Chapter 3

Analysis of Gene Expression Data Using Biclustering Algorithms

Fadhl M. Al-Akwaa

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/48150

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 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 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 becomes 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 in to subsets that share biological significance.

1.1. What is Clustering?

A large number of clustering definitions can be found in the literature. The simplest definition is shared among all and includes one fundamental concept: the grouping together of similar data items into clusters[1].

Clustering is an important explorative statistical analysis of gene expression data. It aims to identify and group genes that exhibit similar expression patterns over several conditions and also group the conditions based on the expression profiles across set of genes. The successful clustering approach should guarantee two criteria which are homogeneity high similarity between elements in the same cluster, and separation – low similarity between elements from different clusters. When homogeneity and separation are precisely defined, those are two opposing objectives: The better the homogeneity the poorer the separation, and vice versa [2]. Several algorithmic techniques were previously used for clustering gene expression data, including hierarchical clustering [3], self organizing maps [4], and graph theoretic approaches [5].

1.1.1. K-means

K-means is a classical clustering algorithm [6] invented in 1956 to classify or to group objects (genes) based on attributes or features (experimental conditions) into K number of groups (clusters). K is positive integer number and assumed to be known.

K-means computational approach starts by placing K points into the space represented by the objects that are being clustered. These points represent initial group centroids. We can take any random objects as the initial centroids or the first K objects in sequence can also be used as the initial centroids. Then the K means algorithm will do the four steps below until convergence:

1. Determine the centroids coordinate.
2. Determine the distance of each object to the centroids using the Euclidean distance.
3. Group the objects based on minimum distance.
4. Iterate the above steps till no object moves its assigned group.

Each iteration of k-means modifies the current partition by checking all possible modifications of the solution, in which one element is moved to another cluster. This is done by reducing the sum of distances between objects and the centers of their clusters. This procedure is repeated until no further improvement is achieved (No object move the group) and all the objects are grouped into the final required number of clusters.

A disadvantage of K-means algorithm could be perceived in the need to specify the number of clusters K as a parameter value prior to running the algorithm. In cases where there is no expectation about K, user has to make trails with several values of K or use external techniques to guess the no of clusters may be exist.

1.1.2. Hierarchical clustering (HCL)

Hierarchical clustering does not partition the genes into subsets. Instead it builds a down-top hierarchy of clusters using agglomerative methods or top - down hierarchy of clusters using divisive methods. The traditional graphical representation of this hierarchy is called dendrogram tree. The divisive method begins at the root and starts to breaks up clusters whose having low similarity. Whereas, the Agglomerative method begins at the leaves of the tree and starts with an initial partition into single element clusters and successively merges clusters until all elements belong to the same cluster [3]. (See Figure 1) The agglomerative method is widely used than the divisive one which is not generally available, and rarely has been applied. The idea of the agglomerative method can be summarized as following: Given a set of N items (genes in our case) to be clustered, and an N*N distance (or similarity) matrix [7],

1. Assign each item to a cluster, so you have N clusters, each containing just one item.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

In Step 3, distance or similarity measurements between the merged clusters and all the other clusters can be calculated in one of three schemes: single-linkage, complete linkage and average-linkage.

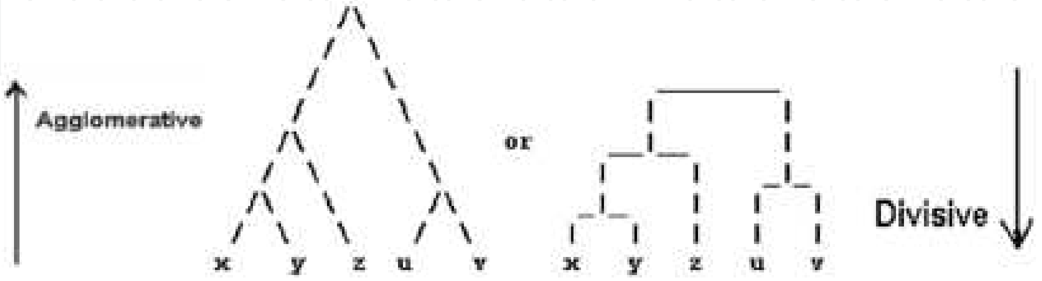


Figure 1. HCL: Agglomerative and Divisive Methods.

1.2. Biclustering

Traditional clustering approaches such as k-means and hierarchical clustering put each gene in exactly one cluster based on the assumption that all genes behave similarly in all conditions. However, recent understanding of cellular processes shows that it is possible for subset of genes to be co expressed under certain experimental conditions, and at the same time; to behave almost independently under other conditions. From this context, a new two mode clustering approach called biclustering or co-clustering has been introduced to group the genes and conditions in both dimensions simultaneously.

This allows finding subgroups of genes that show the same response under a subset of conditions, not all conditions. Also, genes may participate in more than one function, resulting in one regulation pattern in one context and a different pattern in another.

Example, if a cellular process is only active under specific conditions and there is a gene participates in multiple pathways that are differentially regulated, one would expect this gene to be included in more than one cluster; and this cannot be achieved by traditional clustering techniques.

Many biclustering methods exist in the literature [8]. Table 1 summarized some of promising biclustering algorithms developed during the last ten years. In brief, we described some of these algorithms according to their prediction strength, their promising results, to what they extend in the community, whether an implementation was available, and the feedback from their authors to explain some ambiguous issues.

1.2.1. Cheng and Church (CC)

CC algorithm[18] is considered to be the first real biclustering implementation after the primary idea has been introduced by Hartigan [19] in 1972.

Algorithm	Approach Time	Complicity	Prediction ability
Bivisu [9]	Exhaustive Bicluster Enumeration	$O(m^2nlogm)$[a]	Coherent values
MSBE [10]	Greedy Iterative Search	$O((n + m)^2)$	Coherent values
Bimax[11]	Divide-and-Conquer	$O(nm\beta log\beta)$	Coherent values
ROBA [12]	Matrix algebra	O(nmLN)	Coherent Evolution
x-motif [13]	Greedy Iterative Search	$nm^{O(log(1/\alpha)/log(1/\beta))}$	Coherent Evolution
SAMBA [14]	Exhaustive Bicluster Enumeration	O(n2)	Coherent Evolution
OPSM [15]	Greedy Iterative Search	$O(nm^3I)$	Coherent Evolution
Plaid[16]	Distribution Parameter Identification	XXX[b]	Coherent values
ISA [17]	Iterative Signature Algorithm	XXX	Coherent values
CC [18]	Greedy Iterative Search	O((n + m)nm)	Coherent values

[a] n and m are the row and column sizes of the expression matrix

[b] not available

Table 1. Biclustering Algorithms Comparison.

CC defines a bicluster as a subset of rows and a subset of columns with a high similarity. The proposed similarity score is called mean squared residue (H) and it is used to measure the coherence of the rows and columns in the single bicluster. Given the gene expression data matrix A = (X;Y); a bicluster is defined as a uniform submatrix (I;J) having a low mean squared residue score as following:

The CC Mean Squared Residue:

$$H(I,J) = \frac{1}{\|I\|\|J\|} \sum_{i \in I, j \in J} \left(a_{ij} - a_{iJ} - a_{Ij} + a_{IJ} \right)^2$$

Where: a_{ij} is gene expression level at row i and column j, a_{iJ} is the mean of row i, a_{Ij} is the mean of column j, a_{IJ} is the overall mean. CC algorithm will identify the submatrix as a bicluster if the score is below a level alpha which is a user input parameter to control the quality of the output biclusters. Generally; CC algorithm performs the following major steps:

1. Delete rows and columns with a score larger than alpha.
2. Adding rows or columns until alpha level is reached.
3. Iterate these steps until a maximum number of biclusters is reached or no bicluster is found [18].

1.2.2. Iterative Signature Algorithm (ISA)

The ISA algorithm [17, 20] is a novel method for the biclustering analysis of large-scale expression data. It is an efficient algorithm based on the iterative application of the signature algorithm presented in [17]. ISA considers a bicluster to be a transcription module which can be defined as a set of coexpressed genes together with the associated set of regulating conditions (Figure 2). Starting with an initial set of genes, all samples (conditions) are scored with respect to this gene set and those samples are chosen for which the score exceeds a certain threshold (usually defined by the user). In the same way, all genes are scored regarding the selected samples and a new set of genes is selected based on another user-defined threshold. The entire procedure is repeated until the set of genes and the set of samples converge and do not change anymore.

Multiple biclusters can be discovered by running the ISA algorithm on several initial gene sets. This approach requires identification of a reference gene set which needs to be carefully selected for good quality results. In the absence of pre-specified reference gene set, random set of genes is selected at the cost of results quality[17].

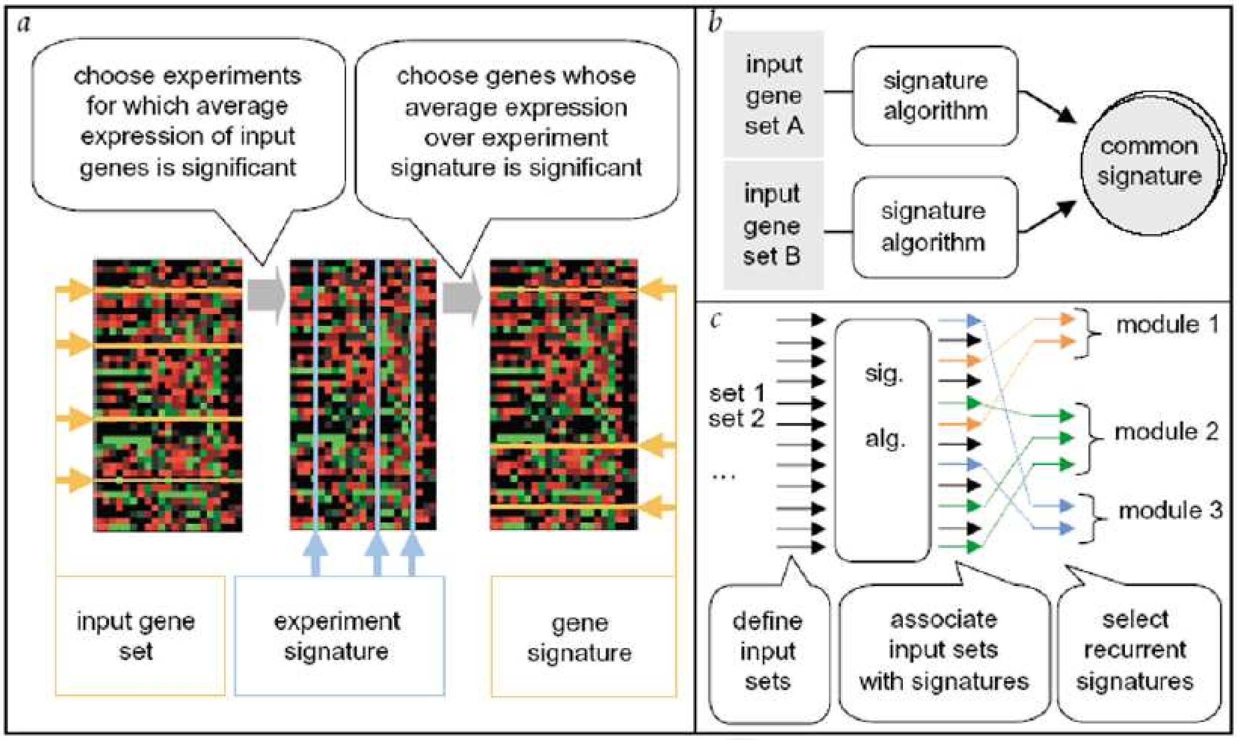

a
choose experiments for which average expression of input genes is significant
choose genes whose average expression over experiment signature is significant
input gene set
experiment signature
gene signature
b
input gene set A
input gene set B
signature algorithm
signature algorithm
common signature
c
set 1
set 2
...
sig.
alg.
module 1
module 2
module 3
define input sets
associate input sets with signatures
select recurrent signatures

Figure 2. The recurrence signature method. a, The signature algorithm. b, Recurrence as a reliability measure. The signature algorithm is applied to distinct input sets containing different subsets of the postulated transcription module. If the different input sets give rise to the same module, it is considered reliable. c, General application of the recurrent signature method. Copyright © [17].

1.2.3. Biclusters Inclusion Maximal (Bimax)

Bimax[11] is a simple binary model and new fast divide-and-conquer algorithm used to cluster the gene expression data. It is presented in 2006 by Computer Engineering and Networks Laboratory ETH Zurich, Switzerland. Bimax discretized the gene expression data matrix and convert it into a binary matrix by identifying a threshold, so transcription levels (genes expression values) above this threshold become ones and transcription levels below become zeros (or vice versa). Then, it searches for all possible biclusters that contain only ones. This can be done by iterating these steps:

1. Rearrange the rows and columns to concentrate ones in the upper right of the matrix.
2. Divide the matrix into two sub matrices.
3. Whenever in one of the submatrices only ones are found, this sub matrix is returned.

1.2.4. Order Preserving Submatrix(OPSM)

The order-preserving submatrix (OPSM) algorithm [15] is a probabilistic model introduced to discover a subset of genes identically ordered among a subset of conditions. It focuses on the coherence of the relative order of the conditions rather than the coherence of actual expression levels. In other words, the expression values of the genes within a bicluster induce an identical linear ordering across the selected conditions. Accordingly, the authors define a bicluster as a subset of rows whose values induce a linear order across a subset of the columns. The time complexity of this model is $O(nm^3I)$ where n andmare the number of rows and columns of the input gene expression matrix respectively and I is the number of biclusters. A disadvantage of OPSM algorithm is that it takes long time for high dimensional datasets. And this is because its time complexity is cubic with regards to the number of columns (dimensions) of the input matrix [15].

1.2.5. Maximum Similarity Bicluster(MSBE)

MSBE Biclustering algorithm [10] is a novel polynomial time algorithm to find an optimal biclusters with the maximum similarity. The idea behind this algorithm is to find subset of genes that are related to a reference gene. The reference gene is known in advance. MSBE algorithm uses the similarity score for a sub-matrix to find the similar expressions in the microarray datasets. And the threshold of the average similarity score is a user input parameter in order to allow the user to control the quality of the biclustering results.

1.3. Clustering or biclustering

Clustering algorithms [21-23] have been used to analyze gene expression data, on the basis that genes showing similar expression patterns can be assumed to be co-regulated or part of the same regulatory pathway. Unfortunately, this is not always true. Two limitations obstruct the use of clustering algorithms with microarray data. First, all conditions are given equal weights in the computation of gene similarity; in fact, most conditions do not contribute information but instead increase the amount of background noise. Second, each gene is assigned to a single cluster, whereas in fact genes may participate in several functions and should thus be included in several clusters[24].

A new modified clustering approach to uncovering processes that are active over some but not all samples has emerged, which is called biclustering. A bicluster is defined as a subset of genes that exhibit compatible expression patterns over a subset of conditions [11].

During the last ten years, many biclustering algorithms have been proposed (see [8] for a survey), but the important questions are: which algorithm is better? And do some algorithms have advantages over others?

Recently Kevin *et al.*[25]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, approaches, time complicity, number of parameters and prediction ability. In addition, 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[21-23], for biclustering; no such extensive comparison exist even after initial trails have been taken [11]. In the end, Biological merit is the main criterion for evaluation and comparison between the various biclustering methods.

In this chapter we attempt to develope a comparative tool (Bicat-Plus) which is showen in Figure 3 that includes the biological comparative methodology and to be as an extension to the BicAT program[26].

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. In addition, 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.

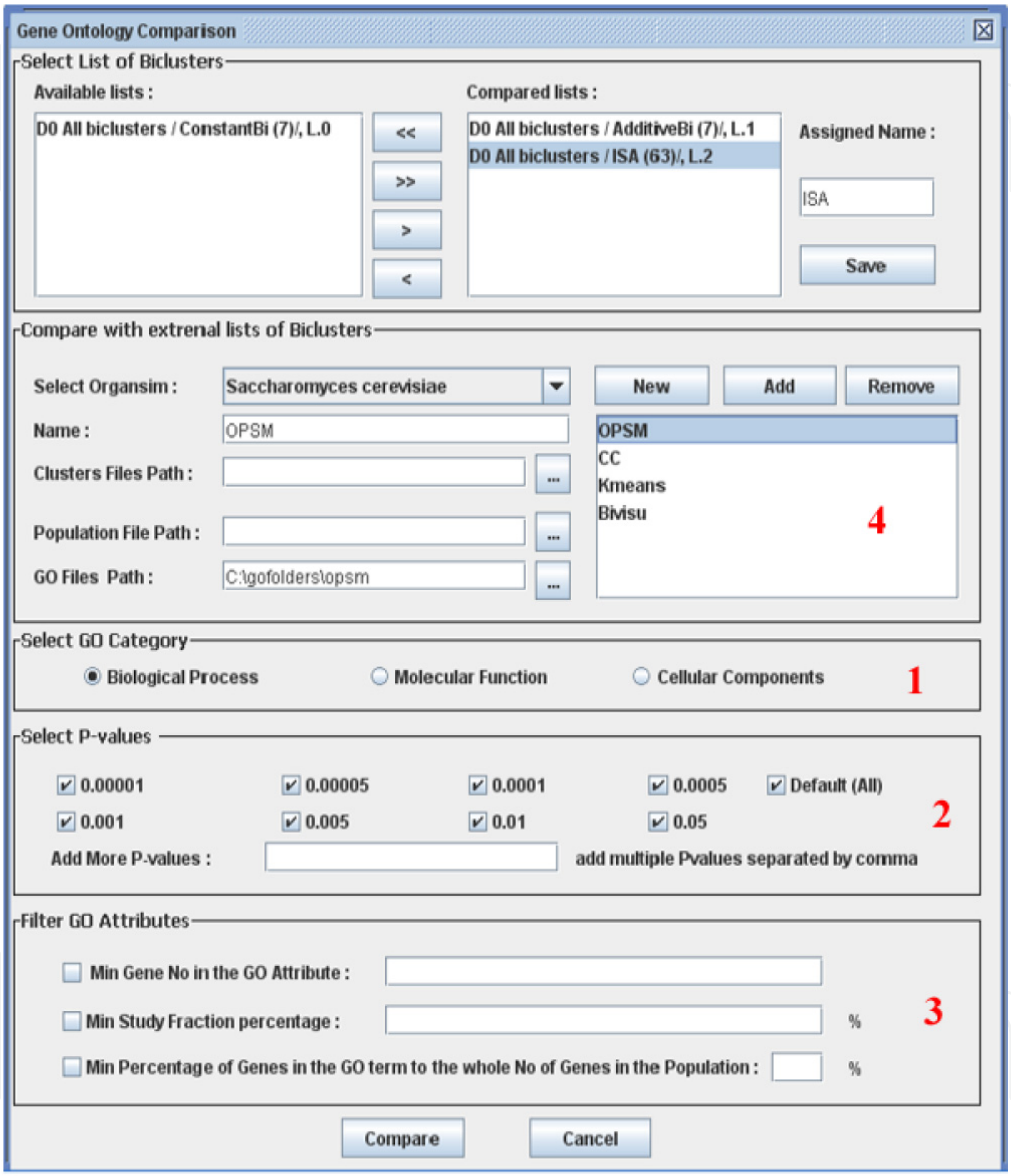
BicAT-Plus has many features, which could be summarized in the following:-

Algorithms required to be compared could be selected from the biclustering list (left list) to the compared list (right list). External biclustering results for other algorithms could be included in the comparison process. In addition, the organism model, selectable significance level, and GO category should be selected. Finally, Comparison criteria have to be selected based on the user biological metric.

1. User could perform biclusters functional analysis using the three Gene Ontology (GO) categories (biological process, molecular function and cellular component) (Figure3 with label number 1).
2. User could evaluate the quality of each biclustering algorithm results after applying the GO functional analysis and display the percentage of the enriched biclusters at different P-values (Figure3 with label number 2).
3. User could compare 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. (Figure3 with label number 3).
4. User could evaluate and compare the results of external biclustering algorithms. 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. (Figure 3 with label number 4).

5. User could display the results using graphical and statistical charts visualizations in multiple modes (2D and 3D).



Gene Ontology Comparison

Select List of Biclusters

Available lists :
D0 All biclusters / ConstantBi (7)/, L.0

Compared lists :
D0 All biclusters / AdditiveBi (7)/, L.1
D0 All biclusters / ISA (63)/, L.2

Assigned Name : ISA

Compare with extrenal lists of Biclusters

Select Organsim : Saccharomyces cerevisiae

Name : OPSM

Clusters Files Path :

Population File Path :

GO Files Path : C:\gofolders\opsm

OPSM
CC
Kmeans
Bivisu

Select GO Category

● Biological Process ○ Molecular Function ○ Cellular Components

Select P-values

☑ 0.00001 ☑ 0.00005 ☑ 0.0001 ☑ 0.0005 ☑ Default (All)

☑ 0.001 ☑ 0.005 ☑ 0.01 ☑ 0.05

Add More P-values : add multiple Pvalues separated by comma

Filter GO Attributes

□ Min Gene No in the GO Attribute :

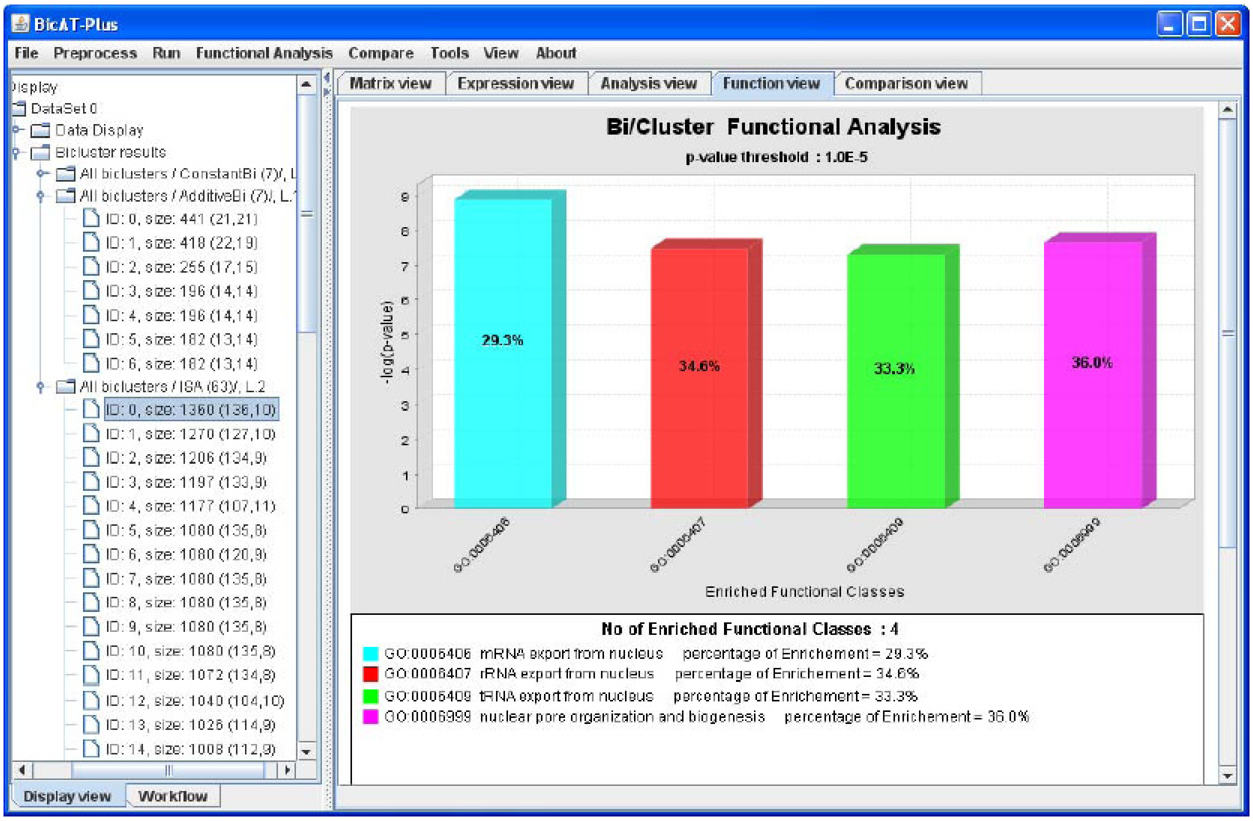
□ Min Study Fraction percentage : %

□ Min Percentage of Genes in the GO term to the whole No of Genes in the Population : %

Figure 3. BicAT-Plus Comparison Panel.

2. 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
File Preprocess Run Functional Analysis Compare Tools View About
Matrix view Expression view Analysis view Function view Comparison view
Bi/Cluster Functional Analysis
p-value threshold : 1.0E-5
-log(p-value)
0 1 2 3 4 5 6 7 8 9
29.3%
34.6%
33.3%
36.0%
GO:0006406
GO:0006407
GO:0006409
GO:0006999
Enriched Functional Classes
Display view Workflow

Display
DataSet 0
Data Display
Bicluster results
All biclusters / ConstantBi (7)/, L
All biclusters / AdditiveBi (7)/, L.1
ID: 0, size: 441 (21,21)
ID: 1, size: 418 (22,19)
ID: 2, size: 255 (17,15)
ID: 3, size: 196 (14,14)
ID: 4, size: 196 (14,14)
ID: 5, size: 182 (13,14)
ID: 6, size: 182 (13,14)
All biclusters / ISA (63)/, L.2
ID: 0, size: 1360 (136,10)
ID: 1, size: 1270 (127,10)
ID: 2, size: 1206 (134,9)
ID: 3, size: 1197 (133,9)
ID: 4, size: 1177 (107,11)
ID: 5, size: 1080 (135,8)
ID: 6, size: 1080 (120,9)
ID: 7, size: 1080 (135,8)
ID: 8, size: 1080 (135,8)
ID: 9, size: 1080 (135,8)
ID: 10, size: 1080 (135,8)
ID: 11, size: 1072 (134,8)
ID: 12, size: 1040 (104,10)
ID: 13, size: 1026 (114,9)
ID: 14, size: 1008 (112,9)

No of Enriched Functional Classes : 4

- 🟦 GO:0006406 mRNA export from nucleus percentage of Enrichement = 29.3%
- 🟥 GO:0006407 rRNA export from nucleus percentage of Enrichement = 34.6%
- 🟩 GO:0006409 tRNA export from nucleus percentage of Enrichement = 33.3%
- 🟪 GO:0006999 nuclear pore organization and biogenesis percentage of Enrichement = 36.0%

Figure 4. 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 (i.e. height = -log (p-value).

2.1. GO overrepresentation programs

Many programs like: BINGO[27], FUNCAT[28], GeneMerge[29] and FuncAssociate[30] were used to investigate whether the set of genes discovered by biclustering methods present significant enrichment with respect to a specific GO annotation provided by Gene Ontology Consortium [31]. 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 described by Steven *et al.*[27] "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's counterpart with replacement, the binomial test, which provides only an approximate P-value, but requires less calculation time."

2.2. Comparative methodologies

BicAT-Plus provides reasonable methods for comparing the results of different biclustering algorithms by:

- Identifying the percentage of enriched or overrepresented biclusters with one or more GO term per multiple significance level for each algorithm. 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. The results are displayed using a histogram for all the algorithms compared at the different preset significance levels, and the algorithm that gives the highest proportion of enriched biclusters for all significance levels is considered the optimum because it effectively groups the genes sharing similar functions in the same bicluster.
- Identifying the percentage of annotated genes per each enriched bicluster.
- Estimating the predictive power of algorithms to recover interesting patterns. Genes whose transcription is responsive to a variety of stresses have been implicated in a general Yeast response to stress (awkward). Other gene expression responses appear to be specific to particular environmental conditions. BicAT-Plus compares biclustering methods on the basis of their capacity to recover known patterns in experimental data sets. For example, Gasch et al.[32] measure changes in transcript levels over time responding to a panel of environmental changes, so it was expected to find biclusters enriched with one of response to stress (GO:0006950), Gene Ontology categories such as response to heat (GO:0009408), response to cold (GO:0009409) and response to glucose starvation(GO:0042149). The details of this comparison strategy are described in the results and in Table 3.

2.3. Comparison Process Steps

The following process diagram shown in Fig 5 summarizes the required steps by the user to compare between the different algorithms using the BicAT-plus:

1. Download BicAT-Plus from (*www.bioinformatics.org/**bicat-plus/***).
2. Load Gene Expression Data to BicAT-Plus then run the selected five prominent biclustering methods with setting parameters as shown in Table 2.
3. Run GO comparison tool in the BicAT-Plus and add the available biclustering algorithms to the compared list as shown in Fig 1.
4. Select 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.

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

Table 2. Default Parameter settings of the compared bi/clustering methods. The definitions of these parameters are listed in their original publications [9, 15, 17-18, 20] respectively.

3. Results & discussion

The above comparison steps is performed on the gene expression data of *S. cerevisiae* provided by Gasch [32]. 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 [33]. For each biclustering algorithm, we used the default parameters as authors recommend in their corresponding publications. See Table 2.

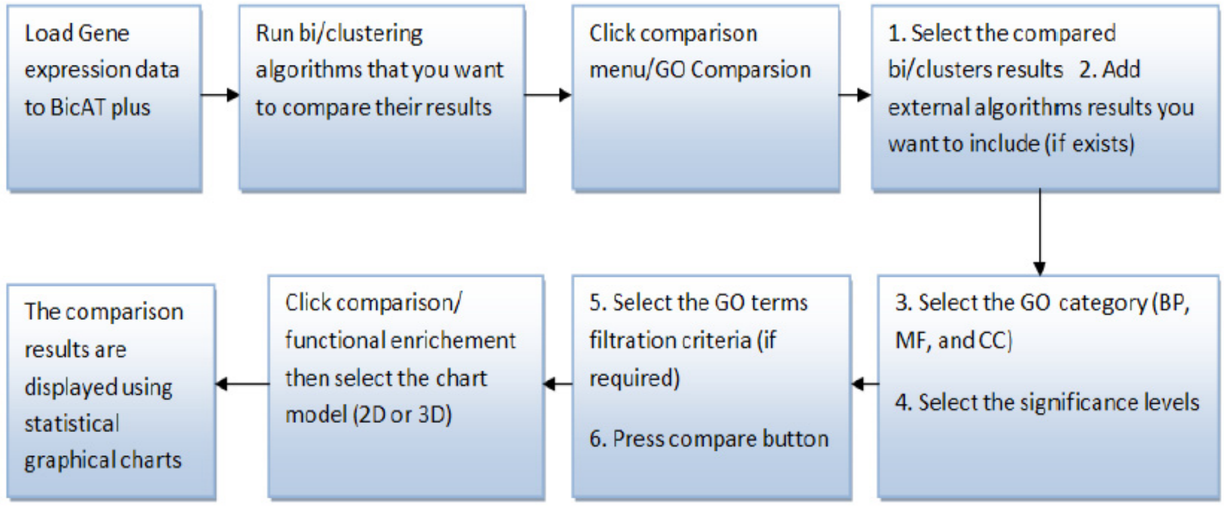

Load Gene expression data to BicAT plus
Run bi/clustering algorithms that you want to compare their results
Click comparison menu/GO Comparsion
1. Select the compared bi/clusters results 2. Add external algorithms results you want to include (if exists)
3. Select the GO category (BP, MF, and CC)
4. Select the significance levels
5. Select the GO terms filtration criteria (if required)
6. Press compare button
Click comparison/ functional enrichement then select the chart model (2D or 3D)
The comparison results are displayed using statistical graphical charts

Figure 5. BicAT-Plus Comparison process steps

3.1. The percentage of enriched function

After applying the above steps on Gasch data[32] , BicAT-plus produce the histogram shown in Fig 6. Investigating this figure, 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 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.

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

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 which are able to group maximum numbers of genes sharing same functions in one bicluster.

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 bi/clusters which have response to the conditions applied in Gasch experiments. In Table 2, the difference between the biclusters/clusters contents were summarized.

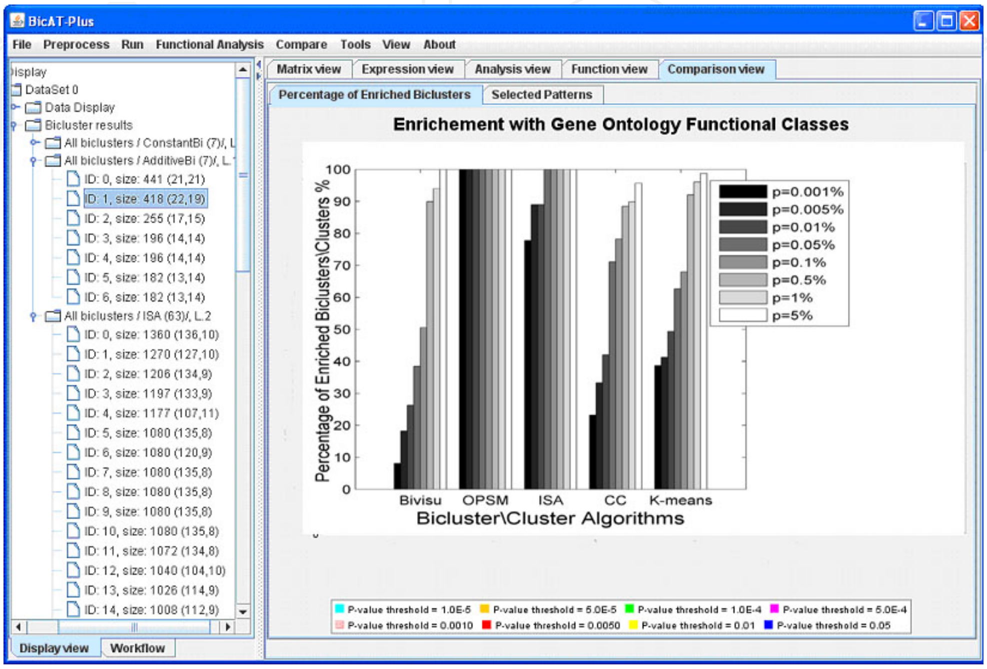

BicAT-Plus
File Preprocess Run Functional Analysis Compare Tools View About
Display
DataSet 0
Data Display
Bicluster results
All biclusters / ConstantBi (7)/, L
All biclusters / AdditiveBi (7)/, L
ID: 0, size: 441 (21,21)
ID: 1, size: 418 (22,19)
ID: 2, size: 255 (17,15)
ID: 3, size: 196 (14,14)
ID: 4, size: 196 (14,14)
ID: 5, size: 182 (13,14)
ID: 6, size: 182 (13,14)
All biclusters / ISA (63)/, L 2
ID: 0, size: 1360 (136,10)
ID: 1, size: 1270 (127,10)
ID: 2, size: 1206 (134,9)
ID: 3, size: 1197 (133,9)
ID: 4, size: 1177 (107,11)
ID: 5, size: 1080 (135,8)
ID: 6, size: 1080 (120,9)
ID: 7, size: 1080 (135,8)
ID: 8, size: 1080 (135,8)
ID: 9, size: 1080 (135,8)
ID: 10, size: 1080 (135,8)
ID: 11, size: 1072 (134,8)
ID: 12, size: 1040 (104,10)
ID: 13, size: 1026 (114,9)
ID: 14, size: 1008 (112,9)
Display view
Workflow
Matrix view
Expression view
Analysis view
Function view
Comparison view
Percentage of Enriched Biclusters
Selected Patterns
Enrichement with Gene Ontology Functional Classes
Percentage of Enriched Biclusters\Clusters %
0
10
20
30
40
50
60
70
80
90
100
Bivisu
OPSM
ISA
CC
K-means
Bicluster\Cluster Algorithms
p=0.001%
p=0.005%
p=0.01%
p=0.05%
p=0.1%
p=0.5%
p=1%
p=5%
P-value threshold = 1.0E-5
P-value threshold = 5.0E-5
P-value threshold = 1.0E-4
P-value threshold = 5.0E-4
P-value threshold = 0.0010
P-value threshold = 0.0050
P-value threshold = 0.01
P-value threshold = 0.05

Figure 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.

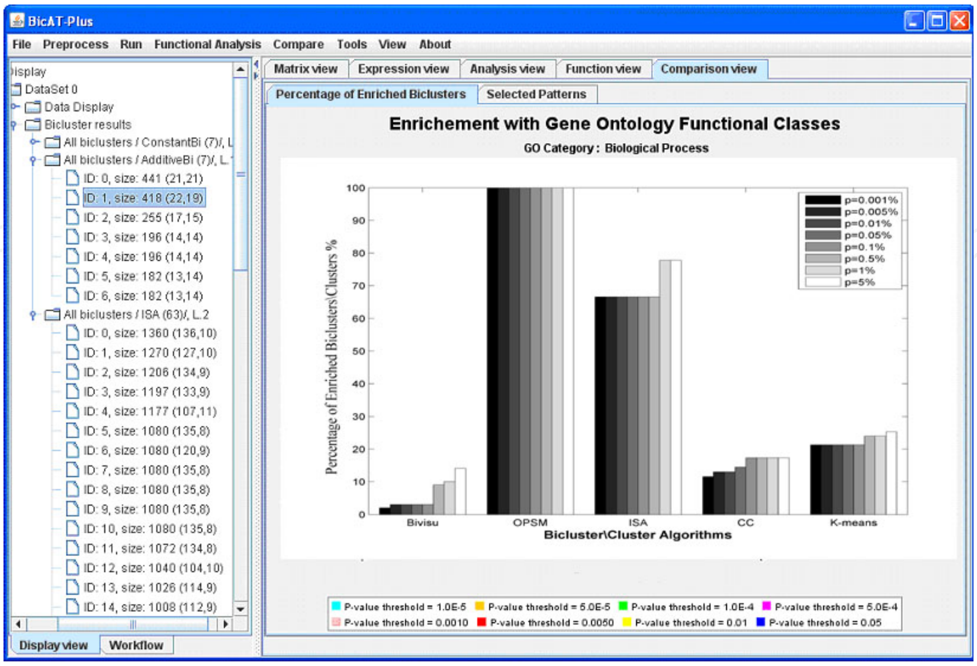

BicAT-Plus
File Preprocess Run Functional Analysis Compare Tools View About
Display
DataSet 0
Data Display
Bicluster results
All biclusters / ConstantBi (7)/, L
All biclusters / AdditiveBi (7)/, L
ID: 0, size: 441 (21,21)
ID: 1, size: 418 (22,19)
ID: 2, size: 255 (17,15)
ID: 3, size: 196 (14,14)
ID: 4, size: 196 (14,14)
ID: 5, size: 182 (13,14)
ID: 6, size: 182 (13,14)
All biclusters / ISA (63)/, L 2
ID: 0, size: 1360 (136,10)
ID: 1, size: 1270 (127,10)
ID: 2, size: 1206 (134,9)
ID: 3, size: 1197 (133,9)
ID: 4, size: 1177 (107,11)
ID: 5, size: 1080 (135,8)
ID: 6, size: 1080 (120,9)
ID: 7, size: 1080 (135,8)
ID: 8, size: 1080 (135,8)
ID: 9, size: 1080 (135,8)
ID: 10, size: 1080 (135,8)
ID: 11, size: 1072 (134,8)
ID: 12, size: 1040 (104,10)
ID: 13, size: 1026 (114,9)
ID: 14, size: 1008 (112,9)
Display view
Workflow
Matrix view
Expression view
Analysis view
Function view
Comparison view
Percentage of Enriched Biclusters
Selected Patterns
Enrichement with Gene Ontology Functional Classes
GO Category : Biological Process
Percentage of Enriched Biclusters\Clusters %
0
10
20
30
40
50
60
70
80
90
100
Bivisu
OPSM
ISA
CC
K-means
Bicluster\Cluster Algorithms
p=0.001%
p=0.005%
p=0.01%
p=0.05%
p=0.1%
p=0.5%
p=1%
p=5%
P-value threshold = 1.0E-5
P-value threshold = 5.0E-5
P-value threshold = 1.0E-4
P-value threshold = 5.0E-4
P-value threshold = 0.0010
P-value threshold = 0.0050
P-value threshold = 0.01
P-value threshold = 0.05

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 large than 50%.

Although OPSM show high percentage level of enriched biclusters (as shown in Fig 6, 7), 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 bicluster 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.

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

Table 3. Gene Ontology category per number of annotated genes of the Bicluster/cluster algorithm results for the experimental condition on Gasch Experiments[32].

4. Conclusion

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

Finally there are several aspects of this research that worth further investigation, according to the Studies carried out so far and also introducing new ideas for consideration

1. 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.
2. Extend the BicAT-Plus to provide users with multiple export options for the interested enriched biclusters.
3. Embed the BicAT-Plus as a plug-in in the Cytoscape platform[34] which is open source bioinformatics software for visualizing molecular interaction networks and biological pathways and integrating 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.

Author details

Fadhl M. Al-Akwaa
Biomedical Eng. Dept., Univ. of Science & Technology, Sana'a, Yemen

5. References

[1] Fung G: A Comprehensive Overview of Basic Clustering Algorithms. *Citeseer* 2001:1-37.

[2] Sharan R, Elkon R, Shamir R: Cluster analysis and its applications to gene expression data. *Ernst Schering Res Found Workshop* 2002:83-108.

[3] Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 1998, 95:14863 - 14868.

[4] P. Tamayo DS, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic dierentiation. In: *Proceedings of the National Academy of Sciences of the United States of America,: 1999*. 2907–2912.

[5] Sharan RSaR: Click: a clustering algorithm for gene Expression analysis. In: *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology: 2000*. 307–316.

[6] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: Systematic determination of genetic network architecture. *Nature Genetics* 1999, 22:281-285.

[7] Johnson S: Hierarchical clustering schemes. *Psychometrika* 1967, 32(3):241-254.

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

[9] Cheng KO, Law NF, Siu WC, Lau TH: BiVisu: software tool for bicluster detection and visualization. *Bioinformatics* 2007, 23(17):2342 - 2344.

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

[11] Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E: A Systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 2006, 22(9):1122 - 1129.

[12] A. Tchagang and A. Twefik: Robust biclustering algorithm (ROBA) for DNA microarray data analysis. In: *IEEE/SP 13thWorkshop on Statistical Signal Processing*. 2005: 984–989.

[13] Murali TM, S K: Extracting conserved gene expression motifs from gene expression data. In: *Pac Symp Biocomput*. 2003: 77–88.

[14] A. Tanay RS, M. Kupiec, and R. Shamir, : Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data,. In: *Proceedings of the National Academy of Sciences of the United States of America: 2004*. 2981–2986.

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

[16] H. Wang WW, J. Yang, and P. S. Yu, : Clustering by Pattern Similarity: the pCluster Algorithm. *SIGMOD* 2002.

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

[18] Cheng Y, Church GM: Biclustering of expression data. *Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology* 2000:93 - 103.

[19] Hartigan J: Direct Clustering of a data matrix. *Journal of the American Statistical Association* 1972, 67:123–129.

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

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

[22] Guthke R, Moller U, Hoffmann M, Thies F, Topfer S: Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics* 2005, 21(8):1626-1634.

[23] D'haeseleer P, Liang S, Somogyi R: Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 2000, 16(8):707-726.

[24] Reiss D, Baliga N, Bonneau R: Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* 2006, 7(1):280.

[25] Yip KYaQ, Peishen and Schultz, Martin and Cheung, David W and Cheung, Kei-Hoi: SemBiosphere: A Semantic Web Approach to Recommending Microarray Clustering Services. In: *The Pacific Symposium on Biocomputing*. 2006: 188-199.

[26] Barkow S, Bleuler S, Prelic A, Zimmermann P, Zitzler E: BicAT: a biclustering analysis toolbox. *Bioinformatics* 2006, 22(10):1282-1283.

[27] Maere S, Heymans K, Kuiper M: BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* 2005, 21(16):3448-3449.

[28] Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M *et al*: The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucl Acids Res* 2004, 32(18):5539-5545.

[29] Castillo-Davis CI, Hartl DL: GeneMerge - post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* 2003, 19(7):891 - 892.

[30] Berriz GF, King OD, Bryant B, Sander C, Roth FP: Characterizing gene sets with FuncAssociate. *Bioinformatics* 2003, 19(18):2502-2504.

[31] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25:25 - 29.

[32] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Mol Biol Cell* 2000, 11(12):4241-4257.

[33] http://genome-www.stanford.edu/yeast/_stress

[34] Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003, 13(11):2498-2504.