mathematics

MDPI

Article

A Novel Fuzzy Bi-Clustering Algorithm with Axiomatic Fuzzy Set for Identification of Co-Regulated Genes

Kaijie Xu * and Yixi Wang

School of Electronic Engineering, Xidian University, Xi'an 710071, China; 21022100039@stu.xidian.edu.cn
* Correspondence: kjxu@xidian.edu.cn

Abstract: The identification of co-regulated genes and their Transcription-Factor Binding Sites (TFBSs) are the key steps toward understanding transcription regulation. In addition to effective laboratory assays, various bi-clustering algorithms for the detection of the co-expressed genes have been developed. Bi-clustering methods are used to discover subgroups of genes with similar expression patterns under to-be-identified subsets of experimental conditions when applied to gene expression data. By building two fuzzy partition matrices of the gene expression data with the Axiomatic Fuzzy Set (AFS) theory, this paper proposes a novel fuzzy bi-clustering algorithm for the identification of co-regulated genes. Specifically, the gene expression data are transformed into two fuzzy partition matrices via the sub-preference relations theory of AFS at first. One of the matrices considers the genes as the universe and the conditions as the concept, and the other one considers the genes as the concept and the conditions as the universe. The identification of the co-regulated genes (bi-clusters) is carried out on the two partition matrices at the same time. Then, a novel fuzzy-based similarity criterion is defined based on the partition matrices, and a cyclic optimization algorithm is designed to discover the significant bi-clusters at the expression level. The above procedures guarantee that the generated bi-clusters have more significant expression values than those extracted by the traditional bi-clustering methods. Finally, the performance of the proposed method is evaluated with the performance of the three well-known bi-clustering algorithms on publicly available real microarray datasets. The experimental results are in agreement with the theoretical analysis and show that the proposed algorithm can effectively detect the co-regulated genes without any prior knowledge of the gene expression data.

Keywords: axiomatic fuzzy set (AFS); bi-clustering; gene expression; co-regulated genes; partition matrix

MSC: 68U35

Citation: Xu, K.; Wang, Y. A Novel Fuzzy Bi-Clustering Algorithm with Axiomatic Fuzzy Set for Identification of Co-Regulated Genes. *Mathematics* **2024**, *12*, 1659. https://doi.org/10.3390/math12111659

Academic Editor: Xibei Yang

Received: 25 April 2024
Revised: 19 May 2024
Accepted: 24 May 2024
Published: 26 May 2024

1. Introduction

Gene expression clustering allows for an open-ended exploration of the data, without getting lost among the thousands of individual genes [1]. Traditional (global) clustering methods only analyze genes under all experimental conditions or only analyze conditions of all the genes. In practice, in numerous cellular processes, many genes are regularly co-expressed (co-regulated) [2] under some special conditions [3] but behave differently under different conditions. Consequently, mining local co-expressed valuable patterns becomes a vital objective in discovering genetic pathways that are not very clear when clustered globally [4]. Designing algorithms to mine bi-clusters (co-regulated genes) is crucial for uncovering gene regulatory networks, identifying biomarkers and drug targets, reducing data dimensionality, and advancing personalized medicine and basic biological research. These algorithms enhance our understanding of complex biological processes and improve clinical practices, holding significant scientific and societal value. This is the so-called bi-clustering problem. Bi-clustering extends the traditional clustering techniques

by attempting to find (all) subgroups of genes with similar expression patterns under to-be-identified subsets of experimental conditions when applied to gene expression data.

Bi-clustering can discover valuable co-regulated patterns of genes from plenty of gene expression data, which are more helpful in defining genes functioning together than traditional clustering approaches.

The bi-clustering model measures coherence within the subset of genes and conditions. This model is effective in disclosing the involvement of genes or conditions in multipaths, some of which can only be uncovered under the dominance of more consistent ones [5]. The coherence score is usually defined by building a symmetric function of genes and conditions involved, and therefore bi-clustering is a process of simultaneously clustering genes and conditions. A so-called mean squared residue (MSR), defined by Cheng and Church [6], is first introduced and applied to gene expression data transformed by a logarithm and augmented by the additive inverse. Furthermore, the MSR is also the most commonly used index in bi-clustering, and based on which many bi-clustering algorithms have been developed.

1.1. Literature Review

So far, various algorithms have been developed attempting to solve the bi-clustering problem. Popular bi-clustering algorithms, such as the Cheng and Church (CC) algorithm [6], Flexible Overlapped Clusters (FLOCs) [7], Plaid [8], order-preserving sub-matrix (OPSM) [9], Iterative Signature Algorithm (ISA) [10], conserved gene expression MOTIFs (xMOTIFs) [11], and BiMax [12] have drawn much attention in the literature. Emerging algorithms, such as Bayesian Bi-clustering [13], Maximum Similarity Bi-cluster algorithm (MSB) [14], and QUalitative BI-Clustering algorithm (QUBIC) [15] have not been extensively studied. In a word, as of now, the research on the bi-clustering is still at its initial stage. The real power of this clustering strategy is yet to be fully realized due to the lack of effective and efficient algorithms for reliably solving the general bi-clustering problem.

Among the existing bi-clustering algorithms, the CC and the FLOC algorithms are considered the most effective tools for processing gene expression data to date. The CC algorithm is the earliest and the most studied one, and the emerging algorithms are mostly based on the idea of the CC algorithm. The CC algorithm uses the MSR of a bi-cluster as a similarity measure to greedily extract bi-clusters that satisfy a homogeneity constraint. It generates the row and column cluster randomly and then improves the bi-clusters to minimize the MSR value. Only one bi-cluster is identified each time and then replaced by random numbers before identifying the next cluster [16]. Based on this, and with the aim of improving the generic CC algorithm, Yang et al. proposed another well-known method called FLOC [7], where an additional function is introduced to deal with the missing data and to discover the overlapping bi-clusters [17]. Subsequent studies suggest that the MSR is useful only for identifying certain classes of co-expressed genes, but not adequate to detect other transcriptionally co-regulated genes. Another well-known algorithm named QUBIC [15], which can solve the bi-clustering problem in a more general form, was proposed. The QUBIC algorithm can effectively and efficiently identify all statistically significant bi-clusters that cannot be identified by the other bi-clustering algorithms has turned out to be a more useful tool for the identification of co-regulated genes.

Existing bi-clustering algorithms for mining co-regulated genes face several limitations, including inadequate handling of noise and missing data, sensitivity to parameter selection, high computational complexity, difficulty in interpreting results, poor adaptability to various biological samples, lack of stability and consistency, and limited ability to integrate multiple data types. These issues constrain the effectiveness and reliability of bi-clustering algorithms in practical applications, highlighting the need for further improvement and optimization.

1.2. Brief Introduction of a Fuzzy Bi-Clustering Algorithm with Axiomatic Fuzzy Set

Based on the Axiomatic Fuzzy Set (AFS) theory [18], this paper proposes a novel bi-clustering model for the identification of co-regulated genes. The AFS theory facilitates a way to transform data into fuzzy sets (membership functions) and implement their fuzzy logic operations, which provides a flexible and powerful tool for representing human knowledge and emulating the human recognition process. In recent years, AFS theory has received increasing interest [18]. AFS theory takes the uncertainty of randomness and imprecision of fuzziness as a unified and coherent process so that the membership functions are determined by the observed data. In AFS, the fuzzy sets (membership functions) and their logic operations are impersonally and automatically determined by a consistent algorithm according to the distributions of original data (AFS structures and AFS algebras), which is very different from the traditional fuzzy sets that the membership functions are often given by personal intuition and the logic operations are implemented by a kind of triangular norm (t-norm); the attributes of objects in it can be various data types or sub-preference relations, even human intuition descriptions; the distance function and objective function are not required, and any prior knowledge about the dataset is also not required. For a large dimensionality and a huge number of genes, it is impossible or difficult to define the membership functions just by personal intuition and define distance-based functions to implement fuzzy logic operations. Thus, in this paper, we design a bi-clustering algorithm to discover the co-regulated genes based on AFS theory.

From the design perspective, the sub-preference relations theory of AFS is used to build a fuzzy membership (partition) matrix only based on the distributions of original gene expression data, and it does not require any distance measures and prior knowledge about the gene expression data. Specifically, in the proposed scheme, a reference gene is selected at first. Then, considering the genes as the universe and the conditions as the concept, a fuzzy partition matrix is built by the sub-preference relations theory [18]. Similarly, when considering the genes as the concept and the conditions as the universe, another fuzzy partition matrix can also be built. With the two partition matrices, we define a fuzzy-based similarity criterion to measure the similarity of the co-regulated genes under some special conditions. Subsequently, we design a cyclic optimization algorithm to discover the bi-clusters (co-regulated genes). We believe that this is the first time that such a fuzzy-based similarity criterion has been proposed and the first for solving the bi-clustering problem. In addition, an approach based on Fuzzy C-Means (FCM) [19,20] clustering is proposed to select a number of reference genes. Experimental studies completed on real-world gene expression data demonstrate that the proposed approach achieves better performance compared with that of the several well-known methods used for gene expression.

In brief, the major contribution of the paper is to propose a novel bi-clustering algorithm to discover the co-regulated genes. This algorithm does not require any prior knowledge about the gene expression data. To the best of our knowledge, the idea of the proposed approach has not been exposed in previous studies.

This paper is organized as follows. The bi-clustering-related concepts and novel similarity definitions based on the AFS theory, and the principle of the proposed method are presented in Section 2. Section 3 discusses the performance indexes. Section 4 includes experimental setup and covers an analysis of completed experiments. Section 5 covers some conclusions.

2. Bi-Clustering Algorithm with Axiomatic Fuzzy Set for Identification of Co-Regulated Genes

2.1. Problem Definitions

A commonly used way to visualize microarray data for gene expression analyses is to represent the data set as a matrix with rows representing the genes and columns representing the conditions (or the other way around) with each element of the matrix representing the expression value of a gene under a specific condition. Thus, identifying

groups of genes in a microarray data set that share similar expression patterns under to-be-identified conditions is equivalent to finding submatrices with similar properties.

Let $\boldsymbol{A} = [\cdots, a_{ij}, \cdots] \in R^{N\times M}$ be a microarray expression data matrix with a set of genes $\boldsymbol{G} = [G_1, G_2, \cdots, G_N]^T$ and a set of conditions $\boldsymbol{O} = [O_1, O_2, \cdots, O_M]$, where a_{ij} represents an expression value of a gene (G_i) under a condition (O_j). A bi-cluster is basically a sub-matrix ($\boldsymbol{A}_{IJ}$) that exhibits some similar tendency, which can be expressed by $\boldsymbol{A}(I, J)$, where $I \subset N$ and $J \subset M$ are subsets of genes and conditions, respectively. Let $g^* \in \boldsymbol{G}$ be a reference gene; our goal is to find a subset of genes (co-regulated genes, bi-cluster) that are related to g^*. When the reference gene is not known, we can enumerate all genes in the matrix or randomly select several genes as the reference gene subsets. Similar ideas are also used in [14].

2.2. Similarity Definitions with Membership Degree Based on the AFS

Consider the aforementioned gene expression data matrix $\boldsymbol{A}$. First, we use the AFS theory to build two fuzzy partition matrices [21], which are used to define the similarity matrices in this paper. One of the matrices considers the genes as the universe and the conditions as the concept, and the other one considers the genes as the concept and the conditions as the universe. Assume that the two fuzzy partition matrices are $\boldsymbol{U}_G = [\cdots, \mu_{ij}, \cdots]$ and $\boldsymbol{U}_C = [\cdots, u_{ij}, \cdots]$. The calculation of the membership degree based on the AFS is determined as follows [18,22]:

$$\mu(x) = \sup_{i\in I}\left\{\frac{\mathcal{M}[A_i(x)]}{\mathcal{M}(x)}\right\} \tag{1}$$

where $x \in X$, X is the universe of discourse, $\mathcal{M}$ is a finite and positive measure over σ-algebra, and I is a non-empty indexing set [23].

For a gene expression data matrix ($\boldsymbol{A}_{IJ}$) and a given reference gene ($g^* \in \boldsymbol{G}$), define $\boldsymbol{U}_{Gg^*} = \boldsymbol{U}_G - \boldsymbol{\mu}_{g^*} = [\ldots, \delta_{ij}, \cdots]$, $\delta_{ij} = |\mu_{ij} - \mu_{g^*j}|$. Obviously, δ_{ij} can characterize the similarity between the ith gene and the reference gene under the jth condition, and the smaller the δ_{ij} is, the larger the similarity is, and vice versa. Furthermore, $\boldsymbol{U}_C$ is used as another similarity to jointly discover the local co-regulated genes, and this is the so-called proposed bi-similarity.

Let $\boldsymbol{U}_{Gg^*}(N, M)$ be an $N \times M$ gene similarity matrix and $\boldsymbol{U}_{Gg^*}(I, J)$ be a bi-cluster (sub-matrix) of $\boldsymbol{U}_{Gg^*}(N, M)$. For column $j \in J$, we define the dissimilarity score of the j-column in $\boldsymbol{U}_{Gg^*}(I, J)$ as the range of the column, i.e.,

$$\mu(I, j) = \textstyle\sum_{i\in I}\left[\max(\mu_{I,j}) - \min(\mu_{I,j})\right] \tag{2}$$

The dissimilarity score of $\boldsymbol{U}_{Gg^*}(I, J)$ is

$$\mu(I, J) = \frac{1}{J}\sum_{(i\in I), j=1}^{J}\left[\max(\mu_{I,j}) - \min(\mu_{I,j})\right] \tag{3}$$

Let $\boldsymbol{U}_C(N, M)$ be an $N \times M$ condition similarity matrix and $\boldsymbol{U}_C(I, J)$ be a bi-cluster (sub-matrix) of $\boldsymbol{U}_C(N, M)$. For row $i \in I$, we define the dissimilarity score of the i-row in $\boldsymbol{U}_C(I, J)$ as the range of the row, i.e.,

$$u(i, J) = \textstyle\sum_{j\in J}\left[\max(u_{i,J}) - \min(u_{i,J})\right] \tag{4}$$

The dissimilarity score of $\boldsymbol{U}_C(I, J)$ is

$$u(I, J) = \frac{1}{I}\sum_{(j\in J), i=1}^{I}\left[\max(u_{i,J}) - \min(u_{i,J})\right] \tag{5}$$

Consider a bi-cluster $\boldsymbol{A}(I, J)$. If the dissimilarity score $\boldsymbol{U}_{Gg^*}(I, J)$ is low, and the genes under all the conditions in $\boldsymbol{A}(I, J)$ will have similar expression values. However, many

genes are regularly co-expressed under some special conditions; in other words, mining local co-expressed valuable patterns is more meaningful than that of clustering globally. Thus, we use $\boldsymbol{U}_{G_{g^*}}$ and $\boldsymbol{U}_C$ to jointly discover the local co-expressed valuable patterns. For a bi-cluster $\boldsymbol{A}(I,\ J)$, if both $\boldsymbol{U}_{G_{g^*}}\overline{\boldsymbol{U}}_G(I,J)$ and $\boldsymbol{U}_C(N,M)$ are low, then the genes in $\boldsymbol{A}(I,\ J)$ under the J conditions are co-expressed.

Thus, we have completed the creation of two similarity (gene similarity and condition similarity) matrices. Based on the above analysis, we will report the proposed algorithm for discovering the bi-clusters.

2.3. Bi-Similarity Criterion Based on the Similarity Matrices

The algorithm is an essentially greedy algorithm, and it starts with the whole gene expression data matrix $\boldsymbol{A}(N,M)$ as an initial bi-cluster. In the discovery of bi-clusters, we define a novel bi-similarity criterion based on the similarity matrices above. With the use of the bi-similarity criterion, the co-regulation of the genes in the same bi-clusters becomes enhanced.

Let $\mu(N,M)$ and $u(N,M)$ be the dissimilarity scores of $\boldsymbol{U}_{G_{g^*}}(N,M)$ and $\boldsymbol{U}_C(N,M)$, respectively. our goal is to discover a bi-cluster with small dissimilarity scores, such as $\mu(N,M)/\alpha$ and $u(N,M)/\beta$, where α and β are the scale factors of the column and row of the bi-cluster. Thus, we can call the bi-cluster an $(\alpha\beta)$-bi-cluster. An excellent bi-cluster is generated by deleting and adding rows and columns with some particular rules. The sketch is as follows (Algorithm 1):

Algorithm 1 (Node Deletion)

Input: $\boldsymbol{U}_G$ and $\boldsymbol{U}_C$, the two fuzzy partition matrices of the gene expression data matrix, and $\alpha,\ \beta$, the two scale factors of the column and row for the bi-clusters to be found.
Output: $\boldsymbol{A}(I,\ J)$, an $(\alpha\beta)$-bi-cluster that is a sub-matrix of $\boldsymbol{A}(I,\ J)$ with row set I and column set J with the dissimilarity scores no larger than $\mu(N,M)/\alpha$ and $u(N,M)/\beta$, respectively.
Initialization: I and J are initialized to the gene and condition sets in the gene expression data, and $\boldsymbol{A}_{IJ}=\boldsymbol{A}$; a reference gene g^* is given by the user; the maximum acceptable dissimilarity scores of the column and row: $\mu(N,M)/\alpha$, $u(N,M)/\beta$.
Iteration:
(1). Calculate $\mu(I,j)$ for all $j\in J$, $u(i,J)$ for all $i\in I$, and $\mu(I,J)$, $u(I,J)$. If $\mu(I,J)\le\mu(N,M)/\alpha$ and $u(I,J)\le u(N,M)/\beta$, return $\boldsymbol{A}_{IJ}$.
(2). Find column $j\in J$ with largest

$$\mu(I,j)=\sum_{i\in I}\left[\max\left(\mu_{I,j}\right)-\min\left(\mu_{I,j}\right)\right] \tag{6}$$

and row $i\in I$ with largest

$$u(i,J)=\sum_{j\in J}\left[\max\left(u_{i,J}\right)-\min\left(u_{i,J}\right)\right] \tag{7}$$

remove the column if

$$\frac{\alpha\mu(I,j)}{\mu(N,M)}>\frac{\beta u(i,J)}{u(N,M)} \tag{8}$$

else remove the row by updating either I or J.

Clearly, after node deletion, both the row and column dissimilarity scores of the sub-matrix will be reduced. However, the resulting $(\alpha\beta)$-bi-cluster may not be maximal, in the sense that some rows and columns may be added without increasing the dissimilarity scores. Thus, we design another algorithm (Algorithm 2): to refine the bi-clusters.

Algorithm 2 (Node Addition)

Input: $\boldsymbol{A}_{IJ}$, a sub-matrix of real numbers; I and J signifying an $(\alpha\beta)$-bi-cluster.
Output: $\boldsymbol{A}(I,\ J)$, I', and J' such that $I' \subset I$ and $J' \subset J$ with the property that

$$\mu(I', J') \leq \mu(I, J) \ \ \& \ \ u(I', J') \leq u(I, J) \tag{9}$$

Iteration:
(1). Compute $\mu(I, j)$ for all $j \notin J$, recompute $\mu(I, J)$ and $u(i, J)$, and add the columns $j \notin J$ if $\mu(I, J) \leq \mu(N, M)/\alpha$.

$$\mu(I, J) \leq \frac{\mu(N, M)}{\alpha} \ \ \& \ \ u(I, J) \leq \frac{u(N, M)}{\beta} \tag{10}$$

(2). Compute $u(i, J)$ for all $i \notin I$, recompute the $u(I, J)$ and $\mu(I, J)$, and add the rows $i \notin I$ if

$$\mu(I, J) \leq \frac{\mu(N, M)}{\alpha} \ \ \& \ \ u(I, J) \leq \frac{u(N, M)}{\beta} \tag{11}$$

(3). If nothing is added in the iterate, return the final I and J as I' and J'.

Obviously, after the execution of the node addition algorithm, neither the row dissimilarity score nor the column dissimilarity score will increase. Sometimes, an addition may decrease the score more than any deletion.

2.4. Selection of the Reference Genes

In the algorithm proposed above, the reference genes we are interested in are known in advance. When the reference genes are unknown, we should select a number of genes as the reference genes. In some cases, the reference genes selected are closely related to the quality of the bi-clusters. Furthermore, we usually prefer the size (number of the co-regulated genes and conditions) of the bi-clusters to be as large as possible (discover more co-regulated genes under more conditions). In other words, if a gene has more similar genes under more conditions, then it is more suitable to be a reference. Based on this, we propose a method to select the reference genes.

Firstly, we calculate the fuzzy similarity matrices of all genes under each (jth) condition, and we obtain M similarity matrices; for example,

$$\begin{gathered} \boldsymbol{S}_j = \left[\boldsymbol{s}_1^{(j)},\ \boldsymbol{s}_2^{(j)},\ \cdots, \boldsymbol{s}_i^{(j)}\ \cdots,\ \boldsymbol{s}_N^{(j)}\right] \\ \boldsymbol{s}_i^{(j)} = \left[s_1^{(ji)},\ s_2^{(ji)},\ \cdots,\ s_k^{(ji)},\ \cdots,\ s_N^{(ji)}\right]^T \\ i = 1, 2, \cdots, N;\ \ j = 1, 2, \cdots, M;\ \ \ k = 1, 2, \cdots, N \end{gathered} \tag{12}$$

where T stands for the transpose operation. Let $\boldsymbol{V}_i^{(j)} = \left[\max(\boldsymbol{s}_i^{(j)}),\ \min(\boldsymbol{s}_i^{(j)})\right]^T$ be the prototypes of $\boldsymbol{s}_i^{(j)}$; based on FCM clustering, we transform $\boldsymbol{s}_i^{(j)}$ into a membership matrix as follows:

$$\begin{gathered} \boldsymbol{\Phi}_i^{(j)} = \begin{bmatrix} \varphi_{k1}^{(ji)} & \varphi_{k2}^{(ji)} \end{bmatrix}^T \in R^{2\times N} \\ \varphi_{kc}^{(ji)} = \frac{\left\| s_k^{(ji)} - v_c^{(ji)} \right\|^{\frac{-2}{m-1}}}{\sum_{h=1}^{2} \left(\frac{1}{\left\| s_h^{(ji)} - v_c^{(ji)} \right\|} \right)^{\frac{2}{m-1}}} \\ c = 1, 2;\ k = 1, 2, \cdots, N \end{gathered} \tag{13}$$

where $v_c^{(ji)}$ is the c-th element (cluster prototype) of $\boldsymbol{V}_i^{(j)}$; m is a fuzziness exponent (fuzziness coefficient); and $||\bullet||$ stands for the Euclidean distance. $\varphi_{kc}^{(ji)} \in [0,\ 1]$ is the degree of membership of an individual $s_k^{(ji)}$ belonging to the cluster c and satisfies the following condition.

$$\sum_{c=1}^{2} \varphi_{kc}^{(ji)} = 1 \text{ for } k = 1,\ 2,\ \cdots,\ N \tag{14}$$

Then, we construct a function to compute the mean of the large fuzzy similarity of the ith gene under all the M conditions.

$$\varpi_i = \frac{\sum_{j=1}^{M}\sum_{k=1}^{N} s_k^{(ji)} sign\left[\varphi_{k1}^{(ji)} - 0.5\right]}{\delta\left\{\sum_{k=1}^{N} sign\left[\varphi_{k1}^{(ji)} - 0.5\right]\right\} + M\sum_{k=1}^{N} sign\left[\varphi_{k1}^{(ji)} - 0.5\right]} \tag{15}$$

where δ is a unit pulse response function. Generally, we consider that the genes with large ϖ values are more suitable to be references, since they have more similar genes under more conditions as previously described.

3. Performance Indexes

In order to evaluate the performance of the proposed algorithm, two commonly used performance indexes are briefly discussed.

3.1. Variance Index

Given a gene expression data matrix ($\boldsymbol{A}(N,\ M)$) with a set of rows (genes, N) and a set of columns (conditions, M), a bi-cluster is a sub-matrix ($\boldsymbol{A}(I,\ J),\ I \subset N, J \subset M$) of $\boldsymbol{A}(N,\ M)$. a_{ij} is the value in the data matrix $\boldsymbol{A}$ corresponding to row i and column j. We denote by a_{iJ} the mean of the ith row in the bi-cluster, a_{Ij} the mean of the jth column in the bi-cluster, and a_{IJ} the mean of all elements in the bi-cluster. These values are defined by

$$a_{iJ} = \frac{1}{|J|}\sum\nolimits_{j\in J} a_{ij} \tag{16}$$

$$a_{Ij} = \frac{1}{|I|}\sum\nolimits_{i\in I} a_{ij} \tag{17}$$

$$a_{IJ} = \frac{1}{|I||J|}\sum\nolimits_{i\in I, j\in J} a_{ij} = \frac{1}{|I|}\sum_{i\in I} a_{iJ} = \frac{1}{|J|}\sum\nolimits_{j\in J} a_{Ij} \tag{18}$$

The variance [22] is used to evaluate the quality of each bi-cluster $\boldsymbol{A}(I,\ J)$:

$$VAR(I,\ J) = \sum_{i\in I, j\in J} \left(a_{ij} - a_{IJ}\right)^2 \tag{19}$$

the lower the value returned, the better the quality of the bi-cluster will be; a perfect bi-cluster is a sub-matrix with variance equal to zero.

3.2. Mean Fluctuation Degree Index

Mean Fluctuation Degree (MFD) is used to evaluate the changing trends of the genes under each condition transition. the MFD of a bi-cluster is defined as

$$MFD(I,\ J) = \sqrt{\frac{1}{|I||J|}\sum_{i\in I,\ j\in J}\left(\Theta_{ij} - \frac{1}{|I|}\sqrt{\Theta_{ij}}\right)^2} \tag{20}$$

where

$$\begin{gathered}\Theta_{ij} \in \boldsymbol{\Theta} \\ \boldsymbol{\Theta} = 180\arctan\left(\boldsymbol{\Delta}^{\dagger}\boldsymbol{\Xi}\right)/\pi\end{gathered} \tag{21}$$

$$\boldsymbol{\Delta} = \frac{diag\left\{\max(a_{1j})-\min(a_{1j}), \cdots, \max(a_{ij})-\min(a_{ij}), \cdots\right\}}{M-1} \quad i = 1,2,\cdots,N;\ \ j = 1,2,\cdots,M \tag{22}$$

$$\boldsymbol{\Xi} = [\cdots,\ a_{ij} - a_{i(j-1)},\ \cdots] \in R^{N\times(M-1)} \quad i = 1,2,\cdots,N;\ \ j = 2,\cdots,M \tag{23}$$

where subscript † denotes the Moore–Penrose Inverse of the matrix [24]. Obviously, for a bi-cluster, if the genes (rows) have similar changing trends under each condition transition, its MFD will be relatively smaller. Furthermore, if all genes (rows) in the bi-cluster have completely similar (or the same) changing trends under each condition transition, its MFD will be zero.

In particular, for a single-row (or a single-column) "bi-cluster", its VAR and MFD indexes are also zero; however, such a "bi-cluster" is meaningless. To fairly compare the performance of the algorithms, we will drop the resulting bi-clusters with only one row and one column.

4. Experimental Studies

In the following experiments, we compare the performance of the proposed fuzzy bi-clustering (FBC) method with CC, FLOC, and QUBIC methods, which are the two well-known bi-clustering methods commonly used for gene expression. In the experiments, two well-known publicly available real microarray datasets named Yeast (http://arep.med.harvard.edu/biclustering/yeast.matrix) (Accessed on 5 April 2024) and Gordon 2002 [25] are used, which are the most commonly used datasets in bi-clustering. Both the variance [24] and the mean fluctuation degree are taken as the evaluation indexes which are briefly discussed as follows.

The methods are used to find 100 bi-clusters. A concise description of the values of the parameters used in the experiments is given in Table 1. The methods are repeated 10 times; the means and the standard deviations of the experimental results are presented. The experimental results are plotted in Figures 1 and 2. The left part of each graph displays indicators for several algorithms in discovering all bi-clusters (co-regulated genes), which assess the quality of the identified co-expressed genes. The right part features error bars to evaluate the overall performance of the algorithms, summarizing the effectiveness of all identified bi-clusters.

Table 1. Datasets and parameters used in the experiments.

Datasets		**Yeast**	**Gordon-2002**
Number of	Genes	2884	1626
	Conditions	17	181
Threshold of MSR		300	3000
α		5.0	5.5
β		1.8	3.0

It is evident that the algorithm proposed in this study excels in discovering bi-clusters (co-regulated genes), outperforming other comparable algorithms. The proposed algorithm is effective in discovering the quality of bi-clusters by grouping together genes that have trends with more similar fluctuation degrees. Compared with the CC, FLOC, and QUBIC clustering methods, the proposed method demonstrates significant advantages. Specifically, the experimental results highlight that our method consistently achieves higher accuracy and quality in identifying co-expressed gene bi-clusters. The proposed algorithm shows improved robustness and reliability, effectively handling diverse and complex datasets where traditional methods may falter.

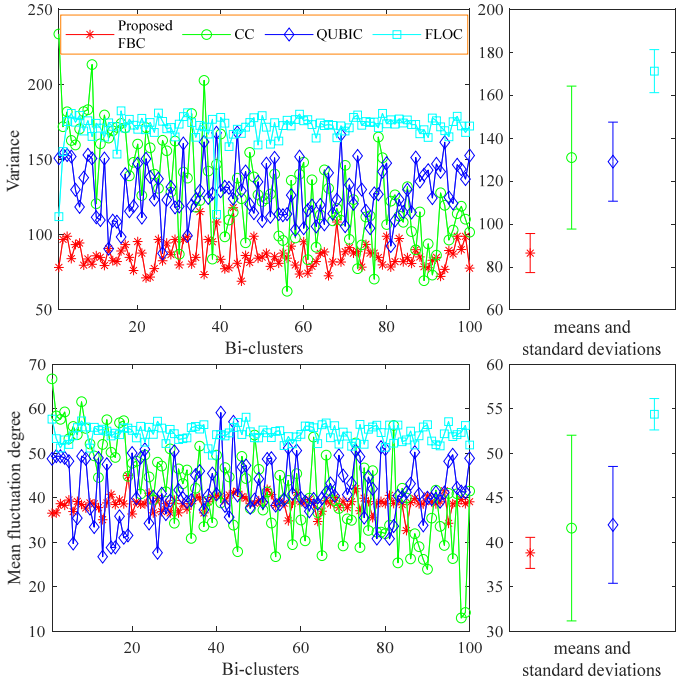

Figure 1. Comparison of the performance of methods on the Yeast dataset.

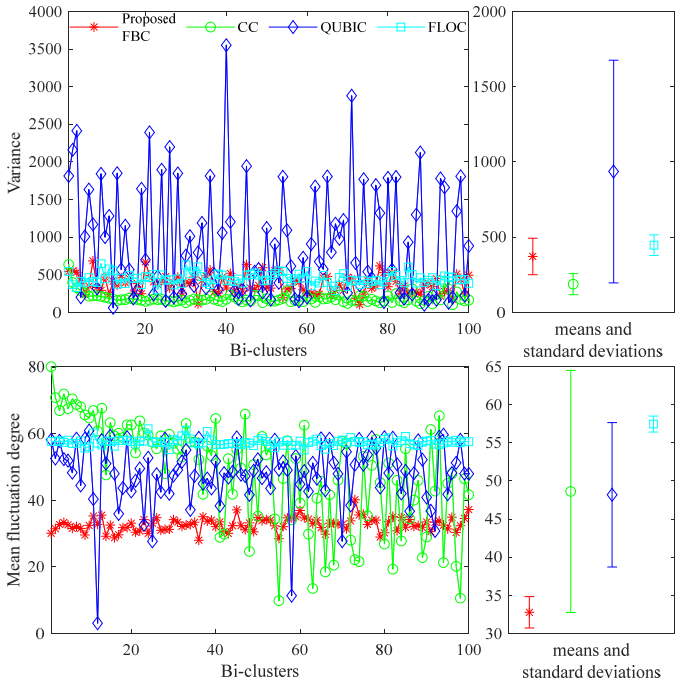

Figure 2. Comparison of the performance of methods on the Gordon 2002 dataset.

To sum up, a robust algorithm to mine co-regulated genes is crucial for uncovering gene regulatory networks, identifying biomarkers and drug targets, reducing data dimensionality, and advancing personalized medicine and basic biological research, which can enhance the understanding of complex biological processes and improve clinical practices, holding significant scientific and societal value.

5. Conclusions

In this research, we designed a bi-clustering algorithm for the identification of co-regulated genes via AFS theory. During the design process, the sub-preference relations

theory of AFS is introduced to construct two fuzzy membership matrices to define a fuzzy-based similarity criterion. With the similarity criterion, a cyclic optimization algorithm is designed to discover the bi-clusters (co-regulated genes). We conducted theoretical analysis and offered a comprehensive suite of experiments. Both the theoretical and experimental results are presented to verify the validity of the proposed method. Experimental results show that the proposed method outperforms the existing algorithms in finding the bi-clusters, and has demonstrated its outstanding performance and great potential for the development of gene expression. To the best of our knowledge, this research scheme is the first proposed, which steadily improves the performance of the bi-clustering.

At the current stage, we have completed a thorough theoretical analysis and conducted a comprehensive suite of experiments to validate our approach. Our theoretical work has laid a strong foundation, and our experimental results have demonstrated the feasibility and potential of our methods under controlled conditions. However, translating these findings into practical applications presents an exciting and valuable avenue for future research.

Future studies could focus on implementing practical experiments in real-world scenarios to assess the robustness and effectiveness of our algorithms. This could involve collaborating with biologists to apply our methods to actual biological datasets, such as those derived from clinical samples or environmental studies. Additionally, exploring the integration of our algorithms with existing bioinformatics tools and pipelines could enhance their usability and impact [26].

Moreover, practical experiments could help identify any unforeseen challenges or limitations that may arise in real-world applications, providing valuable insights for further refinement and optimization of our methods. By bridging the gap between theoretical analysis and practical implementation, we aim to contribute to the development of more robust, reliable, and widely applicable tools for gene expression analysis and other areas of computational biology.

Author Contributions: The authors confirm their contribution to the paper as follows. K.X.: methodology, writing—original draft, software, and visualization; Y.W.: investigation, writing—review and editing, and validation; K.X.: supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Nos. 62101400, 72101075, 72171069, and 92367206), in part by the China Postdoctoral Science Foundation under Grant 2023M732743, and in part by the Shaanxi Fundamental Science Research Project for Mathematics and Physics under Grant 22JSQ032.

Data Availability Statement: Data will be made available upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dhaeseleer, P. How does gene expression clustering work? *Nat. Biotechnol.* **2005**, *23*, 1499–1501. [CrossRef] [PubMed]
2. Pattini, L.; Sassi, R.; Cerutti, S. Dissecting heart failure through the multiscale approach of systems medicine. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 1593–1603. [CrossRef] [PubMed]
3. Mulqueen, R.M.; Pokholok, D.; Norberg, S.J.; Torkenczy, K.A.; Fields, A.J.; Sun, D.; Sinnamon, J.R.; Shendure, J.; Trapnell, C.; O'Roak, B.J.; et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **2018**, *36*, 428–431. [CrossRef] [PubMed]
4. Mishra, D.; Shaw, K.; Mishra, S. Gene expression network discovery: A pattern based biclustering approach. In Proceedings of the 2011 International Conference on Communication, Computing & Security, ACM, Rourkela, Odisha, India, 12–14 February 2011; pp. 307–312.
5. Yang, J.; Wang, H.; Wang, W. Enhanced biclustering on expression data. In Proceedings of the Third IEEE Symposium on Bioinformatics and Bioengineering, Bethesda, MD, USA, 10–12 March 2003; pp. 321–327.
6. Cheng, Y.; Church, G.M. Biclustering of expression data. In Proceedings of the Conference on Intelligent Systems for Molecular Biology (ISM), San Diego, CA, USA, 19–23 August 2000; pp. 93–103.
7. Yang, J.; Wang, H.; Wang, W.; Yu, P.S. An improved biclustering method for analyzing gene expression profiles. *Int. J. Artif. Intell. Tools* **2005**, *14*, 771–789. [CrossRef]
8. Lazzeroni, L.; Owen, A. Plaid models for gene expression data. *Stat. Sin.* **2000**, *12*, 61–86.

9. Ben-Dor, A.; Chor, B.; Karp, R.; Yakhini, Z. Discovering local structure in gene expression data: The order-preserving submatrix problem. *J. Comput. Biol.* **2003**, *10*, 373–384. [CrossRef] [PubMed]
10. Bergmann, S.; Ihmels, J.; Barkai, N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **2003**, *67*, 031902. [CrossRef]
11. Murali, T.M.; Kasif, S. Extracting conserved gene expression motifs from gene expression data. In Proceedings of the Pacific Symposium on Biocomputing, Lihue, HI, USA, 3–7 January 2003; pp. 77–88.
12. Prelić, A.; Bleuler, S.; Zimmermann, P.; Wille, A.; Bühlmann, P.; Gruissem, W.; Hennig, L.; Thiele, L.; Zitzler, E. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **2006**, *22*, 1122–1129. [CrossRef]
13. Gao, C.; McDowell, I.C.; Zhao, S. Context specific and differential gene co-expression networks via Bayesian biclustering. *Comput. Biol.* **2016**, *12*, e1004791. [CrossRef]
14. Liu, X.; Wang, L. Computing the maximum similarity bi-clusters of gene expression data. *Bioinformatics* **2006**, *23*, 50–56. [CrossRef]
15. Li, G.; Ma, Q.; Tang, H.; Paterson, A.H.; Xu, Y. QUBIC: A qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res.* **2009**, *37*, e101. [CrossRef] [PubMed]
16. Shruthi, M.P.; Saravana, K.E. A survey on biclustering. *Int. J. Innov. Res. Sci. Technol.* **2016**, *3*, 2349–6010.
17. Khalid, B.; Allab, K. Bi-clustering continuous data with self-organizing map. *Neural Comput. Appl.* **2013**, *22*, 1551–1562.
18. Liu, X.; Jia, W.; Wang, Y.; Guo, H.; Ren, Y.; Li, Z. Knowledge discovery and semantic learning in the framework of axiomatic fuzzy set theory. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, 1268–1292. [CrossRef]
19. Lian, C.; Ruan, S.; Denoeux, T.; Li, H.; Vera, P. Spatial evidential clustering with adaptive distance metric for tumor segmentation in FDG-PET images. *IEEE Trans. Biomed. Eng.* **2018**, *65*, 21–30. [CrossRef] [PubMed]
20. Xu, K.; Pedrycz, W.; Li, Z.; Nie, W. Constructing a virtual space for enhancing the classification performance of Fuzzy clustering. *IEEE Trans. Fuzzy Syst.* **2018**, *27*, 1779–1792. [CrossRef]
21. Xu, K.J.; Pedrycz, W.; Li, Z.W.; Nie, W.K. High-accuracy signal subspace separation algorithm based on gaussian kernel. *IEEE Trans. Ind. Electron.* **2019**, *66*, 491–499. [CrossRef]
22. Madeira, S.C.; Oliveira, A.L. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB)* **2004**, *1*, 24–45. [CrossRef]
23. Ren, Y.; Song, M.; Liu, X. New approaches to the fuzzy clustering via AFS theory. *Int. J. Inf. Syst. Sci.* **2007**, *3*, 307–325.
24. Stanev, D.; Moustakas, K. Simulation of constrained musculoskeletal systems in task space. *IEEE Trans. Biomed. Eng.* **2017**, *65*, 307–318. [CrossRef]
25. Li, X.; Wong, K.-C. Evolutionary multiobjective clustering and its applications to patient stratification. *IEEE Trans. Cybern.* **2019**, *49*, 1680–1693. [CrossRef] [PubMed]
26. Shrimankar, D.D.; Durge, A.R.; Sawarkar, A.D. Heuristic analysis of genomic sequence processing models for high efficiency prediction: A statistical perspective. *Curr. Genom.* **2022**, *23*, 299–317. [CrossRef] [PubMed]