Contributions to transcriptomic data analysis and gene regulation network inference

ISSSB'2011 - NII Shonan meeting - November 13-17 2011

Céline Rouveirol

Overview

Inferring regulation networks from transcriptomic static data

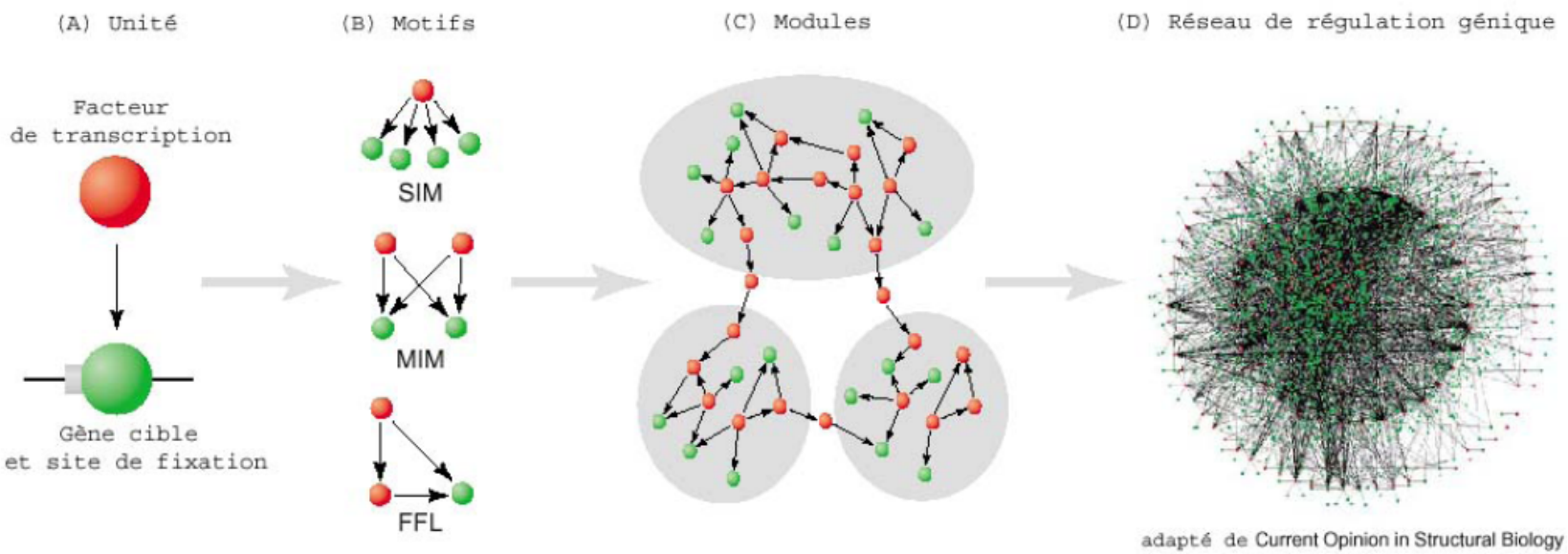
- Motivations and goals
- Methods
- Evaluation

Finding dense regions in binary contexts

- Motivation and goals
- Methods
- Experiments and results

Regulation network inference

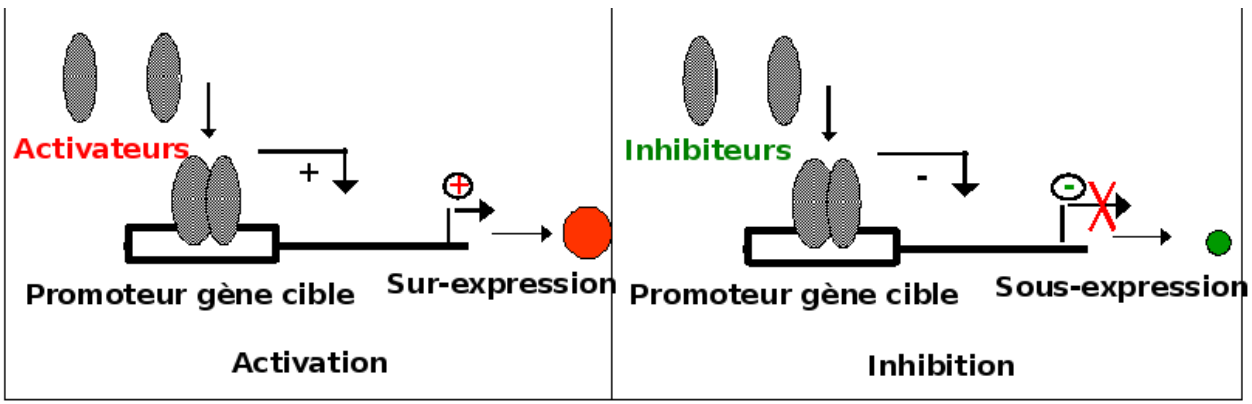
Buiding a regulation graph for a biological process



Transcriptional regulation

A **transcription factor is** :

- a proteine, **that binds to specific sequences of DNA** adjacent to the genes that they regulate
- controls the flow (**activates and/or respress**) this gene's transcription



Cahier des charges

- learning cooperative regulation relations from gene expression only
- no time series data available
- without any a priori assumption concerning the gene expression distribution
- local approach (one network / gene)

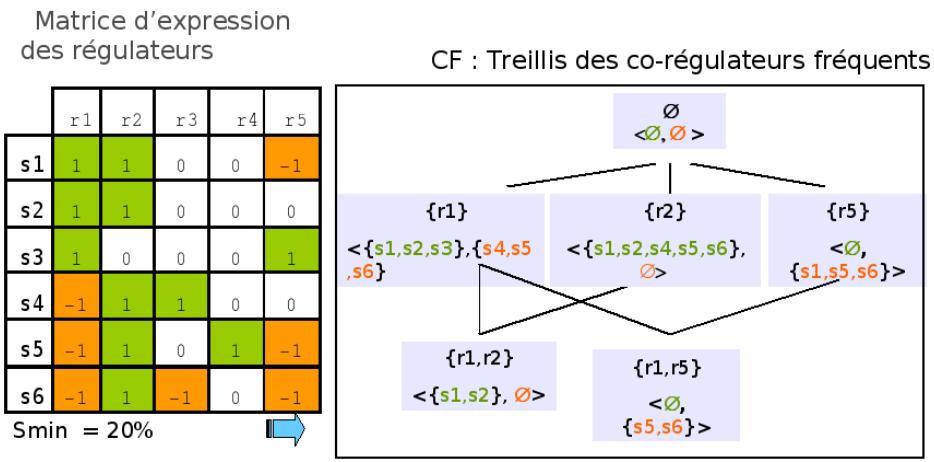
LICORN

LICORN (Elati et al., Bioinformatics 2007) follows these 3 steps:

1. Build a set of candidate co-regulators (predicate invention) for all genes ;
2. Build a set of candidate regulation networks for each target gene g ;
3. Select the best candidate network(s) for each gene g, and assign a significance score to this (these) network(s)

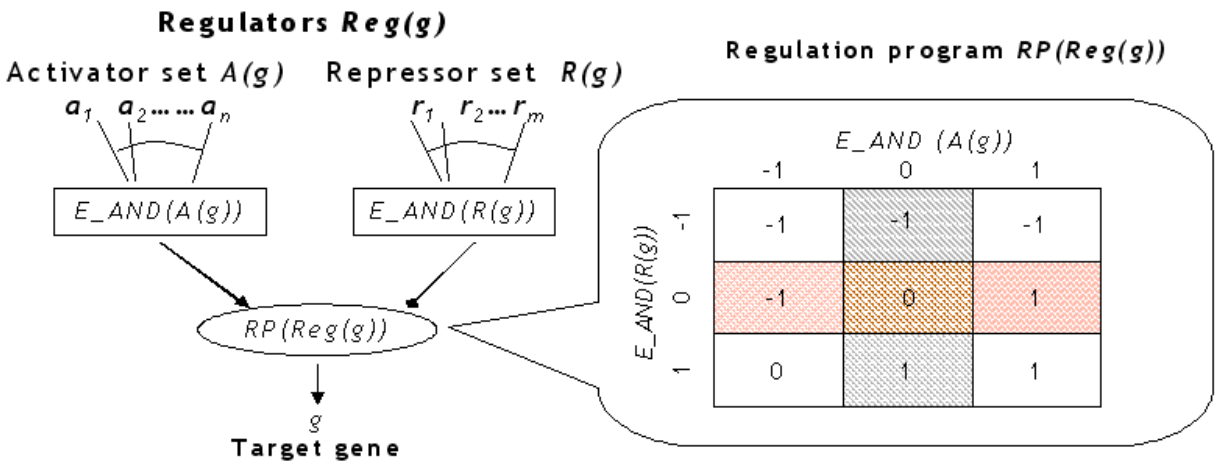
Co-regulators computation

- Goal: find all formulas of a language L that satisfy a constraint q on a dataset r, $Th(r, L, q)$.
- r : discrete matrix r of m observations described with n attributes $A = g_1, g_2, \ldots, g_n$ ($n >> m$)
  L : language describing itemsets on A (2^A).
  q : constraint of interest, *e.g.* frequency of a pattern p in r : p is frequent if $freq(p) \geq min_{supp}$
- Extension of Apriori (Agrawal et al., 1994) for computing frequent/closed itemsets from discrete data



Cooperative regulation model

- potentially, several cooperative activators/repressors
- AND-agregation for activators/repressors + deterministic function for computing target gene state given the aggregated states of its activators/repressors.



Generating candidate co-regulators for a target gene

Let C be a co-regulator and g be a target gene. $\mathcal{S}_x(C)$ and $\mathcal{S}_y(g)$ denote their support for the values $x, y \in \{-1, 1\}$.

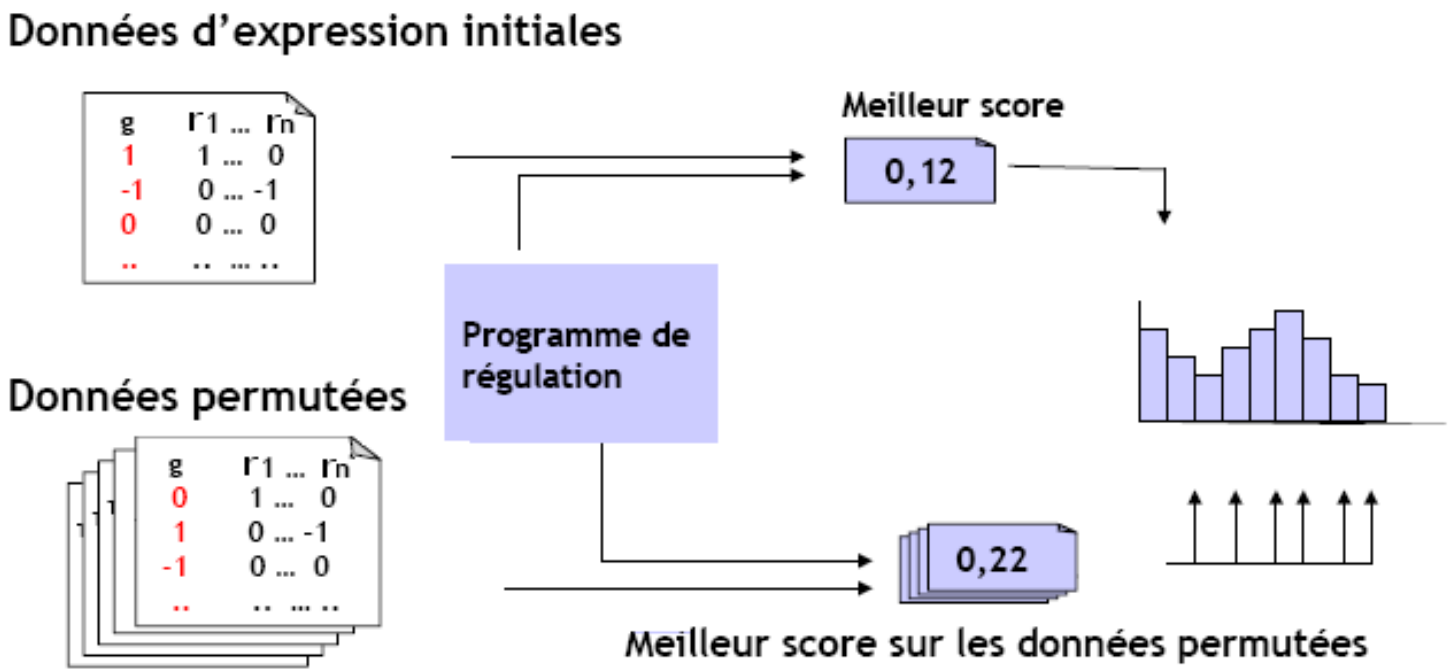
Definition (Overlap constraint)

C in state x *co-varies* with g in state y, denoted $\mathbf{cov}(\mathcal{S}_x(C), \mathcal{S}_y(g))$ if and only if $\frac{|\mathcal{S}_y(g) \cap \mathcal{S}_x(C)|}{|\mathcal{S}_y(g)|} \geq min_{overlap}$, a user-defined minimum overlap threshold.

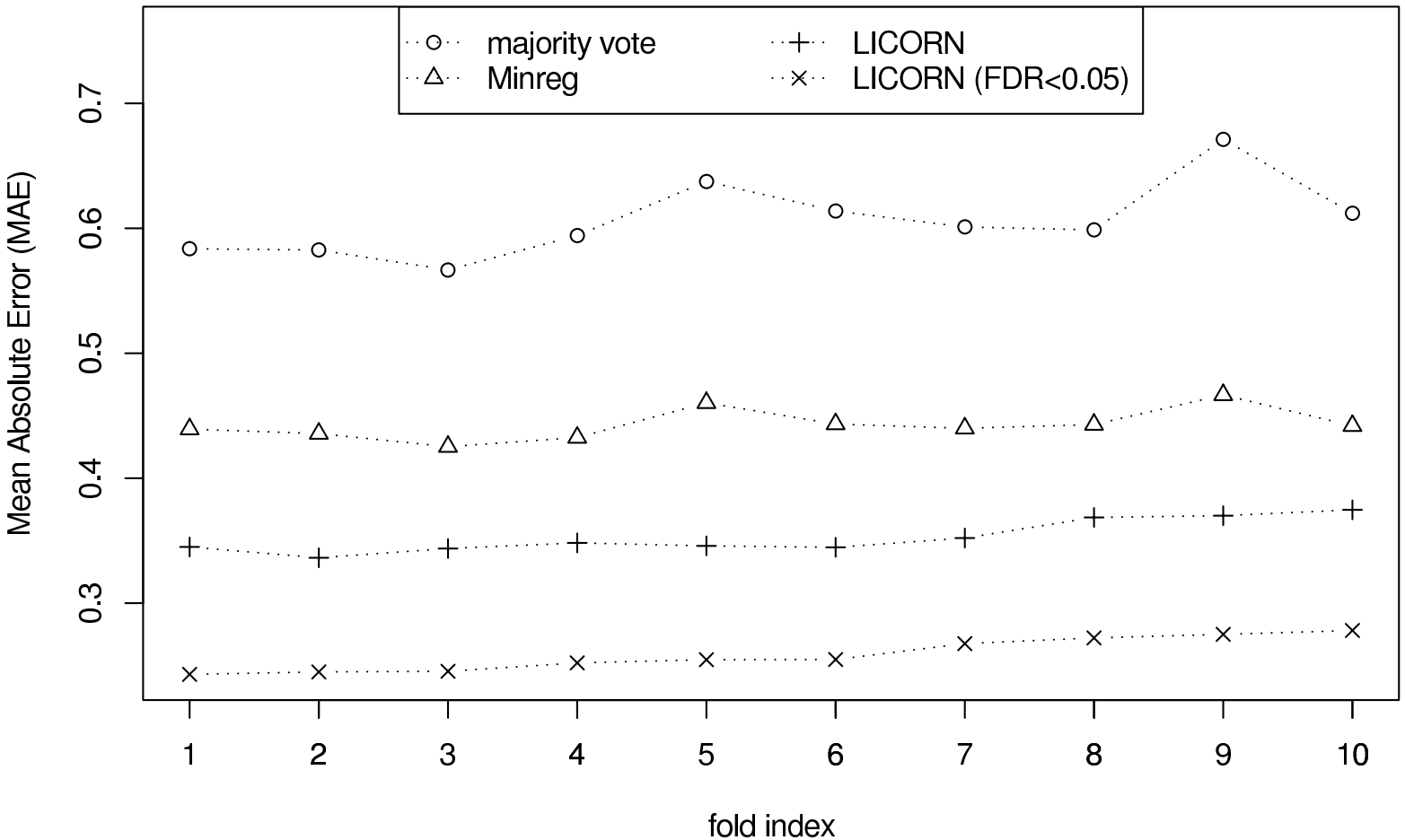
Best-first search for the k-best co-activators and repressors of g.

Assessment of candidate networks

- Rank candidate networks ((A, I) pairs) wrt a local score (MAE)
- Select (n-)best network(s)
- Associate a statistical sigificance to those networks : non-paramteric approach, permutation-based (Benjamini et al., 2001).

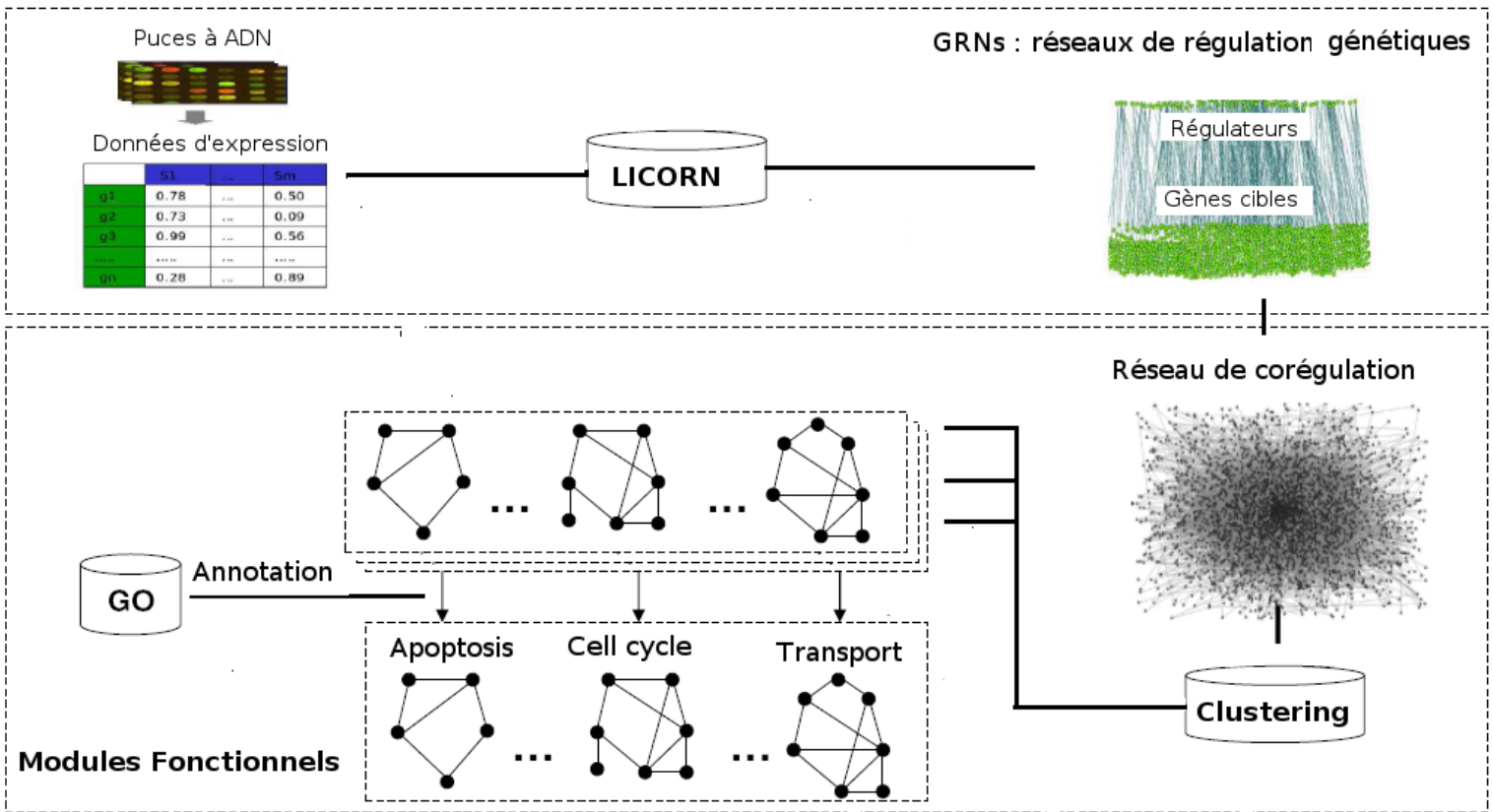


10-CV evaluation



From local to global patterns

(Birmelé et al. BMC 2008)

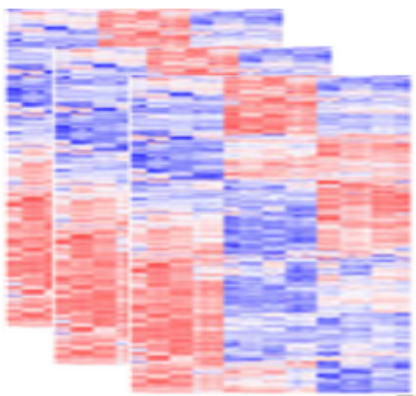


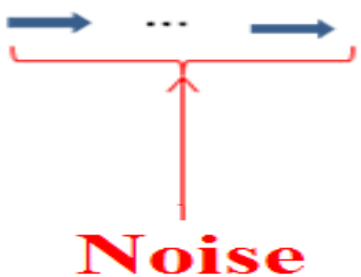
On-going work - Perspectives

- Combine local networks to build a gloabl regulation graph (ILP, frequent graph mining, ...)
- Integrate other information sources (promoter sequence, genomic alterations, miRNA, proteins, epigenetic, ...)
- More powerful evaluation for networks : select networks that are supported by some domain model

Context

- A bioinformatics task from gene expression dataset:
  - Mining co-expressed genes (Sets of genes that are jointly expressed) $\rightarrow$ discretisation + extraction of frequent/closed/maximal itemsets (e.g. Apriori [Bor02]).

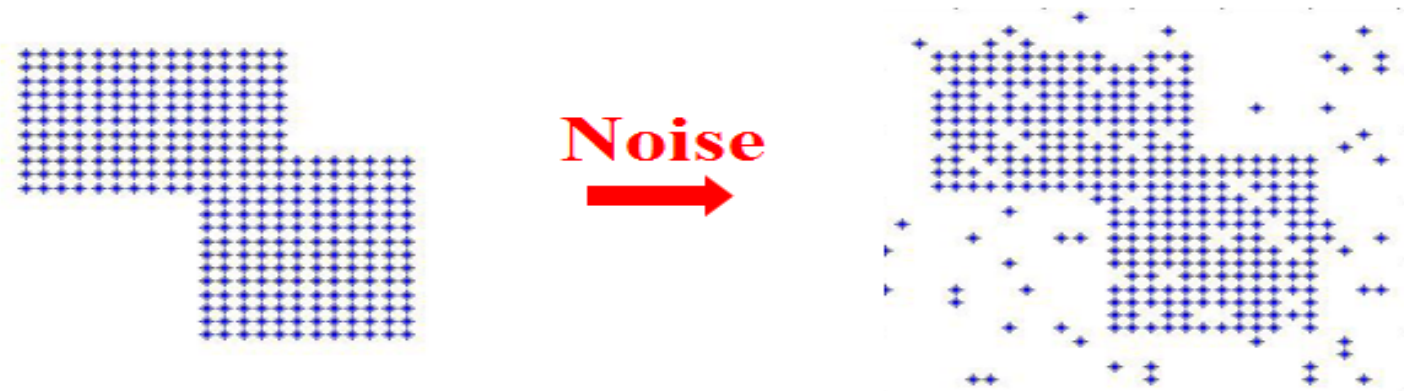




	p_1	p_2	...	p_m
f_1	1	0	...	0
f_2	0	1	...	1
f_3	1	0	...	0
...	...	...	...	...
f_{n-1}	1	0	...	0
f_n	0	1	...	1

The problem

- **Effect of noise**
  - Shattering relevant itemsets into a small irrelevant itemsets $\rightarrow$ explosion in the number of resulting itemsets.



- **Aim and intuition**
  - Mine efficiently a small number of maximal regions of 1, *p*otentially overlapping, and verifying density and minimal support constraints.
  - by combining data mining methods with graph algorithms.

Related work

- **Complete approaches**
  - Methods based on the level-wise principle [Man04, Bes05, Bes06, Liu06, Che06]
  - Handle anti-monotone constraints to prune the space search
  - Quasi-biclique methods [Uno08].

  X Large number of itemsets extracted.
  X Very expensive in execution time for dense data
- **Non-complete approaches**
  - Bi-clustering methods [Pre06]

  X Difficulty in the choice of parameters.
  - Heuristic methods [Mou11]

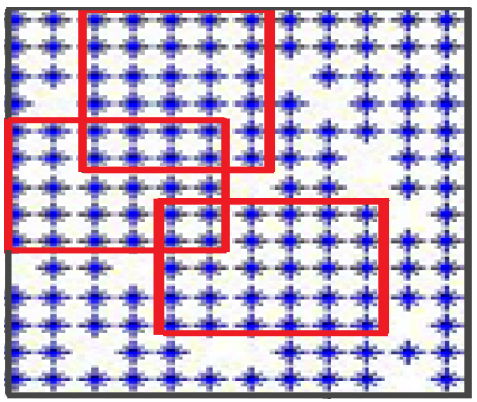
  X Still too many results and redundancy.

HANCIM : Hybrid Approach for Noisy Contexts Itemset Mining

- Consists of two main steps:
  - Identification of a seed pattern s_i
  - Construction of a dense region (O, A) such that $s_i \subseteq A$
- Extracts the maximal regions $M = (A, O)$ such that:
  - Density : $density(M) \geqslant \delta$
  - Minimal support : $\frac{|O|}{|O_{context}|} \geqslant \sigma$.

Seed patterns and the adaptative support

- Use all maximal frequent patterns of D as seeds [Mou11]
  X A high redundancy in the obtained results.
  X Quite expensive especially for dense contexts.



- Seed patterns should :
  - be small enough to be easy to compute
  - favour the extraction of diverse seeds (small overlap) to avoid redundancy in the resulting regions

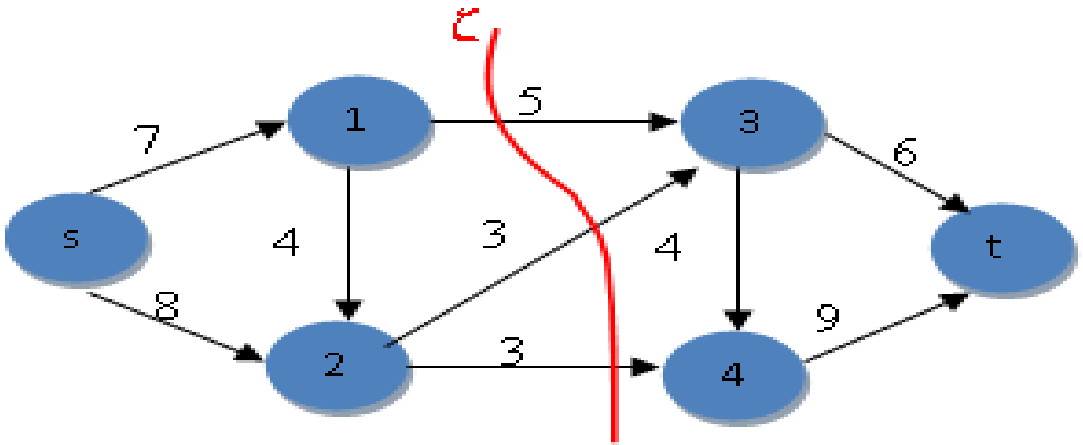
Seed patterns and the adaptative support (ctd.)

- O set of observations, A set of attributes, two contexts on A and O : D and D_{cs} (updated after each new seed is extracted)

while $\exists$ *seed pattern* $s_i \subseteq D_{cs}$ **do**
Compute the region $(O', A') \subseteq D$ such as $(s_i \subseteq A')$ and $(\frac{|O'|}{|O|} \geqslant \sigma)$ and (density$((O', A'))\geq \delta$)
$Updated\_D_{cs}$: D_{cs} where all elements of (A', O') are set to zero
Support = $\frac{Support \times density(UpdatedD_{cs})}{density(D_{cs})}$; $D_{cs} = updated\_D_{cs}$;
end

Searching for dense regions

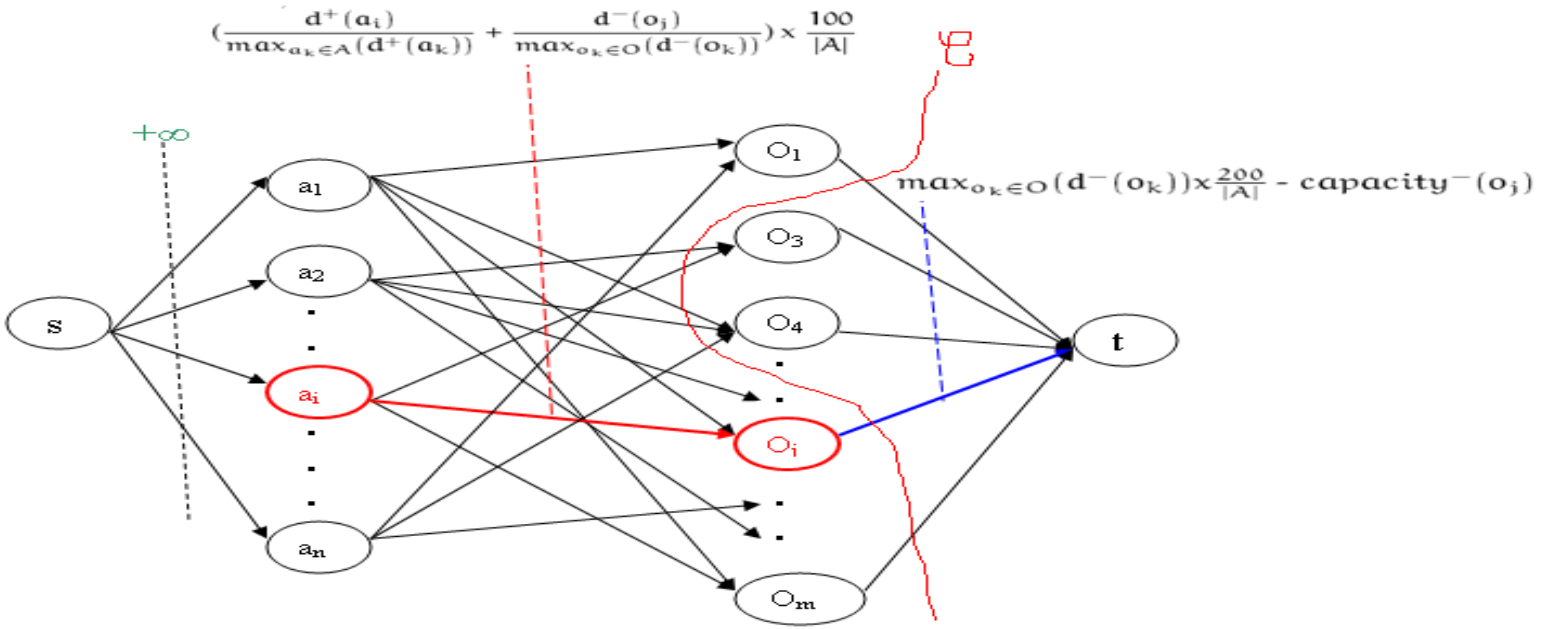
- Searching for a maximal dense region including a seed pattern s_i
- Based on graph algorithms: maximal flow/minimal cut.



Searching for dense regions

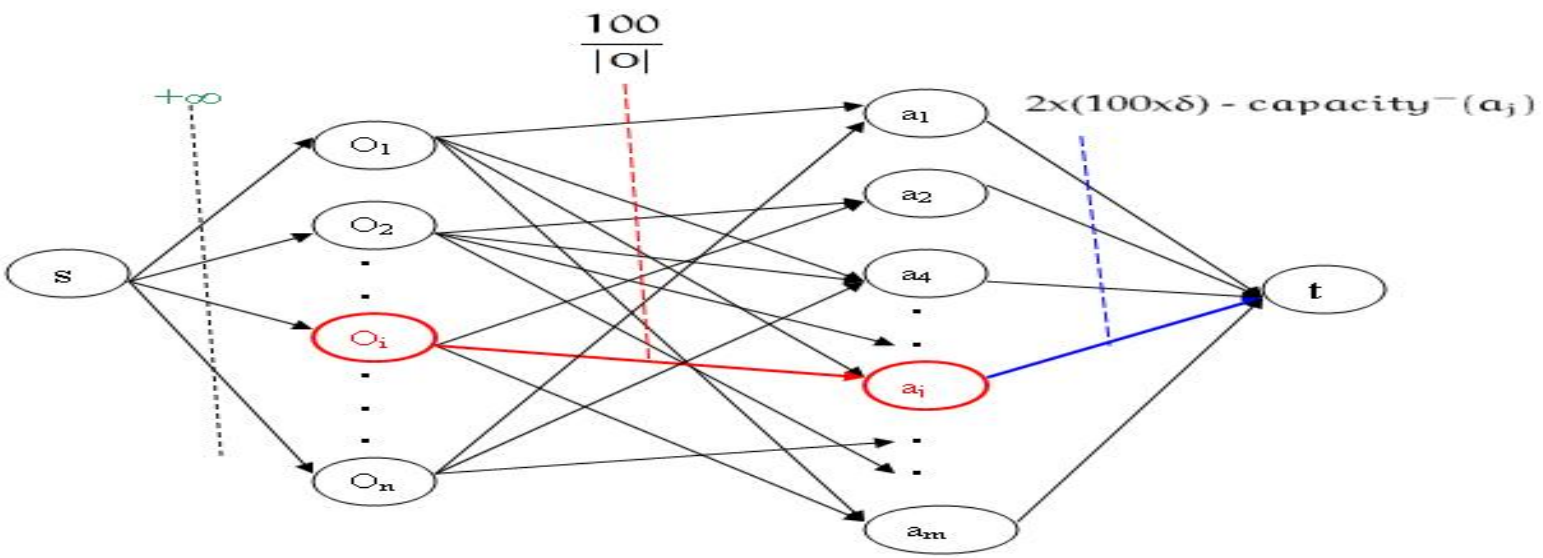
Searching for the maximal dense region including a seed pattern s_i

- Construct the augmented and weighted bipartite graph corresponding to s_i
- Compute a minimal st-cut : push-relabel [Che97]
  $\Rightarrow$ a dense subgraph $G_0=(O_0, s_i)$ where the observations O_0 are strongly linked to the attributes s_i



Searching for dense regions

- Construct the augmented and weighted bipartite graph corresponding to O_0
- Compute a minimal st-cut
  $\Rightarrow$ a dense subgraph $G_1=(O_0, A_1)$ where $s_i \in A_1$ and each attribute of A_1 has a density greater than δ



Experiments on gene expression datasets

- Detection of co-expression relationships between genes from the **Gasch dataset** [Gas00]
  - Expression measures of 2993 genes over 173 observations.
  - Discretization model described by [Pre06].
  - Parameters: σ=20% and δ=80%
- Running time :
  - *Bimax* : calculation stoped after 1 week
  - *HANCIM* : results obtained after 12 minutes.
- Comparison with 100 biclusters published by Bimax [Pre06].

Experiments on real data

- Calculate the enrichment of extracted biclusters in Gene Ontology terms (GO)[Che98].
- The 100-top biclusters extracted :
  - *Bimax* : have p-values ranging between $3e^{-2}$ and $3e^{-4}$.
  - *HANCIM* : have p-values less than e^{-5}.
- The best annotated bicluster:
  - *Bimax* : has a p-value equal to $3e^{-4}$.
  - *HANCIM* : has a p-value equal to e^{-38}.

	$< e^{-2}$	$< e^{-3}$	$< e^{-4}$	$< e^{-5}$	$< e^{-10}$	$< e^{-20}$
HANCIM	94%	48%	28%	18%	7%	4%
BiMax	34%	5%	0%	0%	0%	0%

Conclusion & Pespectives

- A new approach based on max. flow/min. cut algorithms for mining patterns in noisy contexts.
- The results are very promising regarding:
  - quality and size of the extracted patterns
  - reasonable running time
  - annotation quality of results
- Perspectives:
  - Adapt weigth one of the bipartite graph to bias search towards regions that take domaie knowledge into account
  - Links with 'noisy closure'

Collaborators

LIPN, Univ. Paris Nord
I. Chebil
L. Létocart
K. Mouhoubi

Univ. Evry
M. Elati
P. Neuvial
R. Nicolle

Institut Curie
E. Barillot
F. Radvanyi

IGM, Orsay
M. Bolotin