A Data Mining Approach to Highlight Relations Between Functional Modules

Pierre-Nicolas Mougel[1] **Marc Plantevit**[1] **Christophe Rigotti**[1] **Olivier Gandrillon**[2]
Jean-François Boulicaut[1]

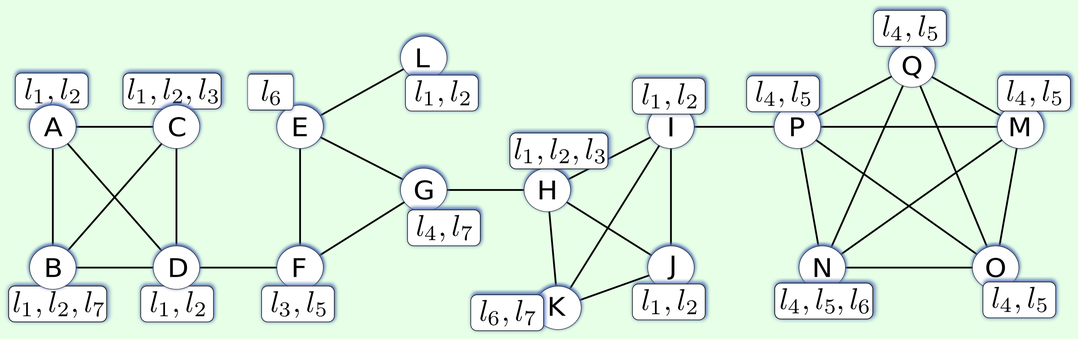
[1] **Combining Team** - Université de Lyon, CNRS, LIRIS, UMR5205, INRIA
[2] **BM2A Team** - Université de Lyon, CNRS, CGMC, UMR5534

Context

Graphs representing entities in interaction (protein/protein interaction network) with labels associated to vertices (biological situations in which the corresponding genes are overexpressed)

- Example of dataset

Proteins $= \{A, ..., Q\}$, biological situations $= \{l_1, ..., l_7\}$



- Corresponding gene expression matrix

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
l_1	•	•	•	•				•	•	•		•					
l_2	•	•	•	•				•	•	•		•					
l_3			•			•		•									
l_4							•						•	•	•	•	•
l_5						•							•	•	•	•	•
l_6					•						•			•			
l_7		•					•				•						

Genericity

This approach might be used on a large class of dataset, for instance:

- Protein/protein interaction graph with biological situations
- Scientific coauthor relationship and conferences
- Wikipedia pages and their contributors

MHCS Intuition

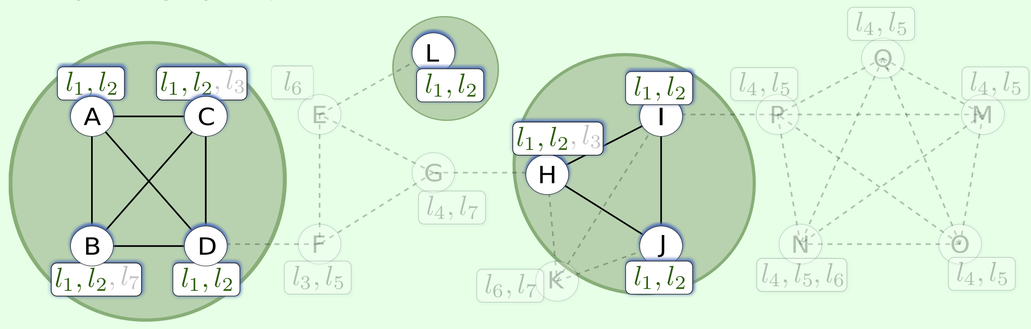
Collection of dense subgraph sharing common properties

- **M**aximal: Most specific patterns
- **H**omogenous: Common properties
- **C**lique: Dense Subgraph
- **S**et: Collection of cliques

MHCS Definition

Set of cliques satisfying 3 constraints:

- $\mathbb{C}_{\alpha}^{lab}$: At least α common labels
- $\mathbb{C}_{\kappa,\beta}^{clique}$: At least κ cliques of size at least β
- $\mathbb{C}^{sep}$: Cliques must be separated (i.e. the union of any pair of cliques is not a clique in graph G)



Interest

Allow to find non trivial relationships between dense subgraphs:

- Local hubs (e.g., transcription factor activating genes falling in different functional categories)
- Disconnected subgraphs sharing similar properties
- Small groups of entities evolving around core groups sharing similar properties (suggest interaction investigation)

Experiments

Biological Dataset

The graph: STRING (`http://string-db.org/`)

- Protein/protein interaction database
- Use human genes

Biological situations from SQUAT (`http://bsmc.insa-lyon.fr/squat/`)

- Binarized gene expression database from SAGE data

$\Rightarrow$ 15,572 vertices (genes), 458,713 edges (protein interactions) and 486 different labels (biological situations where genes are overexpressed)

L2L Significance Scoring

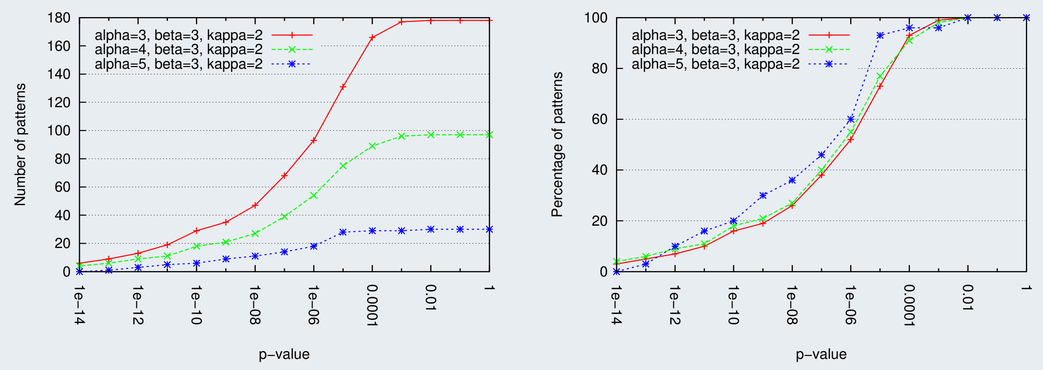


Figure: p-value cumulative distribution function

- 90% of the extracted patterns have a p-value inferior to 10^{-5} (p-value L2L [Newman J. and Weiner A.: L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biology (2005)]).

Example of pattern

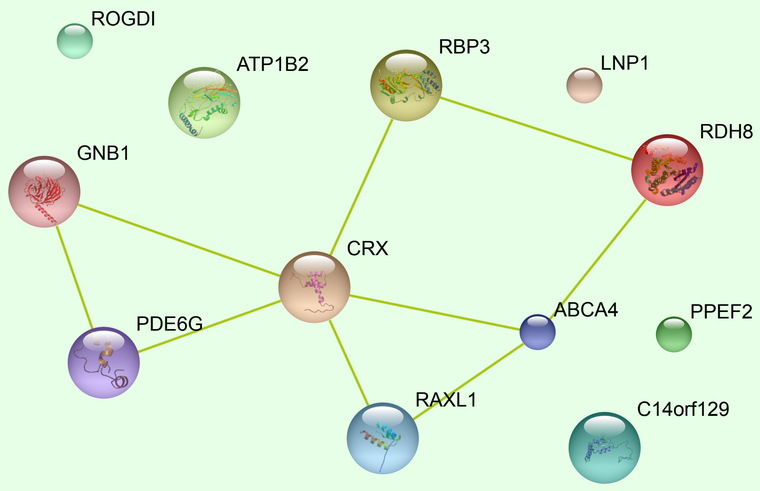


Figure: Genes overexpressed in 3 common biological situations (interaction graph from STRING)

- Most genes harbor a known function in the retina
- Gene CRX behaves as a hub, which is consistent with its transcription factor function

Time Performances

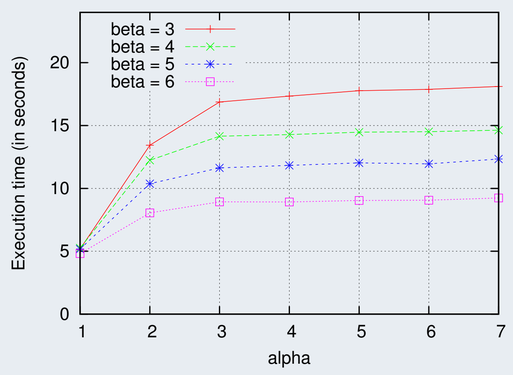


Figure: $\kappa = 2$

- Worst execution time on this dataset: 19 seconds
- Less than 17 minutes on a graph with 479,067 vertices, 386,838 edges and 3,607 distinct labels (same parameters).