Unité de Recherche UTIC
Pôle Bioinformatique

Laboratoire d'Etudes et de Recherche
en Informatique d'Angers

Iterated Local Search for Biclustering of Microarray Data

Wassim AYADI[1, 2] , Mourad ELLOUMI[1] and Jin-Kao HAO[2]

[1] Unit of Technologies of Information and Communication, UTIC, Tunis, Tunisia

[2] Laboratoire d'Etude et de Recherche en Informatique d'Angers, LERIA, Angers, France

PRIB 2010

22-24 Septembre 2010

Nijmegen, The Netherlands

Outline

- Introduction
- Iterated Local Search
- The BILS algorithm
- Results & Conclusion

INTRODUCTION

- DNA microarray datasets are one of the most used data types in Bioinformatics
- It is generally represented by *n* × *m matrix M*
- *Each row* represents a gene and each column represents a data sample or condition

$$M = (m_{ij})_{n \times m}$$

where the value m_{ij} *is the expression of i-th gene in j-th* condition

- Analysing these datastes can give a valuable information on the biological relevance of genes and correlations between them

Why Biclustering?

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

- Given a data matrix *M(I, J)* , biclustering algorithms allow to extract *a* group of biclusters that maximize a given evaluation function
- The biclustering problem is NP-hard

Definitions

- Let m_{ij} *be the expression level of the i-th gene in the j-th* condition
- A *bicluster is a subset of a* data matrix *M* (I,J), I = {1, . . . ,n} and J= {1, . . . , m}
- A *bicluster* is a pair *(I',J') where:*
  - I' is a subset of genes, I' $\subseteq$ I
  - J' is a subset of conditions, J' $\subseteq$ I

Biclustering approaches

- The *systematic search approach:*
  - greedy algorithms
  - divide-and conquer algorithms
  - enumeration algorithms
- The *metaheuristic approach:*
  - neighbourhood-based algorithms
  - evolutionary algorithms

ITERATED LOCAL SEARCH

ILS – procedure Iterated Local Search

```
$s_0 \leftarrow$ GenerateInitialSolution
$s^* \leftarrow$ LocalSearch ($s_0$)
repeat □
    ◦ $s' \leftarrow$ Perturbation($s^*$)
    ◦ $s^{*\prime} \leftarrow$ LocalSearch($s'$)
    ◦ $s^* \leftarrow$ AcceptanceCriterion($s^*$, $s^{*\prime}$)
until termination condition met
```

THE BILS ALGORITHM

BILS – algorithm

Behavior Matrix M’

$$M'[i,l] = \begin{cases} 1 & \text{if } M[i,k] < M[i,q] \\ -1 & \text{if } M[i,k] > M[i,q] \\ 0 & \text{if } M[i,k] = M[i,q] \end{cases}$$

with $i \in [1..n]$, $l \in [1..J'']$, $k \in [1..m-1]$, $q \in [1..m]$ and $q > k+1$.

- The preprocessing step aims to highlight the trajectory patterns of genes
- Each column of *M ‘* represents the trajectory of genes between a pair of conditions in the data matrix *M*
- The whole *M’matrix provides* useful information for the identification of related biclusters
- Genes are considered to be in the same cluster if their trajectory patterns of expression levels are similar across a set of conditions

BILS – algorithm

Initial solution

- s_0: Initial bicluster obtained from the original data matrix using a fast greedy algorithm

BILS – algorithm

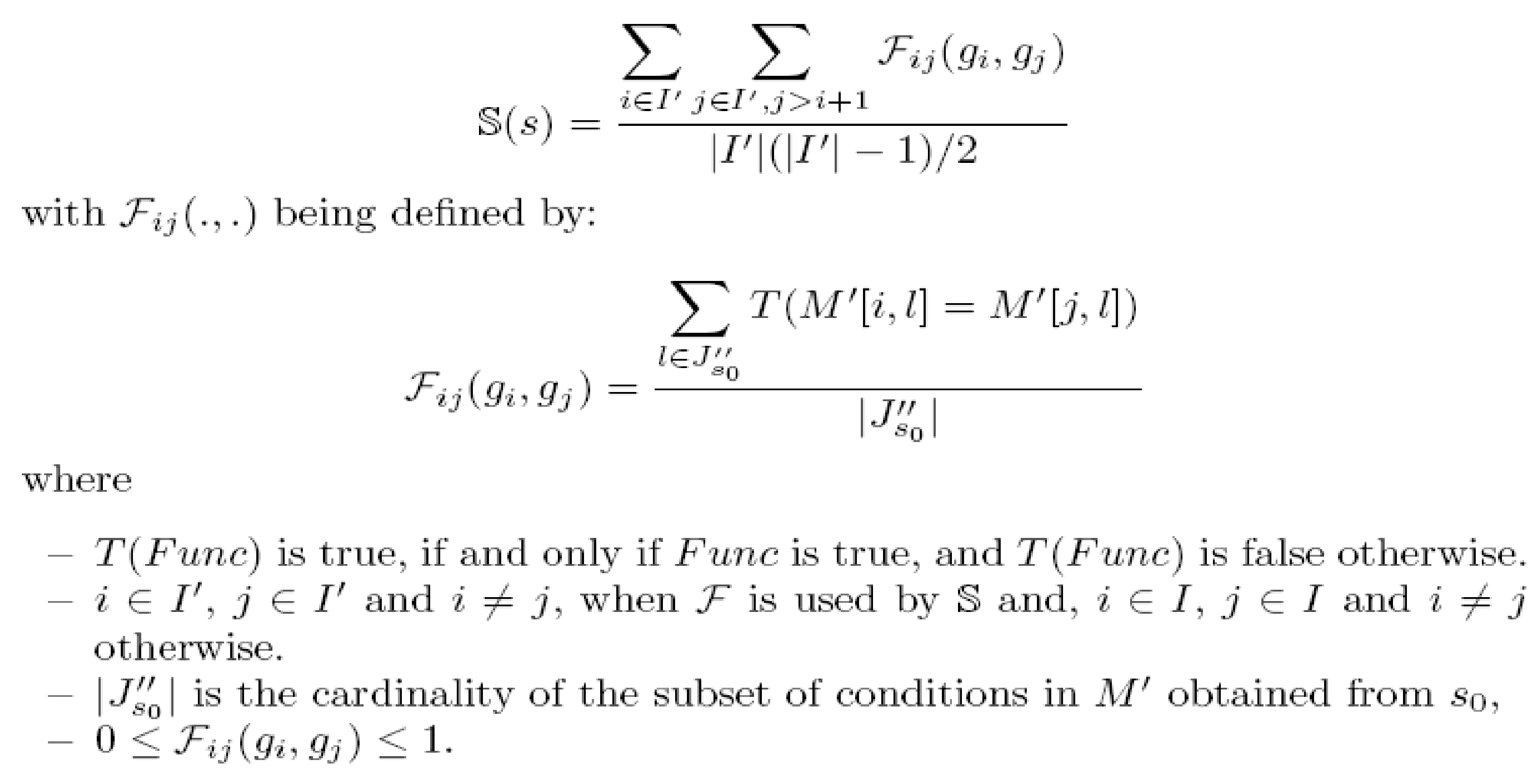
Local search

- s_0 is processed by removing one gene having a *bad* score and adding another gene or several having *good* scores
- Added genes do not belong initially to s_0
- Scores *are* computed by an evaluation function
- Each application of this dual drop/add operation generates a new bicluster (s*) from the current one

BILS – algorithm

Evaluation function

- The quality of a bicluster (s=(I’,J’)) is assessed by an evaluation function $\mathbb{S}$ given below:



$$\mathbb{S}(s) = \frac{\sum_{i \in I'} \sum_{j \in I', j > i+1} \mathcal{F}_{ij}(g_i, g_j)}{|I'|(|I'| - 1)/2}$$

with $\mathcal{F}_{ij}(.,.)$ being defined by:

$$\mathcal{F}_{ij}(g_i, g_j) = \frac{\sum_{l \in J''_{s_0}} T(M'[i,l] = M'[j,l])}{|J''_{s_0}|}$$

where

- $T(Func)$ is true, if and only if $Func$ is true, and $T(Func)$ is false otherwise.
- $i \in I'$, $j \in I'$ and $i \neq j$, when $\mathcal{F}$ is used by $\mathbb{S}$ and, $i \in I$, $j \in I$ and $i \neq j$ otherwise.
- $|J''_{s_0}|$ is the cardinality of the subset of conditions in M' obtained from s_0,
- $0 \leq \mathcal{F}_{ij}(g_i, g_j) \leq 1$.

BILS – algorithm

Perturbation operator

- The perturbation of the best solution (s*) is made to generate a new starting point (s') for the next round of the search
- This perturbation operator changes the best local optimum by:
  - deleting randomly 10% of genes of the best solution
  - adding 10% of genes among the *best* genes (which have *good* scores)
  - The added genes are not included previously in the best solution

BILS – algorithm

Stop condition

- The whole BILS algorithm stops when:
  - the best bicluster reaches a fixed threshold
  - the best solution found is do not change for a fixed number of perturbations

EXPERIMENTAL RESULTS

Yeast data set (Tavazoie et al., 99)

- 2884 genes and 17 conditions
- To obtain the initial solution (the input of BILS), we considered biclusters obtained using CC and OPSM algorithms
- Evaluation of biclusters using the two web tools:
  - Funcassociate for statistic evaluation
  - GoTermFinder is used for biological evaluation

RESULTS

Yeast data set

- Funcassociate: Statistical significance

Algorithm	P-value < 0.001
CC	10%
OPSM	22%
BILS	100%

→ BILS improves all biclusters of CC and OPSM

RESULTS

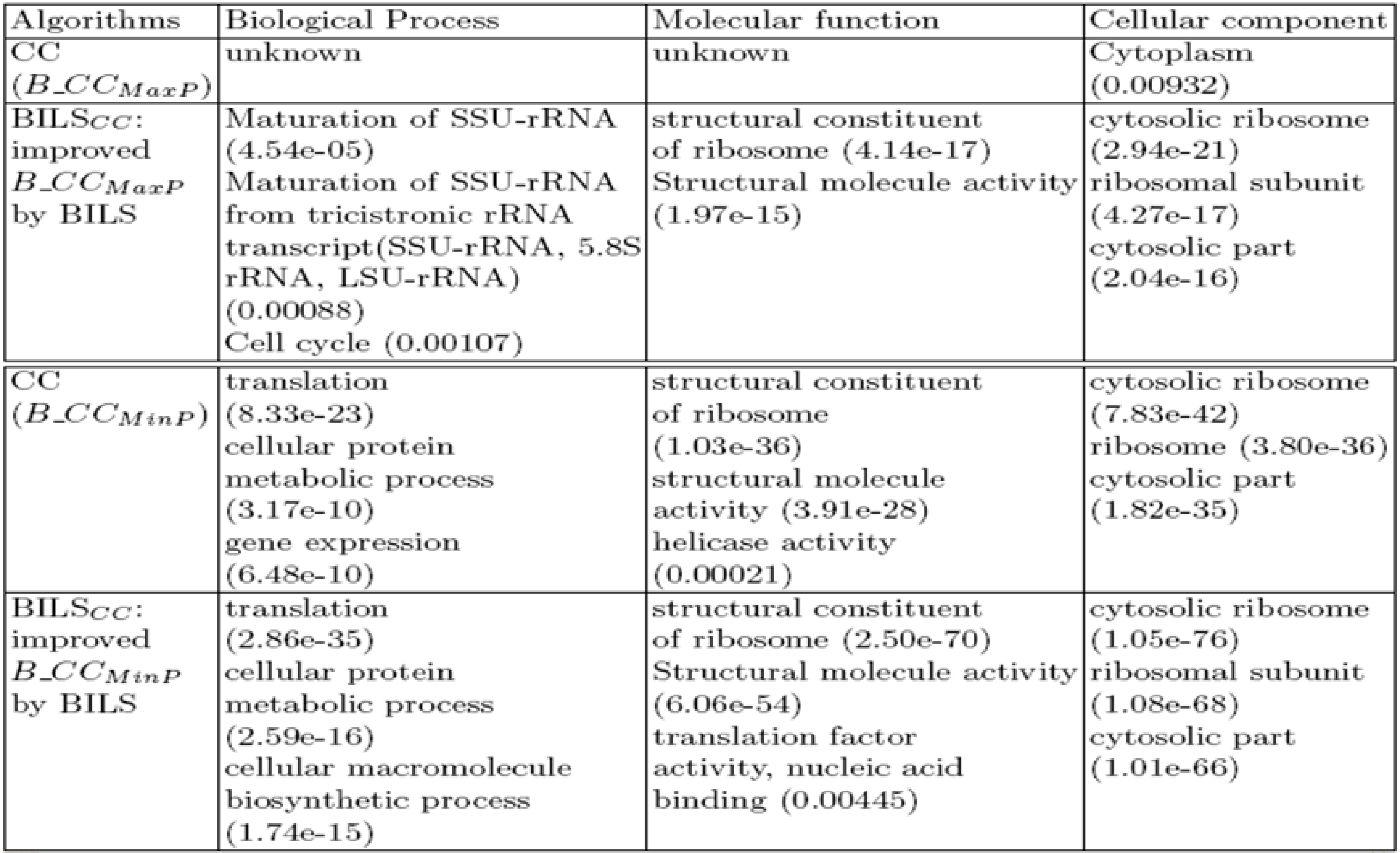
Yeast data set

Algorithms	Maximum p-value	Minimum p-value
CC	0.000010	4.096e-40
BILS	2.220e-17	2.860e-70
OPSM	0.0000012	1.587e-13
BILS	1.156e-10	4.865e-24

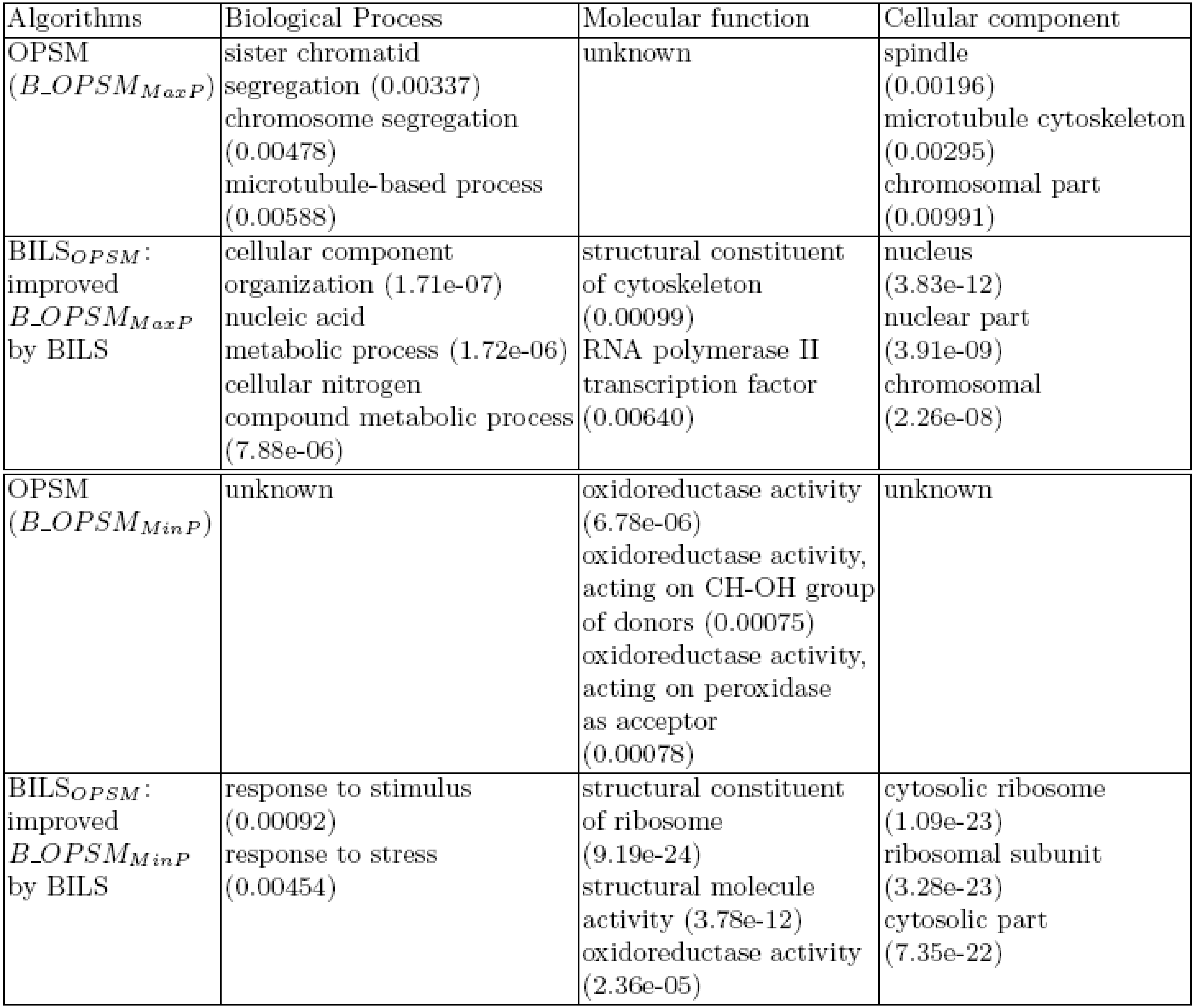
RESULTS

Yeast data set

- GoTermFinder: Biological significance



Algorithms	Biological Process	Molecular function	Cellular component
CC (B_CC_{MaxP})	unknown	unknown	Cytoplasm (0.00932)
BILS$_{CC}$: improved B_CC_{MaxP} by BILS	Maturation of SSU-rRNA (4.54e-05) Maturation of SSU-rRNA from tricistronic rRNA transcript(SSU-rRNA, 5.8S rRNA, LSU-rRNA) (0.00088) Cell cycle (0.00107)	structural constituent of ribosome (4.14e-17) Structural molecule activity (1.97e-15)	cytosolic ribosome (2.94e-21) ribosomal subunit (4.27e-17) cytosolic part (2.04e-16)
CC (B_CC_{MinP})	translation (8.33e-23) cellular protein metabolic process (3.17e-10) gene expression (6.48e-10)	structural constituent of ribosome (1.03e-36) structural molecule activity (3.91e-28) helicase activity (0.00021)	cytosolic ribosome (7.83e-42) ribosome (3.80e-36) cytosolic part (1.82e-35)
BILS$_{CC}$: improved B_CC_{MinP} by BILS	translation (2.86e-35) cellular protein metabolic process (2.59e-16) cellular macromolecule biosynthetic process (1.74e-15)	structural constituent of ribosome (2.50e-70) Structural molecule activity (6.06e-54) translation factor activity, nucleic acid binding (0.00445)	cytosolic ribosome (1.05e-76) ribosomal subunit (1.08e-68) cytosolic part (1.01e-66)



Algorithms	Biological Process	Molecular function	Cellular component
OPSM (B_OPSM_{MaxP})	sister chromatid segregation (0.00337) chromosome segregation (0.00478) microtubule-based process (0.00588)	unknown	spindle (0.00196) microtubule cytoskeleton (0.00295) chromosomal part (0.00991)
BILS$_{OPSM}$: improved B_OPSM_{MaxP} by BILS	cellular component organization (1.71e-07) nucleic acid metabolic process (1.72e-06) cellular nitrogen compound metabolic process (7.88e-06)	structural constituent of cytoskeleton (0.00099) RNA polymerase II transcription factor (0.00640)	nucleus (3.83e-12) nuclear part (3.91e-09) chromosomal (2.26e-08)
OPSM (B_OPSM_{MinP})	unknown	oxidoreductase activity (6.78e-06) oxidoreductase activity, acting on CH-OH group of donors (0.00075) oxidoreductase activity, acting on peroxidase as acceptor (0.00078)	unknown
BILS$_{OPSM}$: improved B_OPSM_{MinP} by BILS	response to stimulus (0.00092) response to stress (0.00454)	structural constituent of ribosome (9.19e-24) structural molecule activity (3.78e-12) oxidoreductase activity (2.36e-05)	cytosolic ribosome (1.09e-23) ribosomal subunit (3.28e-23) cytosolic part (7.35e-22)

CONCLUSIONS

CONCLUSIONS

- A new biclustering algorithm using Iterative Local Search (BILS) was proposed
- BILS employs a new evaluation function
- Experimental results show that the BILS algorithm can successfully improve all biclusters of CC and OPSM according to statistical and biological evaluation criteria

Future work

- Make a study on neighborhoods to introduce more biological knowledge in order to provide more effective guidance of the local search process
- BILS explores the space of biclusters by changing only the subset of genes of a bicluster without changing the conditions of the initial bicluster
- We aim to design similar strategies to optimize the subset of conditions of a bicluster or eventually to optimize simultaneously both the set of genes and conditions

Future work

- Another possible experimentation is to assess the algorithm on a synthetic data

THANK YOU