UNIVERSITY MUSTAPHA STAMBOULI OF MASCARA

Faculty Of Sciences Exactes

Top-k Formal Concepts for identifying Positively and Negatively Correlated Biclusters

Amina HOUARI and Sadok BEN YAHIA

E-mail: amina.houari@univ-mascara.dz

MEDI'2021, June 21, 2021

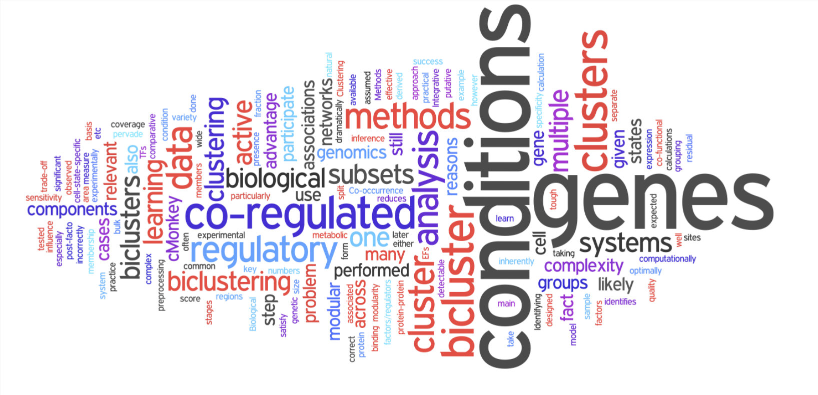
Plan

Outline

Introduction

Context of the research

The Increasing Challenge of Microarray Data



What is Biclustering ?

How to identify genes with similar behavior with respect to different conditions?

Why Biclustering ?

- Key to determine function of genes.
- Key to determine classification of conditions.

Biclustering

- Biclustering identifies subsets of genes and subsets of experimental conditions that share similar expression patterns.
- Similar concepts: subspace clustering, coclustering, bidimentional clustering, two-mode clustering.

Problem formulation

Let m_{ij} be the expression level of the $i - th$ gene in the $j - th$ condition

Problem formulation

Let m_{ij} be the expression level of the $i-th$ gene in the $j-th$ condition

Bicluster

A bicluster is a subset of a data matrix $M(I, J)$, I ={1, . . . ,n} and J={1, . . . ,m}

Problem formulation

Let m_{ij} be the expression level of the $i-th$ gene in the $j-th$ condition

Bicluster

A bicluster is a subset of a data matrix $M(I, J)$, I ={1, . . . ,n} and J={1, . . . ,m}

Bicluster

A bicluster is a pair (A,B) where:

- A is a subset of genes,$A \subset I$
- B is a subset of conditions, $B \subset J$

Contribution

(-)

A majority of existing biclustering algorithms for microarrays data focus only on extracting biclusters with positive correlations of genes.

Challenge

Recently, biological studies turned to a trend focusing on the notion of negative correlations [Zhao et al., 2008, Nepomuceno et al., 2015, Odibat and Reddy, 2014].

Biclustering Gene Expression Data

- Biclusters of positive correlations.

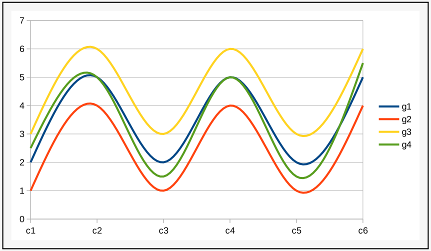


Figure : Examples of positive correlations.

Biclustering Gene Expression Data

- Biclusters of negative correlations [Zhao et al., 2008, Nepomuceno et al., 2015].

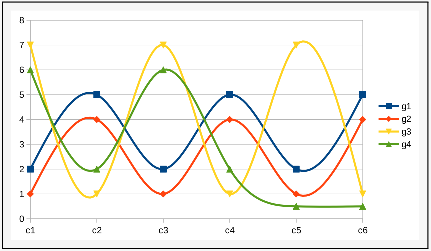


Figure : Examples of negative correlations.

Contibution

(+)

Formal Concept Analysis

Contibution

(+)

Formal Concept Analysis

(+)

To extract :

- Biclusters of positive correlations.
- Biclusters of negative correlations.

Outline

Formal Context

A binary table as a formal context

A triple $\mathcal{K}$=($\mathcal{O}$,$\mathcal{I}$,$\mathcal{R}$), where:

- $\mathcal{O}$: A set of objets : genes,
- $\mathcal{I}$: A set of attributes : Conditions,
- $\mathcal{R} \subseteq \mathcal{O} \times \mathcal{I}$ a binary relation $(o, i) \in \mathcal{R}$, shows which objects have which attributes.

Formal Context

Where:

1. $\mathcal{O}$= {1,2,3,4,5}
2. $\mathcal{I}$= {A,B,C,D,E}
3. $r1$: {(1),(A,C,D)}

	A	B	C	D	E
1	×		×	×	
2		×	×		×
3	×	×	×		×
4		×			×
5	×	×	×		×

Table : Example of a formal context

A maximal rectangle as a formal concept

$$\{3,5\}' = \{A,B,C,E\}$$
$$\{A,B,C,E\}' = \{3,5\}$$

$(\{3,5\},\{A,B,C,E\})$ is a **Formal Concept**

	A	B	C	D	E
1	1	0	1	1	0
2	0	1	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1
5	1	1	1	0	1

A maximal rectangle as a formal concept

A Galois connection to characterize formal concept

$A^{'} = \{o \in O \mid \forall\, g \in \mathcal{A},\ (g,o) \in \mathcal{R}\}$
$B^{'} = \{g \in G \mid \forall\, o \in \mathcal{B},\ (g,o) \in \mathcal{R}\}$

$\{3,5\}^{'} = \{A, B, C, E\}$
$\{A, B, C, E\}^{'} = \{3,5\}$

$(\{3,5\}, \{A, B, C, E\})$ is a **Formal Concept**

	A	B	C	D	E
1	1	0	1	1	0
2	0	1	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1
5	1	1	1	0	1

A maximal rectangle as a formal concept

A Galois connection to characterize formal concept

$A^{'} = \{o \in O \mid \forall\ g \in \mathcal{A},\ (g, o) \in \mathcal{R}\}$
$B^{'} = \{g \in G \mid \forall\ o \in \mathcal{B},\ (g, o) \in \mathcal{R}\}$

(A, B) is a formal concept with **extent** $A^{'} = B$ and **intent** $A = B^{'}$

$\{3, 5\}^{'} = \{A, B, C, E\}$
$\{A, B, C, E\}^{'} = \{3, 5\}$

$(\{3, 5\}, \{A, B, C, E\})$ is a **Formal Concept**

	A	B	C	D	E
1	1	0	1	1	0
2	0	1	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1
5	1	1	1	0	1

FCA-based Biclustering

(+)

FCA as a kind of biclustering for binary data. It provides pattern (bicluster) extraction from a binary relation, namely, a formal concept.

FCA-based Biclustering

(+)

FCA as a kind of biclustering for binary data. It provides pattern (bicluster) extraction from a binary relation, namely, a formal concept.

$FC = (A, B)$ is a concept if:
A : is an extent: objects share same attributes.
B : is an intent: attributes shared by the set of objects (extent).

FCA-based Biclustering

(+)

FCA as a kind of biclustering for binary data. It provides pattern (bicluster) extraction from a binary relation, namely, a formal concept.

$FC = (A, B)$ is a concept if:
A : is an extent: objects share same attributes.
B : is an intent: attributes shared by the set of objects (extent).

(+)

In its gene expression data applications:
The concept's **extent** represent maximal sets of **genes** related to a maximal set of **samples** (concept's **intent**).

Outline

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Strong suits of Top-BicMiner

- A new discretization method for microarray data.
- Extraction of biclusters with positive and negative correlations using FCA.

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Top-BicMiner: Principal

1. **Phase 1: "The discretization phase"**
2. **Phase 2: "The mining Phase"**
3. **Phase 3: "The filtering Phase"**
4. **Phase 4: "Extracting positively / negatively-correlated genes"**

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Top-BicMiner: Principal

1. **Phase 1: "The discretization phase"**
   - Discretize the original microarray data into a behavior data matrix (behavior matrix).
   - Discretize the behavior data matrix into two binary data matrices.
2. **Phase 2: "The mining Phase"**
3. **Phase 3: "The filtering Phase"**
4. **Phase 4: "Extracting positively / negatively-correlated genes"**

Extracting Biclusters of positive and negative correlations
Top-BicMiner algorithm

Top-BicMiner: Principal

1. **Phase 1: "The discretization phase"**
2. **Phase 2: "The mining Phase"**
   - Extracting formal concepts from the two binary contexts.
3. **Phase 3: "The filtering Phase"**
4. **Phase 4: "Extracting positively / negatively-correlated genes"**

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Top-BicMiner: Principal

1. **Phase 1: "The discretization phase"**
2. **Phase 2: "The mining Phase"**
3. **Phase 3: "The filtering Phase"**
   - The resulting biclusters are filtered using the TOPSIS multi-criteria (coupling, cohesion, stability, separation and distance. We have to **maximize**: stability, cohesion and separation. And **minimize:** coupling and distance.)
4. **Phase 4: "Extracting positively / negatively-correlated genes"**

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Example (Pre-processing)

	c_1	c_2	c_3	c_4	c_5
g_1	4	5	3	6	1
g_2	8	10	6	12	2
g_3	3	3	3	3	3
g_4	7	1	9	0	8
g_5	14	2	18	0	16

Table : Example of gene expression matrix (M_1).

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
g_1	1	-1	1	-1	-1	1	-1	1	-1	-1
g_2	1	-1	1	-1	-1	1	-1	1	-1	-1
g_3	0	0	0	0	0	0	0	0	0	0
g_4	-1	1	-1	1	1	-1	1	-1	-1	1
g_5	-1	1	-1	1	1	-1	1	-1	-1	1

Table : 3-state data matrix (M_2).

Extracting Biclusters of positive and negative correlations
Top-BicMiner algorithm

Example (Pre-processing)

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
$g1$	1	0	1	0	0	1	0	1	0	0
$g2$	1	0	1	0	0	1	0	1	0	0
$g3$	0	0	0	0	0	0	0	0	0	0
$g4$	0	1	0	1	1	0	1	0	0	1
$g5$	0	1	0	1	1	0	1	0	0	1

Table : $\mathcal{M}3^{+}$.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
g_1	1	-1	1	-1	-1	1	-1	1	-1	-1
g_2	1	-1	1	-1	-1	1	-1	1	-1	-1
g_3	0	0	0	0	0	0	0	0	0	0
g_4	-1	1	-1	1	1	-1	1	-1	-1	1
g_5	-1	1	-1	1	1	-1	1	-1	-1	1

Table : 3-state data matrix (M_2).

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
$g1$	0	1	0	1	1	0	1	0	1	1
$g2$	0	1	0	1	1	0	1	0	1	1
$g3$	0	0	0	0	0	0	0	0	0	0
$g4$	1	0	1	0	0	1	0	1	1	0
$g5$	1	0	1	0	0	1	0	1	1	0

Table : $\mathcal{M}3^{-}$.

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Example (Pre-processing)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
g_1	1	-1	1	-1	-1	1	-1	1	-1	-1
g_2	1	-1	1	-1	-1	1	-1	1	-1	-1
g_3	0	0	0	0	0	0	0	0	0	0
g_4	-1	1	-1	1	1	-1	1	-1	-1	1
g_5	-1	1	-1	1	1	-1	1	-1	-1	1

Table : 3-state data matrix (M_2).

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
$g1$	1	0	1	0	0	1	0	1	0	0
$g2$	1	0	1	0	0	1	0	1	0	0
$g3$	0	0	0	0	0	0	0	0	0	0
$g4$	0	1	0	1	1	0	1	0	0	1
$g5$	0	1	0	1	1	0	1	0	0	1

Table : $\mathcal{M}3^+$.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
$g1$	0	1	0	1	1	0	1	0	1	1
$g2$	0	1	0	1	1	0	1	0	1	1
$g3$	0	0	0	0	0	0	0	0	0	0
$g4$	1	0	1	0	0	1	0	1	1	0
$g5$	1	0	1	0	0	1	0	1	1	0

Table : $\mathcal{M}3^-$.

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Example (Pre-processing)

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
g_1	1	-1	1	-1	-1	1	-1	1	-1	-1
g_2	1	-1	1	-1	-1	1	-1	1	-1	-1
g_3	0	0	0	0	0	0	0	0	0	0
g_4	-1	1	-1	1	1	-1	1	-1	-1	1
g_5	-1	1	-1	1	1	-1	1	-1	-1	1

Table : 3-state data matrix (M_2).

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
$g1$	1	0	1	0	0	1	0	1	0	0
$g2$	1	0	1	0	0	1	0	1	0	0
$g3$	0	0	0	0	0	0	0	0	0	0
$g4$	0	1	0	1	1	0	1	0	0	1
$g5$	0	1	0	1	1	0	1	0	0	1

Table : $\mathcal{M}3^+$.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
$g1$	0	1	0	1	1	0	1	0	1	1
$g2$	0	1	0	1	1	0	1	0	1	1
$g3$	0	0	0	0	0	0	0	0	0	0
$g4$	1	0	1	0	0	1	0	1	1	0
$g5$	1	0	1	0	0	1	0	1	1	0

Table : $\mathcal{M}3^-$.

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Example (The Mining phase)

Extracting formal concepts from the two binary contexts obtained from the previous step.

Formal Concepts (FCs)					
$\mathcal{M}3^+$			$\mathcal{M}3^-$		
ID concept	extent	intent	ID concept	extent	intent
FC 1^+	g1, g2	C1, C3, C6, C8	FC 1^-	g4, g5	C1, C3, C6, C8, C9
FC 2^+	g4, g5	C2, C4, C5, C7, C10	FC 2^-	g1, g2	C2, C4, C5, C7, C9, C10
			FC 3^-	g1, g2, g4, g5	C9

Table : Extracted Formal concepts from the formal contexts.

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Example (The filtering phase)

A multi-criteria to be aggregated, namely, coupling, cohesion, stability, separation and distance. We have **to maximize** the following criteria: stability, cohesion and separation. In addition, the criteria **to minimize** are coupling and distance.

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Example (Negatively-correlated genes extraction phase)

Consider coherent formal concepts having an intersection size greater or equal to a given intersection threshold $\alpha 1$.

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Example (Negatively-correlated genes extraction phase)

Consider coherent formal concepts having an intersection size greater or equal to a given intersection threshold $\alpha 1$.

Suppose that $\alpha 1$ =70% and using our example we have:

Extracting Biclusters of positive and negative correlations
Top-BicMiner algorithm

Example (Negatively-correlated genes extraction phase)

Consider coherent formal concepts having an intersection size greater or equal to a given intersection threshold $\alpha 1$.

Suppose that $\alpha 1$ =70% and using our example we have:

$FC1^{+} \bigcap FC1^{-} = C1, C3, C6, C8;$
$FC1^{+} \bigcap FC2^{-} = \emptyset;$
$FC1^{+} \bigcap FC3^{-} = \emptyset;$
and
$FC2^{+} \bigcap FC1^{-} = \emptyset;$
$FC2^{+} \bigcap FC2^{-} = C2, C4, C5, C7, C10;$

$FC2^{+} \bigcap FC3^{-} = \emptyset.$

Extracting Biclusters of positive and negative correlations

Top-BicMiner algorithm

Example (Negatively-correlated genes extraction phase)

The biclusters become:
$Bic1 = ((g1, g2, g4, g5), (C1, C3, C6, C8))$ and
$Bic2 = ((g1, g2, g4, g5), (C2, C4, C5, C7, C10))$.

$maxbic(Bic1, Bic2) = ((g1, g2, g4, g5), (C1, C2, C3, C4, C5, C6, C7, C8, C10))$.

Outline

Experimenal Evaluation

Biclusters validation

The used datasets

- Yeast Cell-Cycle dataset [Tavazoie et al., 1999] (Nature Genetics).
- Human B-Cell Lymphoma dataset [Alizadeh et al.,2000]

Experimental Evaluation

Biclusters validation

The used datasets

- Yeast Cell-Cycle dataset [Tavazoie et al., 1999] (Nature Genetics).
- Human B-Cell Lymphoma dataset [Alizadeh et al.,2000]

Statistical significance

- **Coverage:** Total number of cells in a microarray data matrix covered by the obtained biclusters
- **P-value:** Probability that genes of a bicluster have common biological characteristics.

Experimenal Evaluation

Biclusters validation

The used datasets

- Yeast Cell-Cycle dataset [Tavazoie et al., 1999] (Nature Genetics).
- Human B-Cell Lymphoma dataset [Alizadeh et al.,2000]

Statistical significance

- **Coverage:**Total number of cells in a microarray data matrix covered by the obtained biclusters
- **P-value:** Probability that genes of a bicluster have common biological characteristics.

Biological significance

Measuring the quality of biclusters, by checking whether the genes of a bicluster have common biological characteristics.

Experimental Evaluation

Statistical significance

Coverage

Human B-cell Lymphoma			
Algorithms	**Total Coverage**	**Gene Coverage**	**Condition Coverage**
BiMine	8.93%	26.15%	100%
BicFinder	44.24%	55.89%	100%
CC	36.81%	91.58%	100%
Trimax	8.50%	46.32%	11.46%
NBF	73.75 %	100%	100%
TOP-BICMINER	**75.02 %**	**100%**	**100%**

Experimenal Evaluation
Statistical significance

Coverage

Yeast Cell-Cycle			
Algorithms	**Total Coverage**	**Gene Coverage**	**Condition Coverage**
BiMine	13.36%	32.84%	100%
BicFinder	55.43%	76.93%	100%
CC	**81.47%**	**97.12%**	100%
Trimax	15.32%	22.09%	70.59%
NBF	77.17 %	97.08%	100%
Top-BicMiner	79.08 %	96.22%	100%

.

Our algorithm is competitive with surveyed algorithms.

Experimenal Evaluation

Statistical significance

FuncAssociate: *P-Value (Yeast Cell-Cycle dataset)*

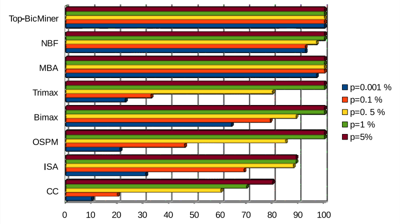


Figure : Proportions of Biclusters significantly enriched by GO annotations

.

The **Top-BicMiner** result shows that 100% of extracted biclusters are statistically significant with adjusted p-value $<0.001\%$.

Experimenal Evaluation

Biological significance

GoTermFinder: Biological significance

Yeast Cell-Cycle

	Bicluster 1	Bicluster 2
Biological process	cytoplasmic translation (53.1%, 7.80e-44) maturation of SSU-rRNA (32.1%, 6.30e-25) gene expression (96.3%, 5.20e-35)	amide biosynthetic process (59.7%, 5.03e-19) cleavage involved in rRNA processing (19.4%,1.14e-10) rRNA 5'-end processing (13.4%, 1.12e-08)
Molecular function	RNA binding (72.8%, 7.61e-30) heterocyclic compound binding (74.1%, 1.16e-12) RNA-dependent ATPase activity(6.2%, 7.39e-06)	structural constituent of ribosome (53.7%, 7.84e-35) binding (77.6%, 5.81e-05) organic cyclic compound binding (73.1%, 7.66e-10)
Cellular component	intracellular ribonucleoprotein complex (97.5%, 3.11e-74) 90S preribosome (29.6%, 7.94e-26) nucleolus (37.0%, 6.15e-17)	preribosome (47.8%, 2.22e-33) large ribosomal subunit (38.8%, 7.66e-20) cytosol (58.2%, 7.37e-12)

.

The results on this real-life data set show that our proposed algorithm can identify biclusters with a high biological relevance.

Outline

Conclusion

A summary of the contribution

- Biclustering is useful for bioinformatics.
- NP-Hard.
- New FCA-based biclustering algorithm for both: positive and negative correlations.
- A new discretization methods for microarray data.
- Experimental study shows that the proposed algorithms can identify biclusters with a high quality (statistical and biological criteria).

Future work

Perspectives...

- Apply our algorithms on other domains of application.
- Another possible experimentation to assess the performance of our algorithm on big data.

Future work

Perspectives...

- Apply our algorithms on other domains of application.
- Another possible experimentation to assess the performance of our algorithm on big data.

Thank You For Your Attention

Experimenal Evaluation : The used datasets

- **Yeast Cell-Cycle dataset:** a very popular dataset in the gene expression analysis community. It contains **2884** genes and **17** conditions.
- **Human B-cell lymphoma dataset:** contains **4026** genes and **96** conditions.

Nepomuceno, J. A., Troncoso, A., and Aguilar-Ruiz, J. S. (2015).
Scatter search-based identification of local patterns with positive and negative correlations in gene expression data.
Appl. Soft Comput., 35:637–651.

Odibat, O. and Reddy, C. K. (2014).
Efficient mining of discriminative co-clusters from gene expression data.
Knowl. Inf. Syst., 41(3):667–696.

Zhao, Y., Yu, J., Wang, G., Chen, L., Wang, B., and Yu, G. (2008).
Maximal subspace coregulated gene clustering.
Knowledge and Data Engineering, IEEE Transactions on, 20(1):83–98.