Bicluster Detection using Strength Pareto Front Evolutionary Algorithm

Maryam Golchin
School of information and communication technology
Griffith University
Southport, Australia
+61 7 5552 8800
maryam.golchin@griffithuniversity.edu.au

Alan Wee-Chung Liew
School of information and communication technology
Griffith University
Southport, Australia
+61 7 5552 8800
a.liew@griffith.edu.au

ABSTRACT

Biclustering has many applications in various fields such as pattern classification, information retrieval, data mining and functional annotation. Biclustering extracts accurate information from gene expression datasets by clustering rows and columns of a dataset simultaneously. In this paper, a new multi-objective evolutionary biclustering framework based on strength Pareto front evolutionary algorithm (SPEA2) is proposed. A heuristic search is added into SPEA2 to delete and add genes and conditions into randomly generated biclusters. In order to select the best bicluster among Pareto front solutions, k-mean algorithm is used. The new population is generated based on mutation and crossover. The performance of the proposed method is evaluated using synthetic and real datasets and compared with several well-known biclustering methods. The experimental results show better performance and significant enrichment of detected biclusters.

CCS Concepts

• Computing methodologies→Machine learning→Learning paradigms→Unsupervised learning→Cluster analysis.

Keywords

Biclustering; heuristic search, SPEA2; gene expression data.

1. INTRODUCTION

In a gene expression matrix, rows describe gene annotation and columns describe condition annotation as it is shown in Figure 1. Clustering algorithms use the gene expression level (e_{ij}) of gene expression matrices to group genes into biologically meaningful groups. These methods have been widely used in discovery of diseases [1] and gene expression profiling [2]. The genes behave similarly if they share the same pattern or structure. However, this similarity sometimes happens only under a subset of conditions. Another property of clustering algorithms is that clustering techniques assign a gene to one cluster only, either row-wise or column-wise [3].

These properties are disadvantages for this research so the biclustering technique was proposed [4]. The goal of biclustering is to discover sub-groups of genes and conditions simultaneously such that these groups show considerable homogeneity from microarray data [5]. Here, a gene can be clustered in several biclusters under differing subsets of conditions so biclusters can overlap and there is no need to use all the conditions at the same time. A survey of biclustering techniques can be found in [6].

In multi-objective optimization [7, 8], two or more conflicting objectives are optimized in a fitness function. A solution dominates any other solution when all the objective values are better or equal to other solution objective values and at least one objective value is better than the corresponding value (blue circles in Figure 2 are dominated results and red circles are non-dominated results). There are various trade-offs of non-dominated solutions in a multi-objective optimization problem, which constitute the Pareto front of the problem. Figure 2 shows a 2D Pareto front with respect to two objectives where increasing one objective causes the decrease of the other objective and vice-versa.

Cheng and Church (CC) [4] were the first to introduce the concept of biclustering to gene expression data analysis. They used a greedy search and a heuristic algorithm to locate δ-biclusters in a top-down manner. They iteratively removed and added rows and columns until the mean square residue (MSR) error remained below δ.

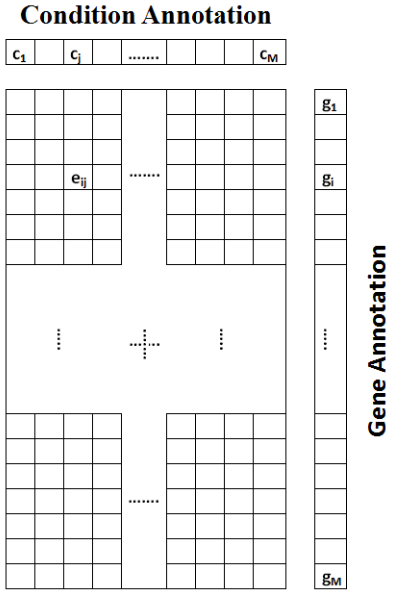

Figure 1. An example of gene expression data

ACE '16 Canberra, ACT Australia
Copyright 2016 ACM 978-1-4503-4042-7/16/02 ...$15.00.
http://dx.doi.org/10.1145/2843043.2843050

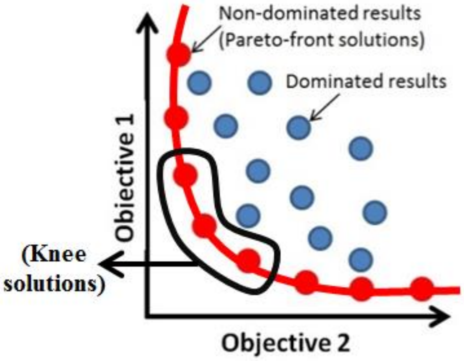

Figure 2. Red circles and line constitute Pareto front. Red circles are non-dominated results and blue circles are dominated results.

There are several multi-objective biclustering algorithms introduced in literature [9-12]. Liu et. al., [9] proposed a dynamic multi-objective immune optimization biclustering technique. They detected maximum biclusters with minimized MSR and maximized row variance. In [10], Coelho and França used artificial immune network (aiNet) to propose a multi-objective multi-population aiNet (MOM-aiNet) in parallel. Their method detected maximum biclusters while minimizing overlap in large datasets. A multi-objective evolutionary algorithm using non-dominated sorting genetic algorithm II (NSGAII) and a heuristic search was proposed in [11] to find maximum biclusters that minimized MSR and maximized row variance. They reported better results in comparison to several well-known methods. In [12], a multi-objective NSGAII with local search strategies is proposed. Mitra and Banka used homogeneity and size as their main objectives. In the aforementioned methods, the number of detected genes in a bicluster is unreasonably big (sometimes this number is one fourth of a dataset) which increases the probability of having irrelevant genes in biclusters.

To our knowledge, only NSGA-II has been used in multi-objective evolutionary algorithm in biclustering detection algorithms. Strength Pareto front evolutionary algorithm (SPEA2) behaves very similar to NSGAII with more uniform and wider distribution of results. In addition, SPEA2 has advantages on higher dimensional objective spaces [13].

In [14], Roh and Park proposed an evolutionary computation algorithm using the order preserving sub-matrix constraint (ECOPSM). Their approach was based on a ranking matrix. Using evolutionary computation algorithm they searched for biclusters with a certain condition length. They reported better results in comparison to Cheng and Church algorithm. Divina and Aguilar-Ruiz [15] proposed a genetic algorithm search technique. They detected biclusters with bigger size, higher row variance and smaller MSR. However, combining these conflicting objectives is not possible in a single function. Biclustering optimization with two or more conflicting objectives leads to a multi-objective biclustering problem.

This work is based on our previous work [16]. The main difference is how we select the final bicluster among Pareto front solutions. In our previous work we used a merit function to select the bicluster among the parent front. Our approach is based on the multi-objective evolutionary algorithm SPEA2 [13]. The objectives include MSR to be minimized and the size of bicluster to be maximized. We use binary encoding to represent biclusters and k-mean method to select the best bicluster in population among Pareto front solutions. In order to generate the next population, crossover and mutation with a heuristic search is applied. The performance of the proposed method is compared with several well-known biclustering techniques for simulated and real microarray datasets.

2. THE PROPOSED METHOD

The flowchart of the proposed algorithm is given in Figure 3. Let $G=\{g_1,g_2,\ldots,g_N\}$ represents a set of genes and $C=\{c_1,c_2,\ldots,c_M\}$ represents a set of conditions. The data is represented as an $N\times M$ expression matrix as in Figure 1. In this method we used binary encoding scheme with a fixed size $N+M$ to show each bicluster. The first N bits represented genes and the remaining M bits represented the conditions of the gene expression matrix. One showed the existence of the corresponding gene or condition in a random generated bicluster otherwise it was set to zero. The generation of a predefined number of biclusters constituted the initial population.

In order to generate the Pareto front of solutions, strength value, raw fitness, distance to neighbors, and final fitness values were calculated. Then the biclusters with raw fitness equal to zero which showed the non-dominated results moved to archive population. If the number of non-dominated results was smaller than the archived size, the archive filled with dominated results with smaller fitness function which increased the diversity of the algorithm. The Pareto front constituted by the members of archive with raw fitness equaled to zero. In order to generate the next population and the offspring for binary tournament selection, one-point crossover and bit string mutation were performed. Genes and conditions underwent the crossover and mutation separately. For each offspring, multi-objective fitness function was calculated based on Equation (6).

2.1 Fitness Function

For the multi-objective optimization we selected MSR to be minimized and size of the biclusters to be maximized to detect biclusters. In order to calculate MSR the original Cheng and Church's equation [4] was used as in Equation (1). This value measures the difference between the real value of an element and its expected value which is calculated from the row mean, column mean, and bicluster mean. The smaller the MSR is, the stronger the correlation of rows and columns.

$$MSR(I,J)=\frac{1}{|I||J|}\sum_{i\in I,j\in J}(e_{ij}-e_{iJ}-e_{Ij}+e_{IJ})^2 \qquad (1)$$

where e_{ij} is an element of the expression matrix and I and J are number of genes and conditions in bicluster respectively. e_{iJ}, e_{Ij} and e_{IJ} are the mean of the i^{th} row, the mean of the j^{th} column, and the mean of the bicluster (I,J), respectively and are calculated by Equations (2)-(4).

$$e_{iJ}=\frac{\sum_{j\in J}e_{ij}}{|J|} \qquad (2)$$

$$e_{Ij}=\frac{\sum_{i\in I}e_{ij}}{|I|} \qquad (3)$$

$$e_{IJ}=\frac{\sum_{i\in I,j\in J}e_{ij}}{|I||J|} \qquad (4)$$

In order to calculate the second objective, i.e. the size of the bicluster, we used the function as in [15].

$$Volume=w_g\times\frac{\delta}{I}+w_c\times\frac{\delta}{J} \qquad (5)$$

where I and J are the number of genes and conditions of the bicluster, respectively, w_g and w_c are weights of the number of genes and conditions in dataset to balance the number of selected genes and conditions. The value of w_g is set to one as in Divina and Aguilar-Ruiz [15]. w_c varies between one to 10 and it is the ratio of the number of genes and the number of conditions in dataset. *Volume* defines the number of detected elements in the bicluster.

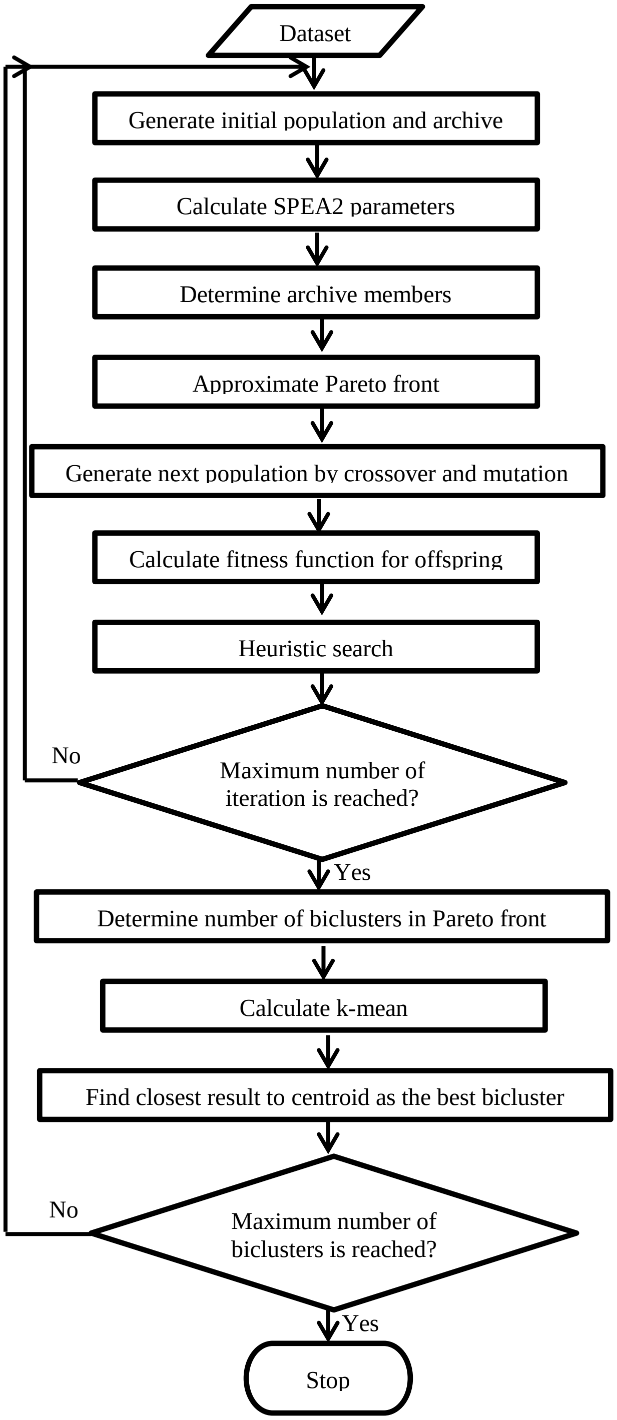
Dataset
Generate initial population and archive
Calculate SPEA2 parameters
Determine archive members
Approximate Pareto front
Generate next population by crossover and mutation
Calculate fitness function for offspring
Heuristic search
No
Maximum number of iteration is reached?
Yes
Determine number of biclusters in Pareto front
Calculate k-mean
Find closest result to centroid as the best bicluster
No
Maximum number of biclusters is reached?
Yes
Stop

Figure 3. Flowchart of proposed method

After determining each objective, the overall fitness function is calculated as follows:

$$fitness = \begin{cases} f_1 = MSR(I,J) \\ f_2 = w_d \end{cases} \tag{6}$$

The smaller the fitness value is, the higher the quality of the detected bicluster.

2.2 Heuristic Search

Due to random generation of biclusters in evolutionary algorithms, the probability of unrelated elements raised at several steps of the algorithm. A greedy local search was applied based on CC method [4] to remove unwanted genes and conditions. The main steps of the algorithm are summarized in Figure 4.

The parameter α determined the rate of element deletion. In [12] they set this value to 1.4 but we used a dynamic value for α which decreased in each step to increase the rate of deletion. This deleted genes and conditions with higher MSR values first and speed up the process of dropping MSR under δ. Pre-selected number of new genes and conditions were added to biclusters until MSR remains under δ.

2.3 Selecting the Best Bicluster

In order to select the final bicluster among Pareto front solutions, we used k-mean algorithm to cluster the Pareto front solutions. The problem that raised here was the number of clusters k. To determine the optimal number of clusters silhouette plot was used as in [17]. Differing values were assigned to k and for each value silhouette width was calculated. The highest average silhouette width indicated the number of clusters. After defining the number of clusters k we applied the k-mean algorithm to the Pareto front biclusters and the centroid of each cluster was calculated. We selected the knee cluster of Pareto front solutions as it is shown in Figure 2. The reason of selecting knee solutions was that they show a good trade-off of two objectives. The best bicluster was the closest bicluster to the center of knee cluster. The value of k is set from 3 to 10 in this algorithm. The values of parameters in the proposed method are listed in Table 1. These values were selected based on the original values in the original method [13].

Table 1. Parameter Selection of Proposed Method

Parameter	Value	Parameter	Value
Maximum Iteration	100	Crossover Probability	0.8
Population Size	50	δ simulated dataset	5
Archive Size	20	δ S.C. yeast dataset	300
Mutation Probability	0.2	δ lymphoma dataset	1200

```
1 Delete all genes satisfying $\frac{1}{|J|}\sum_{i\in I}(e_{ij} - e_{iJ} - e_{Ij} + e_{IJ})^2 > \propto\times MSR(I,J)$.
2 Delete all conditions satisfying $\frac{1}{|I|}\sum_{j\in J}(e_{ij} - e_{iJ} - e_{Ij} + e_{IJ})^2 > \propto\times MSR(I,J)$.
3 Add a number of genes satisfying $\frac{1}{|J|}\sum_{i\in I}(e_{ij} - e_{iJ} - e_{Ij} + e_{IJ})^2 \leq MSR(I,J)$.
4 Add a number of conditions satisfying $\frac{1}{|I|}\sum_{j\in J}(e_{ij} - e_{iJ} - e_{Ij} + e_{IJ})^2 \leq MSR(I,J)$.
```

Figure 4. Pseudo code of proposed heuristic method

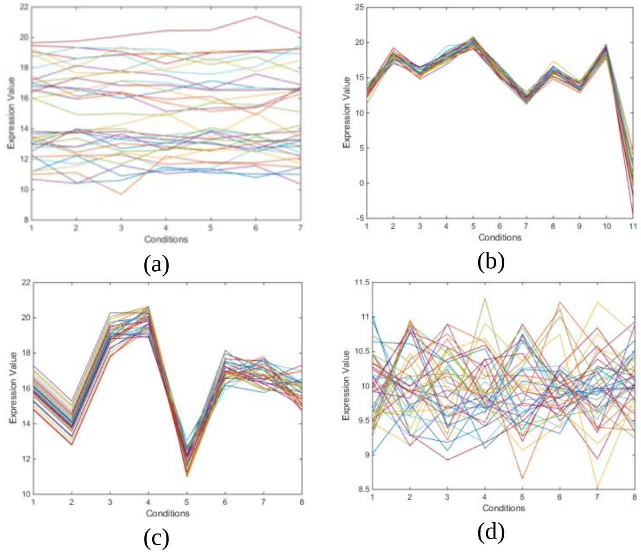

Figure 5. Detected biclusters in simulated dataset (a) Bicluster 1, Size:35×7, MSR:0.22309, Volume:245; (b) Bicluster 2, Size:24×11, MSR:0.71793, Volume:264; (c) Bicluster 3, Size:32×8, MSR:0.24452, Volume:256; (d) Bicluster 4, Size:39×8, MSR:0.20041, Volume:312

In order to qualify the performance of the proposed method, real and simulated datasets were used. The results were compared with several single-objective and multi-objective methods such as CC [4], DA [15], ISA [18] and OPSM [19] as single-objective algorithms and DMOIOB [9], MOM-aiNet [10] and HNSGAII [11] as multi-objective algorithms. We used BicAT toolbox [20] to run CC, ISA and OPSM methods on S. C. yeast and lymphoma dataset. Other results were taken as authors reported in their papers. Furthermore, a biological significance test had been done to study the biological relationship of extracted biclusters using Gene Ontology (GO), GO-TermFinder [21], and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway, GENECODIS [22]. There are three ontologies available in GO, namely, biological process, molecular function and cellular component [23]. Both KEGG and GO term calculate the p-value of genes found in a bicluster which shows the probability of obtaining genes in a bicluster by chance. The smaller the p-value is, the better the result.

2.4 Simulated Data

A simulated dataset with 200 rows and 40 columns was generated with a noisy background and 4 biclusters degraded by a Gaussian noise as in [24]. Bicluster 1 was 40×7 constant row; bicluster 2 and bicluster 3 were 25×10 and 35×8 constant column, respectively; and bicluster 4 was 40×8 constant value. The results of our algorithm on the simulated dataset are shown in Figure 5.

In order to compare the proposed method with CC, OPSM and ISA, we ran BicAT with predefined values of the algorithms on simulated data. Among the algorithms CC had the least informative result and it could not detect any bicluster. Only the scattered parts of bicluster 2 and bicluster 3 were detected by OPSM algorithm. ISA had the best results but it detected only 25 rows and 7 conditions of bicluster 1 and 31 rows and 4 conditions of bicluster 3.

Table 2. Comparison of biclusters of different algorithms for S. C. Yeast dataset

Algorithm	Average MSR	Average volume
Proposed Method	122.6096	802.87
DMOIOB	201.86	2841.08
MOIB	202.32	2638.74
MOM-aiNet	178.28	1831.80
BIC-aiNet	194.65	2556.60
DA	202.68	403.48
CC	204.29	1576.98
ISA	281.59	79.22
OPSM	320.39	1533.31

Table 3. Comparison of biclusters of different algorithms for human B-cell lymphoma dataset

Algorithm	Average MSR	Average volume
Proposed Method	529.8685	1196.5408
DMOIOB	832.79	7106.51
MOIB	839.74	6918.29
MOM-aiNet	1079.26	2826.38
BIC-aiNet	1154.08	11567.61
CC	909.23	606.39
ISA		439.98
OPSM		1494.23

The comparison results of different algorithms for S. C. yeast and human B-cell lymphoma datasets are shown in Table 2 and Table 3 respectively. From the results it is clear that the proposed method achieves smaller results in mean value of MSR compared to other methods. On the other hand, the average size of the biclusters is only larger than DA and ISA. However, the biological significance of these algorithms is very low. The main advantage of the proposed method over the other methods is that it can balance between the number of genes in the bicluster and the number of genes that have biologically significance relationship in the bicluster.

The enrichment of biclusters is verified using GOTermFinder (http://db.yeastgenome.org/cgi-bin/GO/goTermFinder) with three ontologies: biological processes, molecular functions, and cellular components. The p-value shows the probability of finding genes with a specific GO term in a bicluster by chance. The smaller the p-value is, the better the result. GENECODIS [22] is also used to verify our results and calculate the singular KEGG pathway. (http://genecodis.cnb.csic.es/).

The evaluation of the biggest detected bicluster using GO term and KEGG are shown in Table 4 to Table 7. The bicluster includes 86 genes and 15 conditions. The MSR and size are 131.106 and 1290, respectively. The results show that the bicluster is significantly enriched in comparison to background occurrence frequency.

Cellular component refers to a place in the cell where a gene product is activated as in Table 4. Biological process (Table 5) shows the contribution of detected genes or genes products in a biological objective. And finally in Table 6, molecular function shows biochemical activity of a gene product without specifying where or when it actually occurred [23].

Table 4Table 6 are listed significant GO terms in detected biclusters. Cluster frequently shows the number of annotated GO terms in the genes of detected bicluster. In background frequency the number of annotated GO term in the background is shown. Background set includes genes in the database with at least one GO annotation. In addition, the p-value is calculated according to the Equation (7) where N is the total number of genes in the background set, M is the number of annotated genes in the background set, n is the total number of detected genes and k is the number of annotated genes within the detected genes [21].

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}} \tag{7}$$

3. Conclusions

Evolutionary algorithms are powerful search algorithms to find the biclusters in a dataset due to their ability to find near optimal solutions. Multi-objective search strategy is used to handle multiple conflicting objectives. In this paper, we used SPEA2 as multi-objective evolutionary algorithm with a heuristic search. In order to select the best bicluster among Pareto front solution, k-mean clustering algorithm was used. To validate the performance of the proposed method, experiments on simulated and real gene expression dataset were performed and the results were compared with several well-known methods in literature. Two biological significance tests (GO and KEGG) were performed which showed the significant enrichment of detected genes in our method. The proposed method only detects additive patterns in biclusters; for future direction we want to also detect multiplicative patterns in biclusters.

Table 4. Cellular Component Ontology of GOTermFinder

Gene Ontology term	Cluster frequency	Background frequency	P-value
cell part | AmiGO	84 out of 86 genes, 97.7%	5533 out of 7163 background genes, 77.2%	9.04e-06
cell | AmiGO	84 out of 86 genes, 97.7%	5533 out of 7163 background genes, 77.2%	9.04e-06
macromolecular complex | AmiGO	49 out of 86 genes, 57.0%	2193 out of 7163 background genes, 30.6%	4.07e-05
intracellular part | AmiGO	81 out of 86 genes, 94.2%	5252 out of 7163 background genes, 73.3%	7.93e-05
intracellular | AmiGO	81 out of 86 genes, 94.2%	5266 out of 7163 background genes, 73.5%	9.52e-05
protein complex | AmiGO	37 out of 86 genes, 43.0%	1484 out of 7163 background genes, 20.7%	0.00028
nucleolus | AmiGO	15 out of 86 genes, 17.4%	329 out of 7163 background genes, 4.6%	0.00091
transferase complex | AmiGO	15 out of 86 genes, 17.4%	348 out of 7163 background genes, 4.9%	0.00179
catalytic complex | AmiGO	17 out of 86 genes, 19.8%	475 out of 7163 background genes, 6.6%	0.00504
nucleus | AmiGO	44 out of 86 genes, 51.2%	2183 out of 7163 background genes, 30.5%	0.00586

Table 5. Biological Process Ontology of GOTermFinder

Gene Ontology term	Cluster frequency	Background frequency	P-value
ncRNA metabolic process | AmiGO	20 out of 86 genes, 23.3%	524 out of 7163 background genes, 7.3%	0.00123
cellular component biogenesis | AmiGO	30 out of 86 genes, 34.9%	1115 out of 7163 background genes, 15.6%	0.00353
establishment of localization in cell | AmiGO	23 out of 86 genes, 26.7%	757 out of 7163 background genes, 10.6%	0.00885

Table 6. Molecular Function Ontology of GOTermFinder

Gene Ontology term	Cluster frequency	Background frequency	P-value
RNA methyltransferase activity | AmiGO	5 out of 86 genes, 5.8%	36 out of 7163 background genes, 0.5%	0.00543

Table 7. Singular Enrichment Analysis of KEGG Pathway where NGR is the number of annotated genes in the reference list; TNGR is the total number of genes in the reference list; NG is the number of annotated genes in the input list; TNG is the total number of genes in the input list; Hyp is the hypergeometric p-Value; and Hyp* is the corrected hypergeometric p-Value

Genes	NGR	TNGR	NG	TNG	Hyp	Hyp*	Annotations
YDR156W, YNL113W, YDL150W, YOR341W	30	7109	4	85	0.000411578	0.0123474	(KEGG) 03020: RNA polymerase
YDR156W, YER099C, YDR305C, YNL113W, YDL150W, YOR341W	94	7109	6	85	0.000859698	0.0128955	(KEGG) 00230: Purine metabolism
YDR156W, YNL113W, YDL150W, YEL021W, YOR341W	70	7109	5	85	0.00142817	0.0142817	(KEGG) 00240: Pyrimidine metabolism

4. REFERENCES

[1] Rew, D. A. DNA microarray technology in cancer research. *European Journal of Surgical Oncology (EJSO)*, 27, 5 2001), 504-508.

[2] D'haeseleer, P. How does gene expression clustering work? *Nature biotechnology*, 23, 12 2005), 1499-1501.

[3] Zhao, H., Liew, A. W. C., Xie, X. and Yan, H. A new geometric biclustering algorithm based on the Hough transform for analysis of large-scale microarray data. *Journal of Theoretical Biology*, 251, 2 2008), 264-274.

[4] Cheng, Y. and Church, G. M. *Biclustering of expression data*. American Association for Artificial Intelligence (AAAI), City, 2000.

[5] Cheng, K. O., Law, N. F., Siu, W. C. and Liew, A. W. C. Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization. *BMC bioinformatics*, 9, 1 2008), 210.

[6] Zhao, H., Liew, A. W. C., Wang, D. Z. and Yan, H. Biclustering analysis for pattern discovery: Current techniques, comparative studies and applications. *Current Bioinformatics*, 7, 1 2012), 43-55.

[7] Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S. and Coello Coello, C. A. A survey of multiobjective evolutionary algorithms for data mining: Part I. *Evolutionary Computation, IEEE Transactions on*, 18, 1 2014), 4-19.

[8] Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S. and Coello Coello, C. A. A Survey of Multi-Objective Evolutionary Algorithms for Data Mining: Part-II2014).

[9] Liu, J., Li, Z., Hu, X., Chen, Y. and Park, E. K. Dynamic biclustering of microarray data by multi-objective immune optimization. *BMC genomics*, 12, Suppl 2 2011), S11.

[10] Coelho, G. P., de França, F. O. and Von Zuben, F. J. *A multi-objective multipopulation approach for biclustering*. Springer, City, 2008.

[11] Seridi, K., Jourdan, L. and Talbi, E.-G. *Multi-objective evolutionary algorithm for biclustering in microarrays data*. IEEE, City, 2011.

[12] Mitra, S. and Banka, H. Multi-objective evolutionary biclustering of gene expression data. *Pattern Recognition*, 39, 12 2006), 2464-2477.

[13] Zitzler, E., Laumanns, M. and Thiele, L. SPEA2: Improving the strength Pareto evolutionary algorithm. In *Proceedings of* (2001). Eidgenössische Technische Hochschule Zürich (ETH), Institut für Technische Informatik und Kommunikationsnetze (TIK), [insert City of Publication],[insert 2001 of Publication].

[14] Roh, H. and Park, S. *A novel evolutionary algorithm for bi-clustering of gene expression data based on the order preserving sub-matrix (opsm) constraint*. IEEE, City, 2008.

[15] Divina, F. and Aguilar-Ruiz, J. S. Biclustering of expression data with evolutionary computation. *Knowledge and Data Engineering, IEEE Transactions on*, 18, 5 2006), 590-602.

[16] Golchin, M., Davarpanah, S. H. and Liew, A. W. C. *Biclustering analysis of gene expression data using multi-objective evolutionary algorithms*. City, 2015.

[17] Chaudhari, P., Dharaskar, R. and Thakare, V. M. Computing the most significant solution from Pareto front obtained in multi-objective evolutionary. *International Journal of Advanced Computer Science and Applications*, 1, 4 2010), 63-68.

[18] Ihmels, J., Bergmann, S. and Barkai, N. Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20, 13 2004), 1993-2003.

[19] Ben-Dor, A., Chor, B., Karp, R. and Yakhini, Z. Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of computational biology*, 10, 3-4 2003), 373-384.

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

[21] Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M. and Sherlock, G. GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20, 18 2004), 3710-3715.

[22] Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. M. and Pascual-Montano, A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome biology*, 8, 1 2007), R3.

[23] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S. and Eppig, J. T. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25, 1 2000), 25-29.

[24] Gan, X., Liew, A. W. C. and Yan, H. Discovering biclusters in gene expression data based on high-dimensional linear geometries. *BMC bioinformatics*, 9, 1 2008), 209-223.