Improved Biclustering Algorithm Based on Cuckoo Search and Firefly Algorithm

FAN Zhenhao\(^1\) and OU Ling\(^2\)
College of Computer and Information Science, Southwest University, Chongqing 400700, China
E-mail: \(^1\) fan@swu.edu.cn, \(^2\) ouling@swu.edu.cn

Abstract. Biclustering the gene expression data, which focuses on finding a subset of genes and a subset of experimental conditions simultaneously, is a good method to help researchers understanding the data. It can be regarded as optimization problem. Unfortunately, the biclustering result is always fall into local optimal area. To deal with this problem, this paper propose CFB biclustering algorithm. That is, combining Cuckoo Search and Firefly Algorithm for biclustering. The experimental results show that CFB inherits not only the global optimization ability of Cuckoo Search but also the fast convergence speed of Firefly Algorithm, and is more powful to find the biclusters of the gene expression data.

1. Introduction
With the rapid development of biotechnology, people can investigate thousands of genes by measureing their expression levels under different experimental conditions. Usually, these data are organized as a matrix, where each row corresponds to a gene and each column corresponds to a sample or an experimental conditions. The struct of gene expression data is showing in Figure 1

![Figure 1. Struct of gene expression data](image)

Clustering techniques became very important tools for analysising gene expression data. Traditional clustering algorithm allow researchers to improve the understanding of the functions of the genes from an organism, like functional annotation, tissue classification and motif identification[1–3]. However, there are some drawbacks, such as, each gene or experimental condition is assigned to only one cluster, and each gene or experimental condition must be assigned to some cluster. Thus, a traditional clustering method may not be convincing enough.
In order to overcome the limitations, Cheng and Church[4] propose the concept of bicluster firstly. The biclustering algorithm is actually a multi-objective optimization local search algorithm. It’s unreasonable that put the complicated gene data into the biclustering algorithm directly. This paper propose CFB, which combining CS and FA for biclustering.

2. Related work
In this section, we will elaborate the related work about biclustering algorithm and swarm intelligence algorithm.

As Figure 2 shows, unlike traditional clustering method that treat similarity as a function of pairs of genes or pairs of conditions, the bicluster model measures coherence within the subset of genes and conditions. So that we can find out the biological information hidden in local data.

As for a gene matrix $A(X,Y)$, $B(I,J)$ is a submatrix of $A$, which $I$ is a subset of rows(genes), and $J$ is a subset of columns(conditions). The $B(I, J)$ is called a bicluster, as long as its MSR (mean squared residue) is smaller than a given threshold $\sigma$. The MSR design formulas is:

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

where $a_{ij}$ is valid element of $B(I, J)$, and, $a_{iJ}$, $a_{Ij}$ and $a_{IJ}$ are the row average, the column average and the matrix average.

Recent years, there are many algorithms for biclustering. Ayadi[5] classified the biclustering algorithms into systematic search algorithms and stochastic search algorithms. Among the representative systematic search algorithms are CC[4], BiMine[6], OPSM[7], FLOC[8], SAMBA[9], OP-Cluster[10], Motifs[11], Bimax[12], and Paid[13], etc. The stochastic searching algorithms have SEBI[14], SSB[15], CMOPSOB[16], and SAB[17].

Each of these algorithms has its own advantages and disadvantages. Mostly, facing the complex gene data, the systematic search algorithms are easily caught into local optimal solutions. Although the stochastic search algorithms can avoiding local optimal solutions, they are not so stable. So, how to solve this problem is the current research hotspots.

There are some researchers apply swarm intelligence algorithms, Such as PSO[18], FA[19], GA[20], CS[21], to biclustering. Swarm intelligence algorithms can be seen as the stochastic search algorithms essentially. But, different algorithm has its own brillig and drawback. The properly combining can get a better result. In this paper, we choose CS and FA. The former has many outstanding properties, like fewer parameters and greater global optimize ability. And FA can converge faster.
3. CFB

3.1. Encoding

Given $A(X, Y)$, a bicluster $B(I, J)$ can be present by $x_p = (g_1, \ldots, g_i, \ldots, g_m, s_1, \ldots, s_j, \ldots, s_n)$, $p = 1, \ldots, N$. Where $N, m, n$ are the population number, the number of genes and the number of samples of $A$. When the $i$th gene or $j$th sample in $A$ is selected as $B$, $g_i = 1$ or $s_j = 1$, otherwise, $g_i = 0$ or $s_j = 0$. $1 \leq i \leq m$ and $1 \leq j \leq n$.

The original swarm intelligence algorithm solutions (nests, fireflies) are multidimensional continuous values that need to be mapped to corresponding bit strings before they can be used to represent double clusters. The usual practice is to set the upper limit to 1, the lower limit to 0, and then determine whether it is greater than 0.5, and map the real value to a bit value. As Figure 3 shows:

![Figure 3. Continuous solution mapping to bicluster](image)

3.2. Fitness Function

For swarm intelligence algorithms, the design of the fitness function has a decisive influence on the final result. Generally, there are MSR, GV (Gene Volume), CV (Condition Volume) and Var (Row Variance) for evaluating the bicluster. MSR represents the consistency of the fluctuations in the expression of genes, and the smaller the value, the more consistent the changes in the expression levels of genes. Var represents the range of fluctuations in the expression of genes, and the greater the range of fluctuations, the higher the credibility associated with genes. Mostly, the smaller the GV and CV, the smaller the MSR and Var; the larger the GV and CV, the larger the MSR and Var become. However, if the volume of the bicluster is too large or too small, the meaning of the bicluster itself is lost, so the algorithm should be able to find the MSR minimum solution under reasonable GV and CV. For a bicluster $B(I, J)$, its fitness value is:

$$f(B) = MSR(B) + \frac{\lambda}{GV(B)} + \frac{\mu}{CV(B)} + \frac{\omega}{Var(B)}$$  \hspace{1cm} (2)

$$MSR(B) = \frac{1}{k \times l} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{IJ})^2$$  \hspace{1cm} (3)

$$Var(B) = \frac{1}{k \times l} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ})^2$$  \hspace{1cm} (4)

where $a_{iJ}, a_{IJ}$ are the $i$-th row, the $j$-th column and the entire average of $B(I, J)$. $k = |I|$, $l = |J|$ are the volume of the gene and experimental conditions in $B$, i.e. GV and CV. $\lambda$, $\mu$, $\omega$ is to solve the problem of different dimensions. The larger the $\lambda$, $\mu$ and $\omega$, the greater the influence of GV, CV and Var on the fitness value, and the value depends on the data set.

The goal of the algorithm is to try to find a solution with a small fitness value, that is, MSR is as small as possible, and Var are as large as possible.

3.3. Algorithm Describe

Algorithm divide original matrix by CS first, then perform FA clustering on the divided submatrix. Its flow chart as Figure 4 shows.
4. Experiment

4.1. Dataset

In order to evaluate the four algorithms, we did experiments on four datasets. They are Leukaemia dataset, Yeast Cell Cycle dataset, DLBCL (Diffuse Large B-cell Lymphoma) dataset and Breast Cancer dataset. The detail information are showing in Table 1.

| Dataset       | GeneVolume | ConditionVolume | λ    | μ    | ω    |
|---------------|------------|-----------------|------|------|------|
| Leukaemia     | 7128       | 72              | 5.0E02 | 1.0E01 | 3.0E-02 |
| Yeast Cell Cycle | 2884     | 17              | 1.0E06 | 1.0E03 | 3.0E05 |
| DLBCL         | 12625      | 21              | 2.0E07 | 1.0E04 | 3.0E07 |
| Breast Cancer | 13666      | 117             | 2.0E03 | 3.0E02 | 1.0E-01 |

4.2. Results

Figure 5 are the fitness curves of each algorithm on four datasets. From the figures, we can conclude that, CFB inherits not only the global optimization ability of CS but also the fast convergence speed of FA.

Table 2 to Table 5 are the evaluation indexes of the bicluster results of each algorithm on each dataset. It can be seen from the tables that, CFB is good at discovering biclusters with small MSR and large volume on most datasets.

5. GO analysis

For the study of gene expression data, the final problem is to find its biological significance. Therefore, we need to verify whether the gene set in the bicluster participates in some biological processes. GO (Gene Ontology)[22] can be used to investigate if a group of genes belonging to a bicluster presents significant enrichment about a specific GO term. GO provides functional
Figure 5. Fitness curves on each dataset

Table 2. Evaluation indexes on Leukaemia

| Algorithm | MSR     | Volume  | GV  | CV  | Var  |
|-----------|---------|---------|-----|-----|------|
| CS        | 0.1769  | 231488  | 3617| 64  | 0.1797|
| FA        | 0.1776  | 214937  | 3643| 59  | 0.1807|
| CFB       | 0.1769  | 269352  | 3741| 72  | 0.1794|

Table 3. Evaluation indexes on Yeast Cell Cycle

| Algorithm | MSR     | Volume  | GV  | CV  | Var  |
|-----------|---------|---------|-----|-----|------|
| CS        | 578.8211| 10822   | 1546| 7   | 735.4996|
| FA        | 640.3960| 11025   | 1575| 7   | 777.6675|
| CFB       | 529.3312| 9972    | 1662| 6   | 709.6392|

Table 4. Evaluation indexes on DLBCL

| Algorithm | MSR     | Volume  | GV  | CV  | Var  |
|-----------|---------|---------|-----|-----|------|
| CS        | 9139    | 56637   | 6293| 9   | 10308|
| FA        | 11752   | 57942   | 6438| 9   | 13256|
| CFB       | 5019    | 39210   | 6535| 6   | 6025 |

Table 5. Evaluation indexes on Breast Cancer

| Algorithm | MSR     | Volume  | GV  | CV  | Var  |
|-----------|---------|---------|-----|-----|------|
| CS        | 0.3128  | 683793  | 6907| 99  | 0.3172|
| FA        | 0.3118  | 631358  | 6938| 91  | 0.3164|
| CFB       | 0.3189  | 787384  | 6968| 113 | 0.3229|
annotation enrichment analysis of biological processes, cell component, and molecular function. Uploading the biclustered gene set to GO and selecting the species category, GO will returns the gene set and enrichment analysis between each biological process, cell component and molecular function. GO uses P-value to measure the degree of enrichment. When the P-value is less than 0.05, it indicates that the biclustered gene set is enriched in the corresponding GO function category.

In this paper, the enrichment analysis of the biclustering results of three algorithms on DLBCL is carried out in Table 6. Only the most significant terms are shown. For example to the bicluster of CFB, the genes are mainly involved in catalytic activity. The tuple \((n=1910, p=3.28E-31)\) represents that out of 1482 genes in bicluster of CFB, 1910 genes belong to catalytic activity function, and the statistical significance is given by the P-value of \(p=3.28E-31\).

| Algorithm | GV | Biological Processes | Molecular Function | Cell Component |
|-----------|----|----------------------|--------------------|---------------|
| CS        | 6293 | lymphocyte activation | kinase binding | transferase complex |
|           |      | \((n=185, p=1.22E-14)\) \((n=326, p=1.44E-18)\) | \((n=326, p=1.44E-18)\) | \((n=269, p=3.03E-06)\) |
| FA        | 6438 | response to steroid hormone | hormone binding | ribonucleoprotein complex |
|           |      | \((n=187, p=3.14E-18)\) \((n=51, p=1.63E-04)\) | \((n=51, p=1.63E-04)\) | \((n=282, p=9.89E-05)\) |
| CFB       | 6968 | regulation of gliogenesis | catalytic activity | pigment granule |
|           |      | \((n=68, p=1.41E-07)\) \((n=1910, p=3.28E-31)\) | \((n=1910, p=3.28E-31)\) | \((n=59, p=2.05E-06)\) |

### 6. Conclusion

Swarm intelligence techniques are based on collective intelligence of groups of simple agents. A few algorithms are very efficient and they are popular tools for solving real-world problems. In this paper, we propose CFB biclustering algorithm which combining the CS and FA for biclustering. And the experiment results show that the algorithm is more suitable for biclustering than CS and FA.

### References

[1] Pablo A Jaskowiak, Ricardo JGB Campello, and Ivan G Costa Filho. Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 10(4):845–857, 2013.

[2] Pablo A Jaskowiak, Ricardo JGB Campello, and Ivan G Costa. On the selection of appropriate distances for gene expression data clustering. In *BMC bioinformatics*, volume 15, page S2. BioMed Central, 2014.

[3] Amos Tanay, Roded Sharan, and Ron Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(suppl_1):S136–S144, 2002.

[4] Yizong Cheng and George M Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.

[5] Wassim Ayadi, Mourad Elloumi, and Jin-Kao Hao. A biclustering algorithm based on a bicluster enumeration tree: application to dna microarray data. *BioData mining*, 2(1):9, 2009.
[6] Wassim Ayadi, Mourad Elloumi, and Jin-Kao Hao. A biclustering algorithm based on a bicluster enumeration tree: application to dna microarray data. *BioData mining*, 2(1):9, 2009.

[7] Amir Ben-Dor, Benny Chor, Richard Karp, and Zohar Yakhini. Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of computational biology*, 10(3-4):373–384, 2003.

[8] Yang Jiong, Wang Haixun, Wang Wei, and P. Yu. Enhanced biclustering on expression data. In *Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings.*, pages 321–327, March 2003.

[9] Amos Tanay, Roded Sharan, and Ron Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(suppl_1):S136–S144, 2002.

[10] Xiaobo Li, Huijuan Lu, and Mingjun Wang. A hybrid gene selection method for multiclass tumor classification using microarray data. *International Journal Bioautomation*, 17(4), 2013.

[11] TM Murali and Simon Kasif. Extracting conserved gene expression motifs from gene expression data. In *Biocomputing 2003*, pages 77–88. World Scientific, 2002.

[12] Amela Prelič, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Bühlmann, Wilhelm Greissem, Lars Hennig, Lothar Thiele, and Eckart Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.

[13] Laura Lazzeroni and Art Owen. Plaid models for gene expression data. *Statistica sinica*, pages 61–86, 2002.

[14] Federico Divina and Jesus S Aguilar-Ruiz. Biclustering of expression data with evolutionary computation. *IEEE transactions on knowledge and data engineering*, 18(5):590–602, 2006.

[15] Juan A Nepomuceno, Alicia Troncoso, and Jesus S Aguilar-Ruiz. Biclustering of gene expression data by correlation-based scatter search. *BioData mining*, 4(1):3, 2011.

[16] Junwan Liu, Zhoujun Li, Xiaohua Hu, and Yiming Chen. Biclustering of microarray data with mospo based on crowding distance. In *BMC bioinformatics*, volume 10, page S9. BioMed Central, 2009.

[17] Kenneth Bryan, Padraig Cunningham, and Nadia Bolshakova. Application of simulated annealing to the biclustering of gene expression data. *IEEE transactions on information technology in biomedicine*, 10(3):519–525, 2006.

[18] Yuhui Shi and Russell C Eberhart. Empirical study of particle swarm optimization. In *Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*, volume 3, pages 1945–1950. IEEE, 1999.

[19] Xin-She Yang. Firefly algorithm, lévy flights and global optimization. In Max Bramer, Richard Ellis, and Miltos Petridis, editors, *Research and Development in Intelligent Systems XXVI*, pages 209–218, London, 2010. Springer London.

[20] Lawrence Davis. Handbook of genetic algorithms. 1991.

[21] Yang Xin-She and Deb Suash. Cuckoo search via lévy flights. In *2009 World Congress on Nature Biologically Inspired Computing (NaBIC)*, pages 210–214, Dec 2009.

[22] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janaan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.