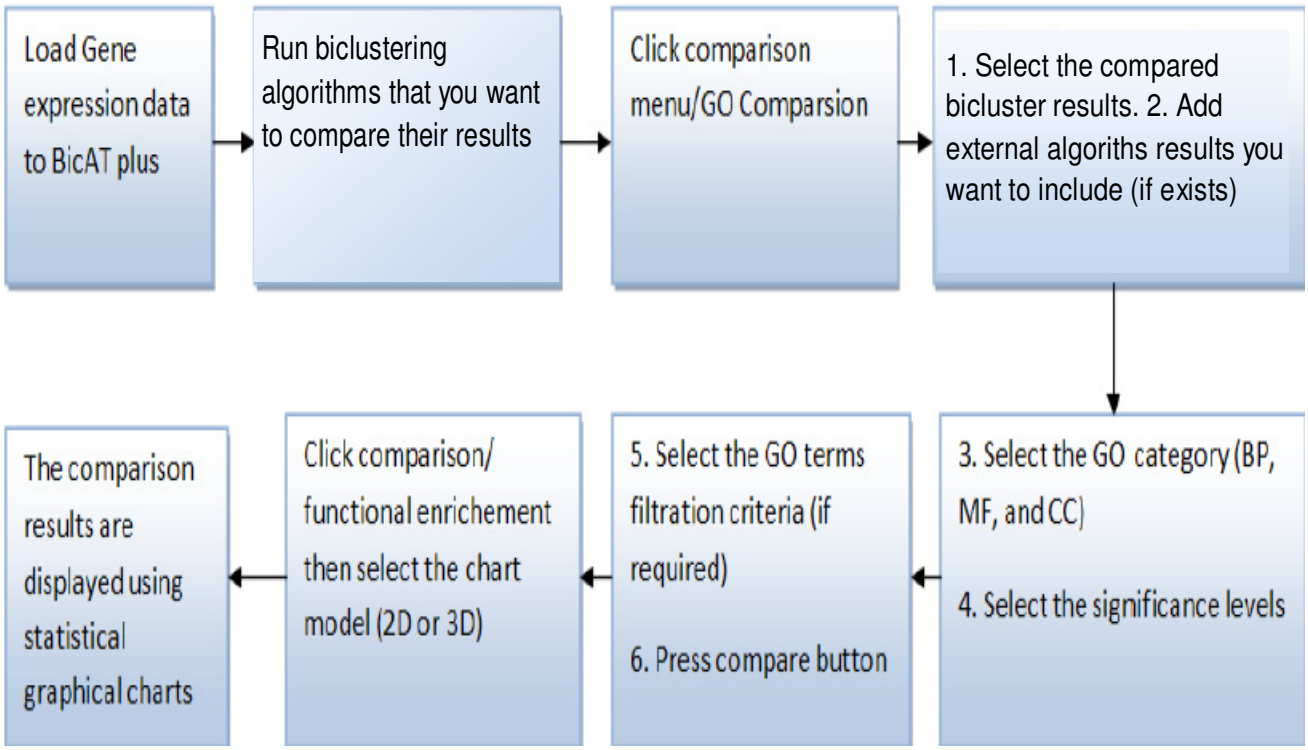

Figure 3. BicAT-Plus Comparison process steps.

Table 1. Default parameter settings of the compares biclustering method.

Biclustering algorithm	Parameter settings
ISA	t_g = 2.0, t_c = 2.0, seeds = 500
CC	δ = 0.5, α = 1.2, M = 100
OPSM	l = 100
BiVisu	E = 0.82, N_r = 10, N_c = 5, P_o = 25
K-means	K=100

The definitions of these parameters are listed in their original publications (Ben-Dor et al., 2003; Cheng and Church, 2000; Ihmels et al., 2002, 2004; Cheng et al., 2007).

enriched biclusters at all significance levels (from 85 to 100%). Next to OPSM, ISA 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 used the filtration criteria developed in the comparative tool by neglecting those biclusters that 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 as described in the equation:

$$\% \text{ study fraction} = \frac{n_{g \in GO}}{n_g} * 100$$

Where, *% study fraction* is the percentage of study fraction of a GO term; $n_{g \in GO}$ is the number of genes sharing the GO term in a bicluster and n_g is the total number of genes in this bicluster.

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 contain a lower number of genes per each category. This filtration will help in identifying the powerful and most reliable algorithms that are able to group maximum numbers of genes sharing same functions in one bicluster.

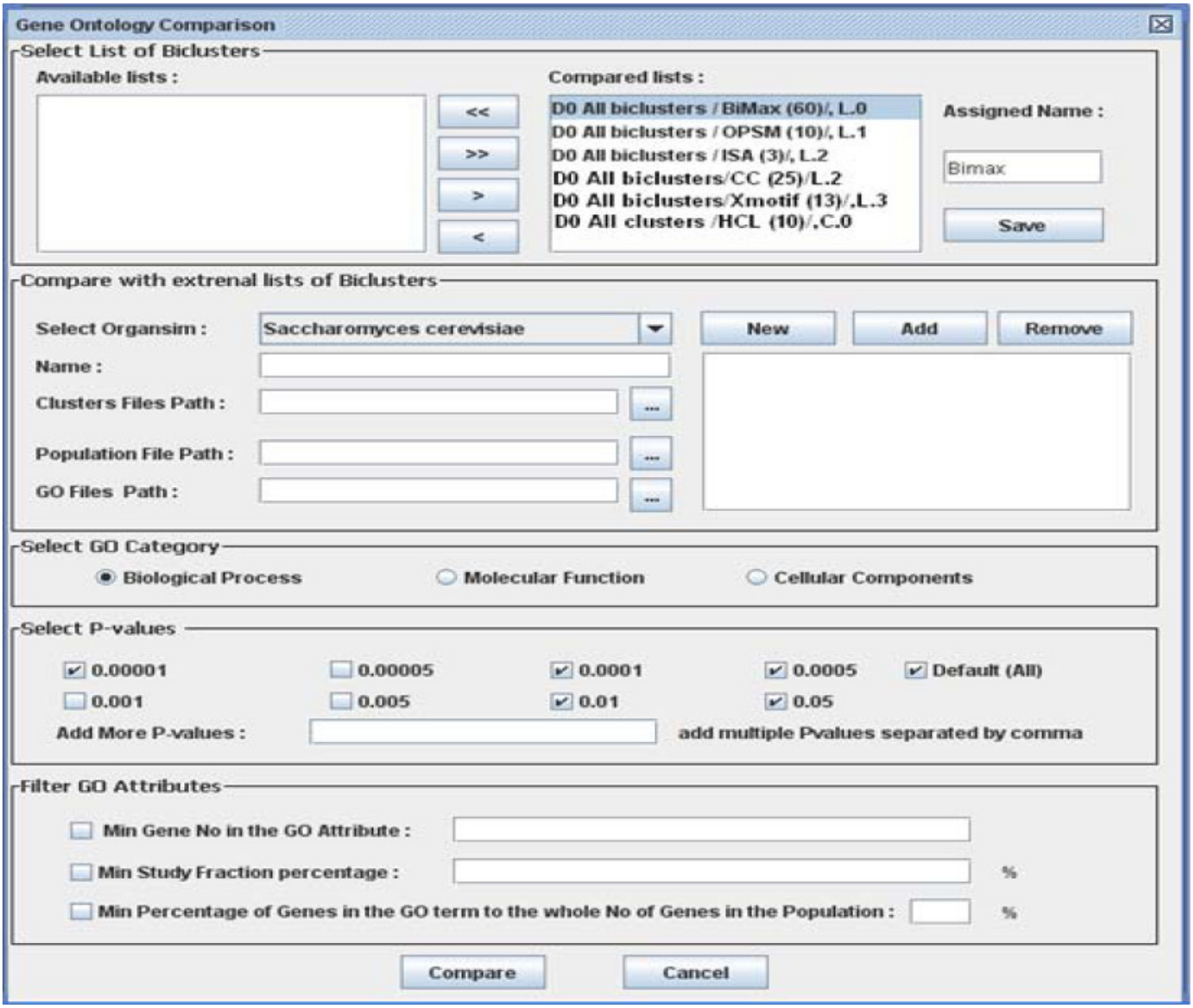

Figure 4. BicAT-Plus comparison dialogue.

The predictability power to recover interested pattern

The user could compare biclusters algorithms based on which of them could recover defined pattern; like which one of them could recover biclusters which have response to the conditions applied in Gasch et al.'s (2000) experiments. In Table 2, the difference between the biclusters/clusters contents were summarized.

Although OPSM showed high percentage level of enriched biclusters (as shown in Figures 6 and 7), its biclusters do not contain any genes within any GO category response to Gasch's experiments. The k-means and Bivisu cluster/bicluster results distinguished a unique GO category, which was GO: 0000304 (response to singlet oxygen) and GO: 0042542 (response to hydrogen peroxide). The powerful usage of these bicluster algorithms is significantly resulted 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.

Conclusion

We have introduced the BicAT-plus with reasonable comparative methodology based on the gene ontology.

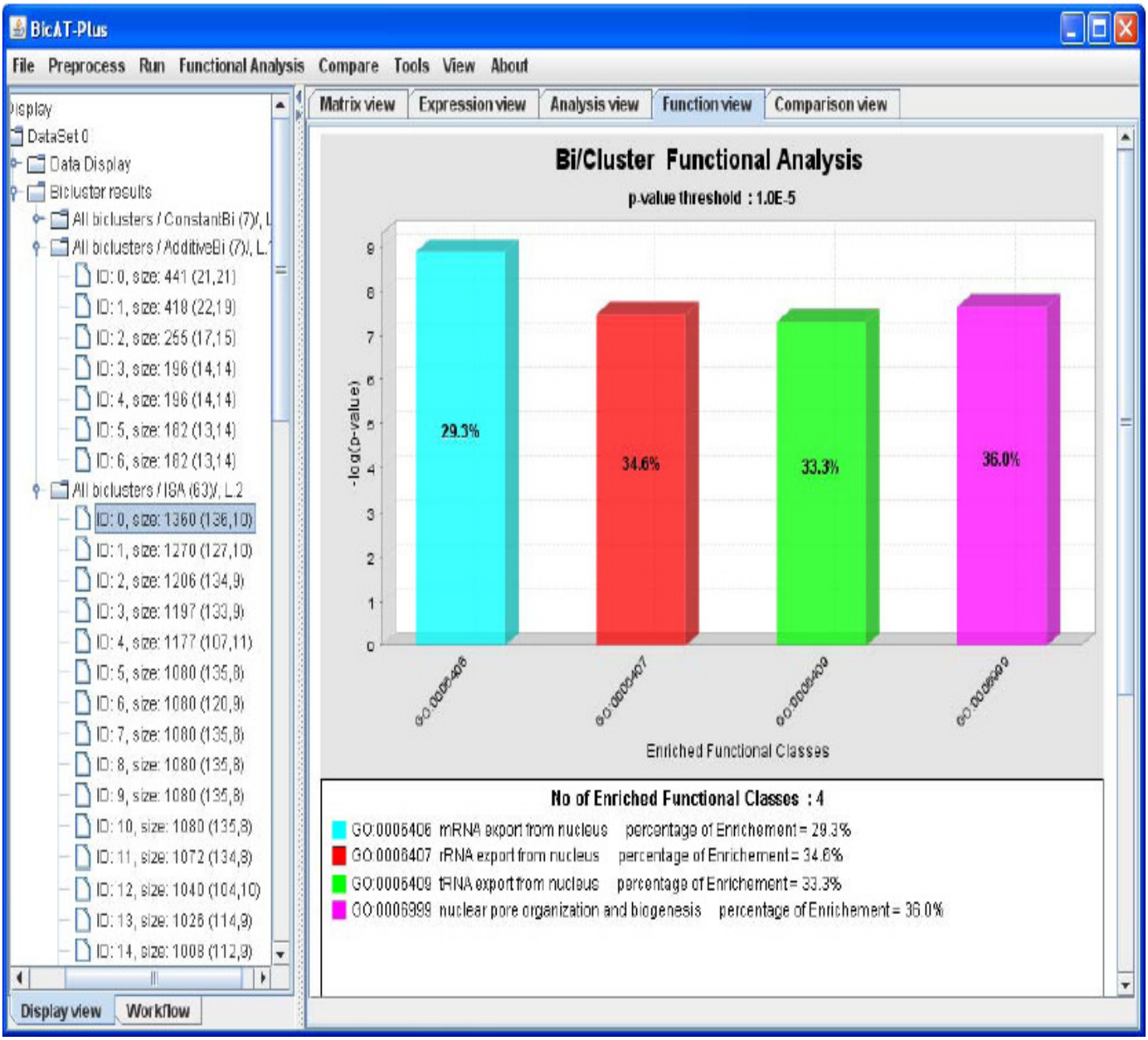

Figure 5. Functional analysis results of the selected bicluster. Each column represents an enriched GO functional class, and the height of the column is proportional to the significance of this enrichment (height = -log (p-value).

To the best of our knowledge, such an automatic comparison tool of the various biclustering algorithms has not been available in 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 could help biologists to extract the best results of each algorithm, and also interpret these results to useful biological meaning. In other words, the algorithms that showed good quality of results (per the dataset) can be used to provide a simple means of gaining information to the functions of many genes whose information are not available currently (unannotated genes). Using BicAT-plus, we can identify the highly enriched biclusters 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

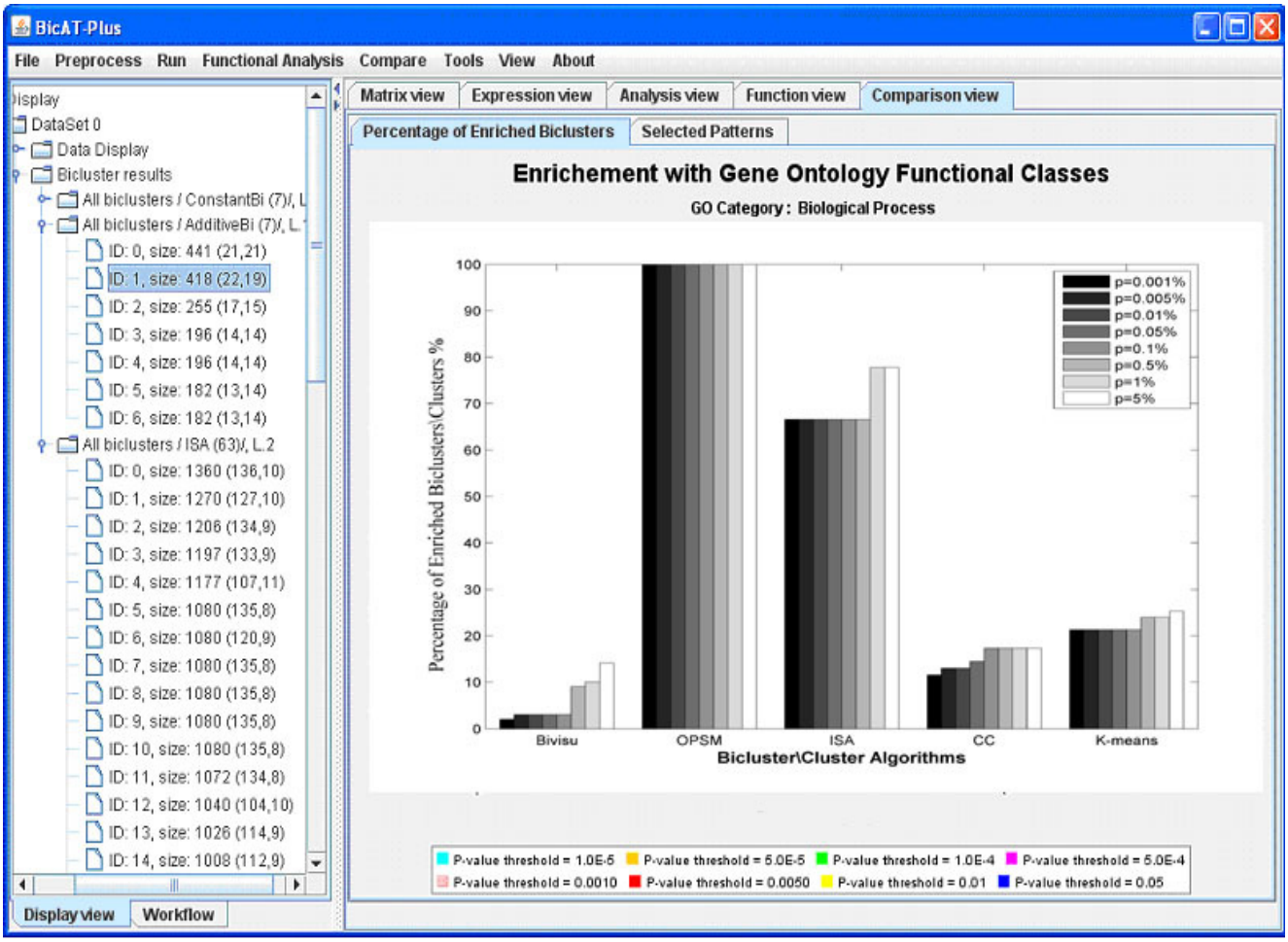

Figure 7. Percentage of significantly enriched biclusters by GO biological process category, by setting the allowed minimum number of genes per each GO category to 10 and the study fraction to more than 50%.

Table 2. Gene ontology category per number of annotated genes of the bicluster/cluster algorithm results for the experiment al condition on Gash et al.'s (2000) experiments.

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

Table 2. Continue

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

points (conditions or samples) with respect to the large number of genes in the microarray dataset.

However, there are several aspects of this research that are worth further investigation. According to the studies carried out so far, new ideas for consideration are introduced as follows:

To enrich the BicAT-plus with more comparative methodologies beside GO, for example, KEGG and promoter analysis, by identifying the transcription factors for the clustered genes.

To extend the BicAT-plus to provide users with multiple export options for the interested enriched biclusters.

To embed the BicAT-plus as a plug-in in the cytoscape platform, which is an open source bioinformatics software for visualizing molecular interaction networks and biological pathways, and to also integrate these networks with annotations, gene expression profiles and other state data? Thus, very promising challenge is to get use of the highly enriched biclusters identified by the BicAT-plus in solving these integrated networks in the cytoscape.

REFERENCES

Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," Nat Genet. 25: 25-29.

Azuaje F (2002), "A cluster validity framework for genome expression data," February 1, Bioinformatics, 18: 319-320.

Barkow S, Bleuler S, Prelic A, Zimmermann P, Zitzler E (2006). "BicAT: a biclustering analysis toolbox," May 15, Bioinformatics, 22: 1282-1283.

Ben-Dor A, Chor B, Karp R, Yakhini Z (2003). "Discovering local structure in gene expression data: the order-preserving submatrix problem," J. Comput. Biol. 10: 373-384.

Berriz GF, King OD, Bryant B, Sander C, Roth FP (2003). "Characterizing gene sets with FuncAssociate, December 12. " Bioinformatics, 19: 2502-2504.

Castillo-Davis CI, Hartl DL (2003). "GeneMerge - post-genomic analysis, data mining, and hypothesis testing," Bioinformatics, 19: 891-892.

Cheng KO, Law NF, Siu WC, Lau TH (2007). "BiVisu: software tool for bicluster detection and visualization," Bioinformatics, 23: 2342 - 2344.

Cheng Y, Church GM (2000). "Biclustering of expression data," Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology, pp. 93-103.

Datta S, Datta S (2003). "Comparisons and validation of statistical clustering techniques for microarray gene expression data," March 1, Bioinformatics, 19: 459-466.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000). "Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes, December 1" .Mol. Biol. Cell, 11: 4241-4257.

Ihmels J, Bergmann S, Barkai N (2004). "Defining transcription modules using large-scale gene expression data," Bioinformatics, 20: 1993-2003.

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

Liu X, Wang L (2007). "Computing the maximum similarity bi-clusters of gene expression data," January 1. Bioinformatics, 23: 50-56.

Madeira SC, Oliveira AL (2004). "Biclustering algorithms for biological data analysis: a survey," IEEE/ACM Trans. Comput. Biol. Bioinform. 1: 24-45.

Maere S, Heymans K, Kuiper M (2005). "BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks," August 15, Bioinformatics, 21: 3448-3449.

Murali TMKS (2003). "Extracting conserved gene expression motifs from gene expression data." in Pac. Symp. Biocomput. 8: 77-88.

Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E (2006). "A Systematic comparison and evaluation of biclustering methods for gene expression data," Bioinformatics, 22: 1122-1129.

Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M, Mewes HW (2004). "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes," October 14. Nucl. Acids Res. 32: 5539-5545.

Szeto L, Liew A, Yan H, Tang S (2003). "Gene Expression data clustering and visualization based on a binary hierarchical clustering framework," Special issue on Biomedical Visualization for Bioinformatics, J. Visual Languages and Comput. 14: 341-362.

Yeung KY, Haynor DR, Ruzzo WL (2001). "Validating clustering for gene expression data, April 1," Bioinformatics, 17: 309-318.

"http://genome-www.stanford.edu\/yeast_stress."