A NEW STUDY ON BICLUSTERING TOOLS, BICLUSTERS VALIDATION AND EVALUATION FUNCTIONS

Haifa Ben Saber¹,² and Mourad Elloumi¹,³

¹Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE) at National Superior School of Engineers of Tunis (ENSIT) - Tunis university, Tunis, Tunisia
²Time université, ³University of Tunis-El Manar, Tunisia

ABSTRACT

There are subsets of genes that have similar behavior under subsets of conditions, so we say that they coexpress, but behave independently under other subsets of conditions. Discovering such coexpressions can be helpful to uncover genomic knowledge such as gene networks or gene interactions. That is why, it is of utmost importance to make a simultaneous clustering of genes and conditions to identify clusters of genes that are coexpressed under clusters of conditions. This type of clustering is called biclustering.

Biclustering is an NP-hard problem. Consequently, heuristic algorithms are typically used to approximate this problem by finding suboptimal solutions. In this paper, we make a new survey on tools, biclusters validation and evaluation functions.

KEYWORDS

Biclustering, heuristic algorithms, microarray data, genomic knowledge.

1. INTRODUCTION

A DNA Microarray is a glass slide covered with a chemical product and DNA samples containing thousands of genes. By placing this glass slide under a scanner, we obtain an image in which colored dots represent the expression level of genes under experimental conditions [1]. This process can be summarized by Figure 1. Simultaneous clustering of rows (genes) and columns (conditions) of this matrix enables to identify subsets of genes that have similar behavior under subsets of conditions, so we say that they coexpress, but behave independently under other subsets of conditions. This type of clustering is called biclustering. Biclustering of microarray data can be helpful to discover coexpression of genes and, hence, uncover genomic knowledge such as gene networks or gene interactions. Biclustering is an NP-hard problem [3]. Consequently, heuristic algorithms are typically used to approximate this problem by finding suboptimal solutions. In this paper, we make a new survey on tools, biclusters validation and evaluation functions.

In this paper, we make a survey on biclustering of gene expression data. The rest of the paper is organized as follows: First, we introduce some definitions related to biclustering of microarray data. Then, we present in section 3 some evaluation functions and biclustering algorithms. Next, we show how to validate biclusters via biclustering tools on microarrays datasets. Finally, we present our conclusion.
2. BICLUSTERING OF MICROARRAY DATA

Let introduce some definitions related to a biclustering of microarray data [3].

Biclusters: Let $I = \{1,2,...,n\}$ be a set of indices of $n$ genes, $I = \{1,2,...,m\}$ be a set of indices of $m$ conditions and $M(I,J)$ be a data matrix associated with $I$ and $J$. A bicluster associated with the data matrix $M(I,J)$ is a couple $M(I',J')$ such that $I' \subseteq I$ and $J' \subseteq J$.

Types of biclusters: A bicluster can be one of the following cases:

- Bicluster with constant values on rows:
  
  \[ m_{ij} = c + a_i \]  
  \[ m_{ij} = c \ast a_i \]
  
  where $c$ is a constant and $a_i$ is the adjustment for the row $i$.

- Bicluster with constant values on columns:
  
  \[ m_{ij} = c + b_j \]  
  \[ m_{ij} = c \ast b_j \]
  
  where $b_j$ is the adjustment for the column $j$.

- Bicluster with coherent values: There are two types of biclusters with coherent values. Those with additive model and those with multiplicative model defined respectively by:

  Those with additive model:
  
  \[ m_{ij} = c + a_i + b_j \]  
  
  Those with multiplicative model:
And those with multiplicative model:

\[ m_{ij} = c \cdot a_i \cdot b_j \quad (2.6) \]

- Bicluster with coherent evolution: It is a bicluster where all the rows (resp. columns) induce a linear order across a subset of columns (resp. rows).

**Groups of biclusters:** A group of biclusters can be one of the following types [4]:

1. Single bicluster,
2. Exclusive rows and columns group of biclusters,
3. Non-overlapping group of biclusters with checkerboard structure,
4. Exclusive rows group of biclusters,
5. Exclusive columns group of biclusters,
6. Non-overlapping group of biclusters with tree structure,
7. Non-overlapping non-exclusive group of biclusters,
8. Overlapping group of biclusters with hierarchical structure,
9. Or, arbitrarily positioned overlapping group of biclusters.

We note also that a natural way to visualize a group of biclusters consists in assigning a different color to each bicluster and in reordering the rows and the columns of the data matrix so that we obtain a data matrix with colored blocks, where each block represents a bicluster. The biclustering problem can be formulated as follows: Given a data matrix \( M \), construct a group of biclusters \( B_{opt} \) associated with \( M \) such that:

\[
f(B) = \max_{B \in BC(M)} f(B) \quad (2.7)
\]

where \( f \) is an objective function measuring the quality, i.e., degree of coherence, of a group of biclusters and \( BC(M) \) is the set of all the possible groups of biclusters associated with \( M \). This problem is NP-hard [4,5].

### 3. EVALUATION FUNCTIONS

An evaluation function is an indicator of the performance of a biclustering algorithm. There are two main classes of evaluation functions: Intra-biclusters evaluation functions and inter-biclusters evaluation functions.

#### 3.1. Intra-biclusters evaluation functions

An intra-biclusters evaluation function is a function that measures the quality of a bicluster, i.e., it quantifies the coherence degree of a bicluster. There are several intra-biclusters evaluation functions.

- The \( E_{AVSS}(I', J') \) is defined as follows[6]:

\[
E_{AVSS}(I', J') = \frac{\sum_{i \in I'} \sum_{j \in J'} s_{ij}}{|I'| |J'|} \quad (3.1)
\]

where \( (I', J') \) is a bicluster, \( s_{ij} \) is a similarity measure among elements of the row \( i \) and the column \( j \) with others elements belonging to \( I' \) and \( J' \). It follows that a number of these functions are particular cases of the **AVerage Similarity Score (AVSS)**.

- The **Average Row Variance** (ARV) is defined as follows [7]:

\[
E_{ARV}(I', J') = \frac{\sum_{i \in I'} \sum_{j \in J'} (m_{ij} - m_{I'J'})^2}{|I'| |J'|} \quad (3.2)
\]
where \( m_{ij} \) is the average over the row \( i \). It follows that the biclusters that contain rows with large changes in their values for different columns are characterized by a large row variance. The ARV guarantees that a bicluster captures rows exhibiting coherent trends under some subset columns.

- The **Mean Squared Residue** (MSR) is defined as follows \([8]\):

\[
E_{\text{MSR}}(I', J') = \frac{\sum_{i \in I'} \sum_{j \in J'} (m_{ij} - m_{iJ} - m_{IJ} + m_{IJ'})^2}{|I'||J'|}
\]  

where \( m_{iJ'} \) is the average over the whole bicluster, \( m_{iJ} \) is the average over the column \( j \), \( m_{iJ'} \) is the average over the row \( i \). The \( E_{\text{MSR}} \) represents the variation associated with the interaction between the rows and the columns in the bicluster. It follows that a low (resp. high) \( E_{\text{MSR}} \) value, i.e., close to 0 (resp. higher than a fixed threshold \( d \)), indicates that the bicluster is strongly (resp. weakly) coherent. The \( E_{\text{MSR}} \) function is inadequate to assess certain types of biclusters. For example, the \( E_{\text{MSR}} \) function is good for biclusters of coherent values with additive model but not for coherent values with multiplicative model.

- The **Volume** (V) is defined as follows \([7]\):

\[
E_V(I', J') = |I'||J'|
\]  

This function enables to have the maximum-sized bicluster that does not exceed a certain coherence value expressed as a MSR score. \( E_v(I', J') \) finds the maximum-sized bicluster that does not exceed a certain coherence value \([9]\) expressed as a MSR score. Hence, discovered biclusters have a high \( E_v(I', J') \) maximized and lower \( E_{\text{MSR}} \) than a given threshold \( \delta \geq 0 \).

- The **Mean Square Error** (MSE) is defined as follows \([10]\):

\[
E_{\text{MSE}}(I, J) = \frac{\sum_{i \in I} \sum_{j \in J} (m_{ij} - m_{Ij} - m_{IJ} + m_{IJ'})^2}{|I||J|}
\]  

where \( m_{ij} \) is the average over the whole matrix, \( m_{ij} \) is the average over the column \( j \) of the whole matrix and \( m_{ij} \) is the average over the row \( i \). This function identifies constant biclusters.

- The **Average Correlation Value** (ACV) is defined as follows \([5, 11]\):

\[
E_{\text{ACV}}(I', J') = \max \left\{ \frac{\sum_{i \in I'} \sum_{j \in J'} |r_{ij}|-|I'|}{|I'||J'|-1}, \frac{\sum_{k \in J'} \sum_{l \in J'} |r_{kl}|-|J'|}{|I'||J'|-1} \right\}
\]  

where \( r_{ij}(i \neq j) \) (resp. \( r_{kl}(k \neq l) \)) is the Pearson’s correlation coefficient associated with the row indices \( i \) and \( j \) (resp. \( k \) and \( l \)) in the bicluster \((I', J') \) \([8]\). The values of \( E_{\text{ACV}} \) belong to \([0;1]\), hence, a high (resp. low) \( E_{\text{ACV}} \) value, i.e., close to 1 (resp. close to 0), indicates that the bicluster is strongly (resp. weakly) coherent. However, the performance of the \( E_{\text{ACV}} \) function decreases when noise exists in the data matrix \([5, 11]\).

- The **Average Spearman’s Rho** (ASR) is defined as follows \([2]\):

\[
E_{\text{ASR}}(I', J') = 2 \max \left\{ \frac{\sum_{i \in I'} \sum_{j \in J'} \rho_{ij}}{|I'||J'|-1}, \frac{\sum_{k \in J'} \sum_{l \in J'} \rho_{kl}}{|I'||J'|-1} \right\}
\]  

\([3.7]\)
where $\rho_{ij}(i \neq j)$ (resp. $\rho_{kl}(k \neq l)$) is the Spearman's rank correlation associated with the row indices $i$ and $j$ in the bicluster $(I',J')$ [12]. The values of the $E_{ASR}$ function belong also to $[-1,1]$, hence, a high (resp. low) $E_{ASR}$ value, i.e., close to 1 (resp. close to -1), indicates that the bicluster is strongly (resp. weakly) coherent. On the other hand, like Spearman’s rank correlation, the $E_{ASR}$ is less sensitive to the presence of noise in data [2]. There are other intra-biclusters evaluation function like the Average Correspondance Similarity Index (ACSI) [2].

### 3.2. Inter-biclusters evaluation functions

An inter-biclusters evaluation function is a function that measures the quality of a group of biclusters, i.e., it assesses the accuracy of an algorithm to recover true implanted biclusters in a data matrix. There are several inter-biclusters evaluation functions. In what follows, we present some of them:

Let $M_1$ and $M_2$ be two groups of biclusters defined as follows:

- $M_1 = \{B_1^{(1)}, B_2^{(1)}, \ldots, B_{K_1}^{(1)}\}$, where $B_i^{(1)} = (G_i^{(1)}, C_i^{(1)})$, $G_i$ and $C_i$ are respectively the $i^{th}$ gene and condition, $1 \leq i \leq K_1$; Set of true implanted biclusters in a data matrix $M$.
- $M_2 = \{B_1^{(2)}, B_2^{(2)}, \ldots, B_{K_2}^{(2)}\}$, where $B_j^{(2)} = (G_j^{(2)}, C_j^{(2)})$, $G_m$ and $C_m$ are respectively the $m^{th}$ gene and condition, $1 \leq m \leq K_2$; Set of the biclusters extracted by a biclustering algorithm.

#### • The $Prelie$ index is defined as follows:

$$I_{Prelie}(M_1, M_2) = \frac{1}{K_1} \sum_{i=1}^{K_1} \max_j S_{Prelie}(B_i^{(1)}, B_j^{(2)})$$

(3.8)

where $S_{Prelie}$ is based on the Jaccard index for two sets and defined as follows:

$$S_{Prelie}(B_i, B_j) = \frac{|G_i \cap G_j|}{|G_i \cup G_j|}$$

(3.9)

This index compares two solutions based on categorization of genes. However, it compares only genes sets.

#### • The $Liu$ and $Wang$ index is defined as follows:

$$I_{Liu&Wang}(M_1, M_2) = \frac{1}{K_1} \sum_{i=1}^{K_1} \max_j S_{Liu&Wang}(B_i^{(1)}, B_j^{(2)})$$

(3.10)

where

$$S_{Liu&Wang}(B_i, B_j) = \frac{|G_i \cap G_j| + |C_i \cap C_j|}{|G_i \cup G_j| + |C_i \cup C_j|}$$

(3.11)

It compares two solutions by considering both genes and conditions.

#### • The $wtJaccard$ index is defined as follows:

$$I_{wtJaccard}(M_1, M_2) = \frac{\sum_{i=1}^{K_1} |B_i^{(1)}| \cdot \max_j S_{Jaccard}(B_i^{(1)}, B_j^{(2)})}{\sum_{i=1}^{K_1} |B_i^{(1)}|}$$

(3.12)

where

$$S_{Jaccard}(B_i, B_j) = \frac{|G_i \cap B_j| + |G_i \cap G_j|}{|G_i| + |B_j| - |G_i \cap C_j|}$$

(3.13)
The Dice index is defined as follows:

\[ I_{\text{Dice}}(M_1, M_2) = \frac{1}{K_1} \sum_{i=1}^{K_1} \max_j S_{\text{Dice}}(B_i^{(1)}, B_j^{(2)}) \]  

(3.14)

where:

\[ S_{\text{Dice}}(B_i, B_j) = \frac{2 \cdot |C_i \cap C_j|}{|C_i| + |C_j|} \]  

(3.15)

which is proposed in [13] and called F-measure in biclustering cases to compute the overall relevance of two bicluster solutions.

The Santamaría index is defined as follows:

\[ I_{\text{Santamaría}}(M_1, M_2) = \frac{1}{K_1} \sum_{i=1}^{K_1} \max_j S_{\text{Dice}}(B_i^{(1)}, B_j^{(2)}) \]  

(3.16)

where:

\[ B_1 \text{ and } B_2 \text{ are two groups of biclusters and the pair } (I, J) \text{ represents the submatrix whose rows and columns are given by the set } I \text{ and } J, \text{ respectively.} \]

The Santamaría index is the most conservative index among above others indices and used for biclustering case [14, 13]. In fact, while the Prelic index compares only object sets and the LW index compares object sets and feature sets independently, the Santamaría index compares two solutions using pairs of genes and conditions.

For gene expression case, the Gene Match Score (GMS) function doesn’t take into account column match. It is given by:

\[ E_{\text{GMS}}(B_1, B_2) = \frac{1}{|B_1|} \sum_{(I, J) \in B_1, (I', J') \in B_2} \frac{|I \cap J|}{|I| \cup |J|} \]  

(3.17)

where \( B_1 \) and \( B_2 \) are two groups of biclusters and the pair \((I, J)\) represents the submatrix whose rows and columns are given by the set \( I \) and \( J \), respectively.

The Row and Column Match Scores (RCMS) assess the method’s accuracy to recover known biclusters and reveal true ones. Thereafter, more similar measures of match scores have been introduced [5, 15, 6]. For instance, the evaluation functions, herein called Row and Column Match Scores, \( E_{\text{RCMS1}} \) and \( E_{\text{RCMS2}} \), are proposed in [6] and [15], respectively and given by:

\[ E_{\text{RCMS1}}(B_1, B_2) = \frac{1}{|B_1|} \sum_{(I, J) \in B_1, (I', J') \in B_2} \frac{|I \cap J| + |J \cap I|}{|I| \cup |J|} \]  

(3.18)

\[ E_{\text{RCMS2}}(B_1, B_2) = \frac{1}{|B_1|} \sum_{(I, J) \in B_1, (I', J') \in B_2} \frac{|I \cap J| + |J \cap I|}{|I| \cup |J||} \]  

(3.19)

All these measures of match score are used to assess the accuracy of an algorithm to recover known biclusters and reveal true ones. Both \( E_{\text{RCMS1}} \) and \( E_{\text{RCMS2}} \) have the advantage of reflecting, simultaneously, the match of the row and column dimensions between biclusters as opposed to \( E_{\text{GMS}} \) that doesn’t take into account column match. They vary between 0 and 1 (the higher the better the accuracy). Let \( B_{\text{opt}} \) denote the set of true implanted biclusters in the data matrix \( M \) and \( B \) the set of the output biclusters of a biclustering algorithm. Thus, \( E_{\text{GMS}}(B_{\text{opt}}, B) \) and \( E_{\text{RCMS1}}(B_{\text{opt}}, B) \)
express how well each of the true biclusters are detected by the algorithm under consideration. 

\[ E_{RCMS2}(B_X,B_Y) \], where \( B_X \) (resp. \( B_Y \)) denotes the set of biclusters detected by the algorithm \( X \) (resp. Algorithm \( Y \)), has the particularity to allow the quantification of how well each bicluster identified by the algorithm \( X \) is contained into some bicluster detected by the algorithm \( Y \).

5. **BICLUSTERING VALIDATION**

There are two types of biclusters validation:

(i) **Statistical validation**: It is used to validate synthetical data

(ii) **Biological validation**: It is used to validate biological data

5.1. **Statistical validation**

Statistical validation can be made by adopting one or many of the following indices:

- **Separation**: It reflects how well the biclusters are separated from each other. Separation between two biclusters \( A \) and \( B \) is defined as follows [62]:

\[
Sep(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}
\]  

(5.1)

- **Coverage**: We distinguish three types of coverage, matrix coverage, genes coverage and conditions coverage:

\[
\text{Matrix coverage} = \frac{\text{Number of the cells covered by the extracted biclusters}}{\text{Total number of cells in the matrix}}
\]  

(5.2)

\[
\text{Genes coverage} = \frac{\text{Number of the genes covered by the extracted biclusters}}{\text{Total number of genes in the matrix}}
\]  

(5.3)

\[
\text{Conditions coverage} = \frac{\text{Number of the conditions covered by the extracted biclusters}}{\text{Total number of conditions in the matrix}}
\]  

(5.4)

- **Compactness**: It assesses cluster homogeneity, with intra-cluster variance[63].

- **Connectedness**: It assesses how well a given partitioning groups data items together with their nearest neighbours in the data space [63].

- **Coherence**: It expresses how well a bicluster is fitted to a specified model. The coherence is computed thanks to compactness and connectedness.

- **Significance**: It is computed thanks to \( p \)-value. Let \( B \) be a bicluster, \( p \)-value is defined as follows [15]:

\[
p\text{-value}_B = 1 - \phi\left( \frac{|1_B|/|B| - p}{\sqrt{|1_B|/|B|}} \right)
\]  

(5.5)

where \( f \) is the standard normal distribution function, \(|1_B|\) is the number of 1’s in the bicluster \( B \) and \( p = k(|1_B||J|) \) of 1’s in \( M(I,J) \), \( k \) is the number of 1’s in the binary matrix \( M_b \). A bicluster \( B \) is considered as potentially significant at a level of significance \( \alpha \) if \( p\text{-value}_B < \alpha \).

5.2. **Biological validation**

Biological validation can qualitatively evaluate the capacity of an algorithm to extract meaningful biclusters from a biological point of view. To assess biologically biclusters, we can use Gene Ontology (GO) annotation [64]. In GO, genes are assigned to three structured, controlled vocabularies, called ontologies: biological process, cellular components and molecular functions.
The GO Consortium (GOC)[64] [65] is involved in the development and application of the GO. In what follows, we briefly report some R tools related to GOC [66, 67]:

- AnnotationDbi: It provides user interface and database connection code for annotation data packages using SQLite data storage.
- FunClust: It is a functional profiling and analysis of microarray expression data based on GO & KEGG.
- GExMap: It is an intuitive visual tool to perform a GO and to test to unveil genomic clusters, graphical interpretations and statistical results in pdf files.
- GO.db annotation: It provides detailed information about the latest version of the GOs and it is updated biannually.
- GOsummaries: It shows GO enrichment results in the context of experimental data.
- GOSTats: It determines which GOs found in gene lists are statistically over/under-represented.
- goTools: It compares the GOs represented by the genes in the three gene lists (biological process, molecular function and cellular component).
- topGO: It provides tools for testing GO terms while accounting for the topology of the GO graph. Different test statistics and different methods for eliminating local similarities and dependencies between GO terms can be implemented and applied.

6. TOOLS

There are also many R microarray biclustering tools. Table 1. presents a few examples on tools and here are some examples [68]:

- arules: It is a mining association rules and frequent itemsets. It provides the infrastructure for representing, manipulating and analyzing transaction data and patterns. It also provides interfaces of the association mining algorithms Apriori and Eclat [69].
- lattice: It is a high-level data visualization system with an emphasis on multivariate data. It supports the creation of trellis graphs to display multivariate relationship between variables, conditioned on one or more other variables via R graphics [69].
- rootSolve: It finds the root of nonlinear functions, solves the steady-state conditions for uni/multi-component and equilibrium analysis of ordinary differential equations via a dynamically running; like gradient and Jacobian matrices, non-linear equations by the Newton-Raphson algorithm.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Biclustering algorithms</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lattice</td>
<td>Galois lattice</td>
<td>[17]</td>
</tr>
<tr>
<td>arules</td>
<td>rules</td>
<td>[71]</td>
</tr>
<tr>
<td>rootSolve, pracma</td>
<td>Newton Raphson</td>
<td>[71]</td>
</tr>
<tr>
<td>blockcluster</td>
<td>Coclustering</td>
<td>[17]</td>
</tr>
<tr>
<td>biclustGUI</td>
<td>CC, Plaid, BiMAX, xMOTIFs, xQuest, Spectral, FABIA, ISA</td>
<td>[20]</td>
</tr>
<tr>
<td>biclust</td>
<td>Plaid, BiMAX, xMOTIFs, xQuest, Spectral</td>
<td>[17]</td>
</tr>
<tr>
<td>BcDiag</td>
<td>biclust, eisa, isa2</td>
<td>[17]</td>
</tr>
<tr>
<td>Algorithm</td>
<td>Description</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td>FABIA, FABIAs, FABIAp,</td>
<td>FABIA</td>
<td>[40]</td>
</tr>
<tr>
<td>NMF</td>
<td>NMF</td>
<td>[70]</td>
</tr>
<tr>
<td>s4vd</td>
<td>s4vd</td>
<td>[26]</td>
</tr>
<tr>
<td>qubic</td>
<td>Rqubic</td>
<td>[38]</td>
</tr>
<tr>
<td>eisa, isa2</td>
<td>ISA</td>
<td>[17]</td>
</tr>
<tr>
<td>BicARE</td>
<td>FLOC</td>
<td>[72]</td>
</tr>
<tr>
<td>ThreeWayPlaid</td>
<td>Plaid for three-dimensional data</td>
<td>[46]</td>
</tr>
<tr>
<td>iBBigs</td>
<td>iBBiG</td>
<td>[44]</td>
</tr>
<tr>
<td>Superbiclust</td>
<td>Ensemble Biclustering</td>
<td>[73, 41]</td>
</tr>
<tr>
<td>HSSVD</td>
<td>HSSVD</td>
<td>[46]</td>
</tr>
<tr>
<td>FacPad</td>
<td>Factor analysis for pathways</td>
<td>[45]</td>
</tr>
<tr>
<td>FastICA</td>
<td>Fast independent component analysis</td>
<td>[74]</td>
</tr>
<tr>
<td>CMonkey</td>
<td>cMonkey</td>
<td>[75]</td>
</tr>
</tbody>
</table>

- pracma: It rootfinds through Newton-Raphson or Secant algorithms [70] via using functions from numerical analysis and linear algebra, numerical optimization, differential equations and some special functions. It also uses Matlab function names where appropriate to simplify porting.
- BicARE: It is based on the FLOC algorithm [23] for biclustering analysis and results exploration.
- BcDiag: It provides methods for data preprocessing, visualization, and statistical validation to diagnostic and visualize in two-dimensional data based on two way anova [40] and median polish residual plots for biclust package output obtained from biclust, eisa-isa2 and fabia packages [17][40]. In addition, the biclust package can be used via biclustGUI, i.e. R commander plug in.
- blockcluster: It performs coclustering of binary, contingency and categorical datasets with utility functions to visualize the coclustered data. It contains a function cocluster which performs coclustering and returns object of appropriate class. It also contains cocluststrategy function which returns an object of class strategy.
- rqubic: It represents an implementation of the QUBIC algorithm [38] for the qualitative biclustering with gene expression data.
- HSSVD: It discovers and compares subgroups of patients and genes which simultaneously display unusual levels of variability. It detects both mean and variance biclusters by testing the biclustering with heterogeneous variance.
- iBBig: It optimizes applying binary data analysis to meta-gene set analysis of gene expression datasets. It extracts iteratively groups of phenotypes from multiple studies that are associated with similar gene sets without requiring prior knowledge of the number or scale of clusters and allows discovery of clusters with diverse sizes.
- NMF: It provides a framework to perform Non-negative Matrix Factorization (NMF). It implements a set of already published algorithms and seeding methods, and provides a framework to test, develop and plug new/custom algorithms. It performs parallel computations on multicore machines.
• s4vd: It performs a biclustering via sparse singular value decomposition (svd) with a nested stability selection. The result is an biclust object and thus all methods of the biclust package can be applied.
• superbiclust: It generates as a result a number of (or super) biclusters with none or low overlap from a bicluster set, i.e. ensemble biclustering [42], with respect to the initialization parameters for a given bicluster solution. The set of robust biclusters is based on the similarity of its elements, i.e. overlap, and on the hierarchical tree obtained via cut-off points.

8. CONCLUSION

The biclustering of microarray data has been the subject of a large research. No one of the existing biclustering algorithms is perfect. The construction of biologically significant groups of biclusters for large microarray data is still a problem that requires a continuous work. Biological validation of biclusters of microarray data is one of the most important open issues. So far, there are no general guidelines in the literature on how to validate biologically extracted biclusters.

REFERENCES


Jiong Yang and al. Enhanced biclustering on expression data.


Tze-Haw Huang ; XingXing Song ; Mao Lin Huang. Optimized data acquisition by time series clustering in opc. IEEE., 2011.


Mohamed Nadif and Gérard Govaert. Block clustering via the block gem and two-way em algorithms. In AICCSA’05, pages 1–1, 2005.


Eleni Mina. Applying biclustering to understand the molecular basis of phenotypic diversity. Phd. Utrecht University Faculty of Science Department of Information and Computing Sciences, 2011.


Haifa BenSaber. Classification non supervisée des donnees des puces a ADN", ESSTT. 2010.


Authors

Professor ELOUMI Mourad : Full Professor in Computer Science Head of the BioInformatics Group (BIG) of The Laboratory of Technologies of Information and Communication, and Electrical Engineering (LaTICE), National High School of Engineers of Tunis (ENSIT), University of Tunis, Tunisia, and Professor at the Faculty of Economic Sciences and Management of Tunis (FSEG), University of Tunis El Manar, Tunisia.

Mrs BEN SABER Haifa : Phd student on the BioInformatics Group (BIG) of The Laboratory of Technologies of Information and Communication, and Electrical Engineering (LaTICE), National High School of Engineers of Tunis (ENSIT), University of Tunis, Tunisia, and Assistant at the Time Université, Tunisia.