Identification of functional modules based on transcriptional regulation structure

E. Birmelé[1] M. Elati[2] C. Rouveirol[2] C. Ambroise[1]

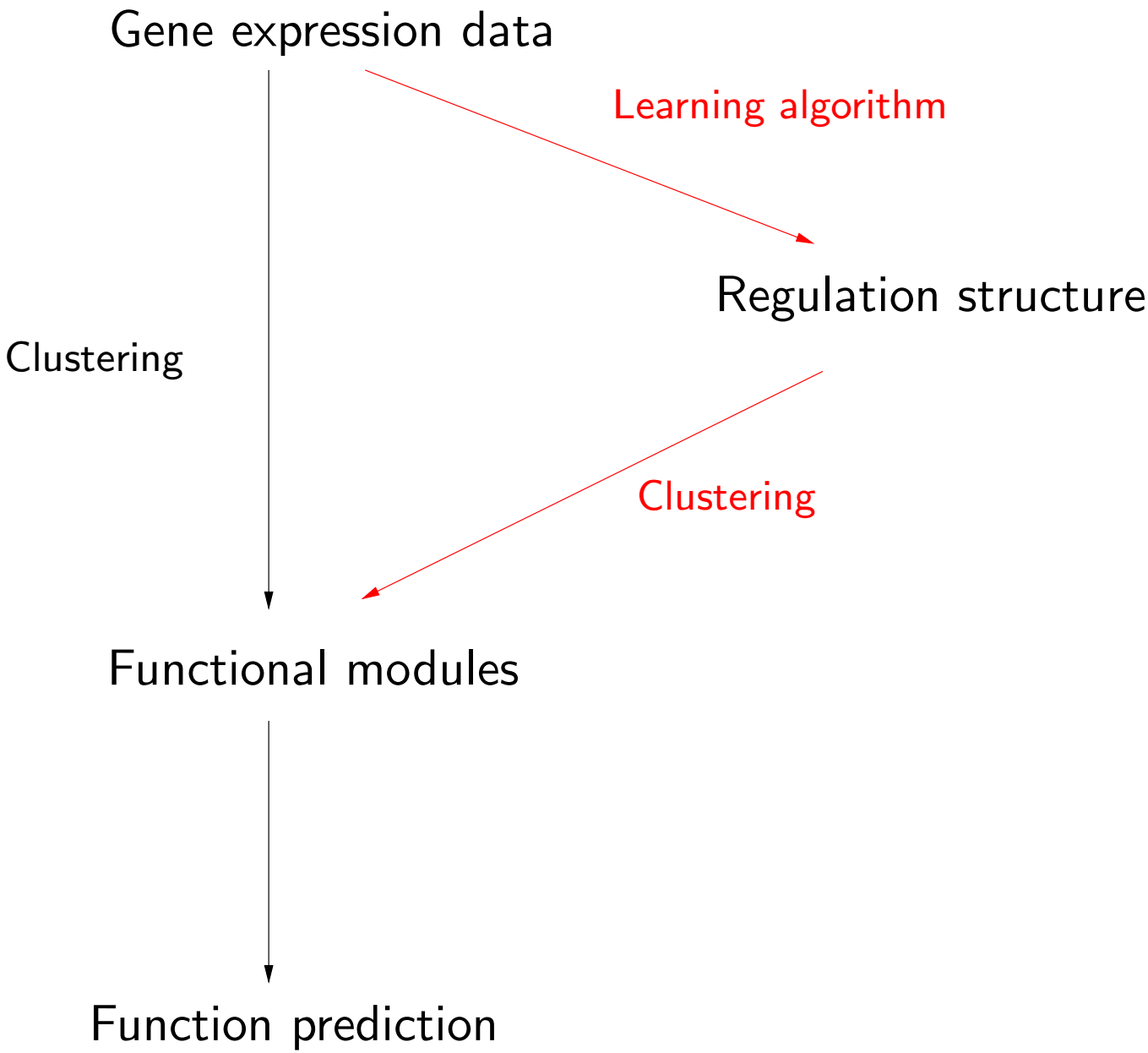
[1]Laboratoire Statistique et Génome
Université d'Evry

[2]LIPN
Université Paris 13

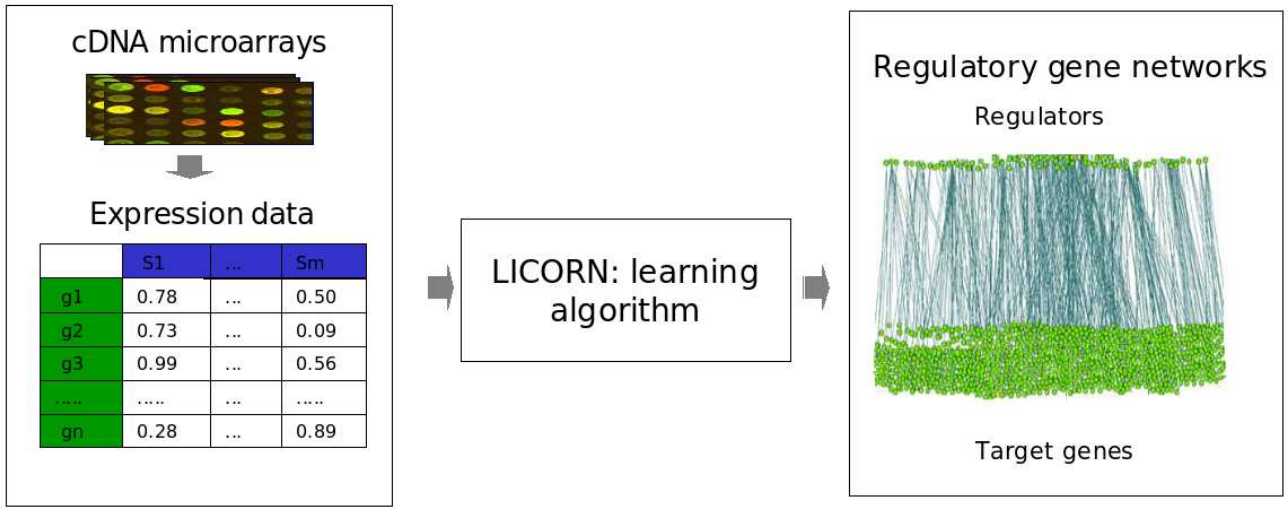
MLSB, 2007

Outline

Introduction



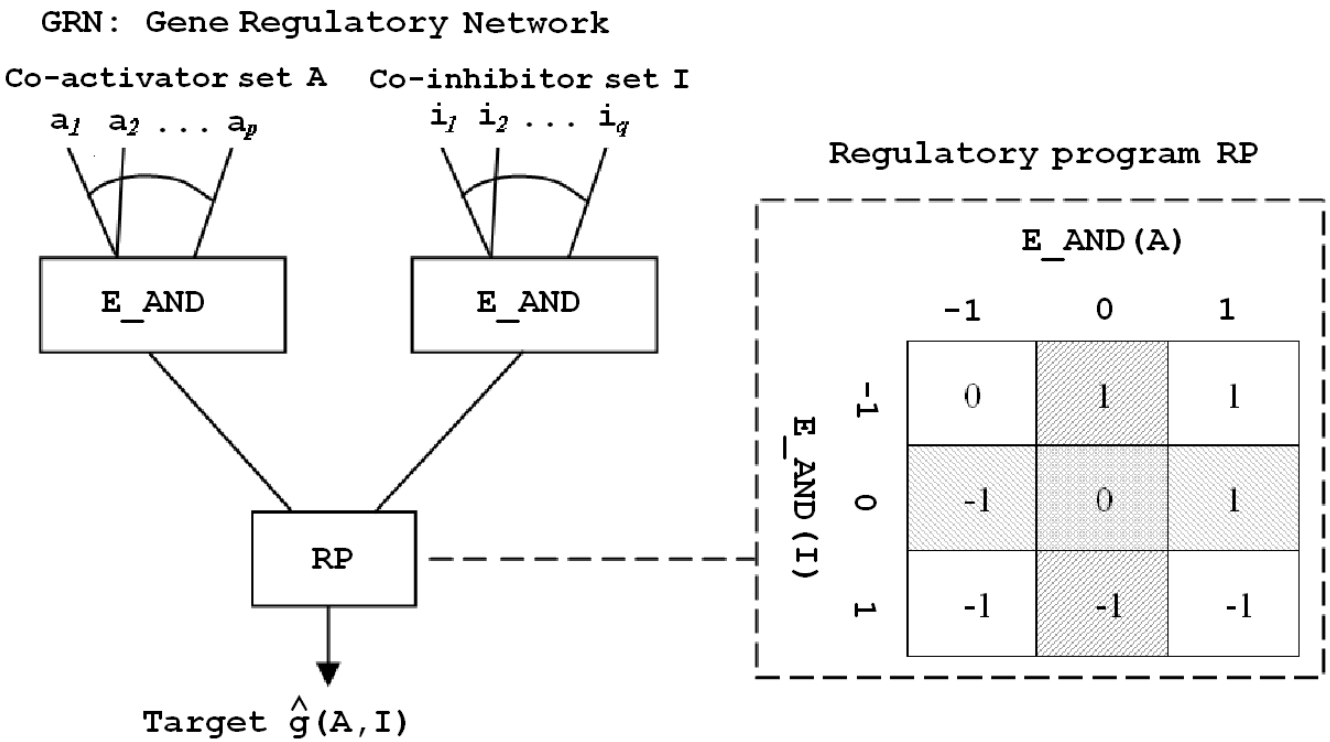
General overview



[1] M. Elati and al., LICORN: learning co-operative regulation networks from gene expression data, Bioinformatics, 2007

Regulation model

- Genes are split into a set $\mathcal{R}$ of regulators and $\mathcal{G}$ of target genes according to litterature.
- Expression values are discretized in a three-value logic: -1(under-expressed), 0(normal) or 1(over-expressed).
- A GRN associated with a target gene g is a pair (A, I) of respectively co-activators and co-inhibitors.



Licorn Step 1: Frequent co-regulator sets

Given the three-valued expression matrix, a co-regulator set $C \in \mathcal{R}$ and its 1- and -1-supports, denoted $\mathcal{S}^1(C)$,$\mathcal{S}^{-1}(C) \in \mathcal{S}$,

C is frequent if and only if

$$\max(|\mathcal{S}^1(C)|, |\mathcal{S}^{-1}(C)|) \geq T_s$$

where T_s is a user-defined minimum support threshold ($\leq 20\%$)

Licorn Step 2: best putative GRN's

Fix an overlap threshold T_o ($\geq 50\%$) and, for each gene g,

- Determine the set $\mathcal{A}(g)$ of frequent co-regulator sets which are over(under)-expressed in the sample with probability $\geq T_o$ when g is over(under)-expressed. The elements of $\mathcal{A}(g)$ are the candidate activator sets.
- Determine the set $\mathcal{I}(g)$ of candidate inhibitor sets.
- For $A \in \mathcal{A}(g)$, $I \in \mathcal{I}(g)$ with $A \cap I = \emptyset$,

$$h_g(A, I) = \sum_{s \in \mathcal{S}} |g_s - \hat{g}_s(A, I)|$$

The pair (A, I) with the best (lowest) score is the best putative GRN for gene g.

Licorn Step 3 - Statistical significance

- Generation of p-values for the GRN's by randomizing the samples.
- Selection of the significant GRN's by the FDR procedure

Summary of the regulation structure learning step

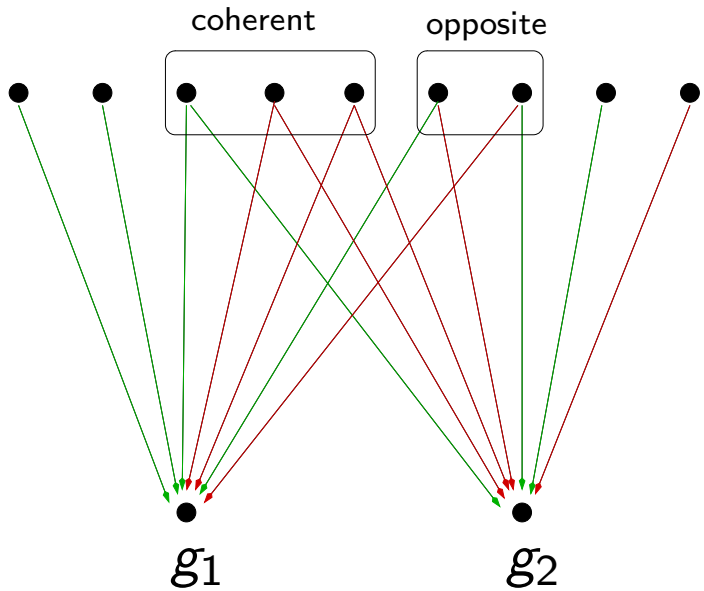
- Find the frequent co-regulator sets
- For each gene, find the best pair (activator set, inhibitor set)
- Correct by multiple hypothesis testing

The three steps to detect modules

- Define a similarity measure ϕ_λ, $\lambda \in [0, 1]$ being the weight of opposite regulations
- Run a clustering algorithm
- Define a score $S(\lambda)$ to choose the best value for λ

The similarity measure

$$\phi(g_1, g_2) = \frac{\#\text{convergent regulators} + \lambda \#\text{opposite regulators}}{\#\text{total number of regulators}}$$



$$\phi(g_1, g_2) = \frac{3+2\lambda}{9}$$

The clustering algorithm

We apply the MCL algorithm[1].

- Suited for weighted graphs
- No prior knowledge about the number of clusters
- Reproducible

[1] van Dongen, Graph clustering by flux simulation, PhD thesis, 2000

The score of a clustering (1)

For each cluster C in the clustering:

- we determine the GO terms (BP ontology) over-represented in C with a rate of 5%.
- we obtain a set T_C of GO terms and a set of associated p-values $\{p_t, t \in T_C\}$.

The score of a clustering (2)

The score of the clustering is

$$S(\lambda) = \sum_{C, c_{min} \leq |C| \leq c_{max}} \sum_{t \in T_C} -\log(p_t) \frac{|C_t|}{|C|}$$

Example in yeast: $\lambda = .3$, $t =$*vacuolar transport*. One cluster C of 44 elements has 7 genes associated to t, which yields a p-value of $5.16e-05$. It's contribution to $S(.3)$ is

$$-\log(5.16e-05)\frac{7}{44}$$

Summary of the clustering step

- For several values of λ,
  - Compute the similarity matrix with parameter λ
  - Run the MCL algorithm
  - Map the clusters to GO terms and compute the score of the clustering
- Choose the clustering with the best score

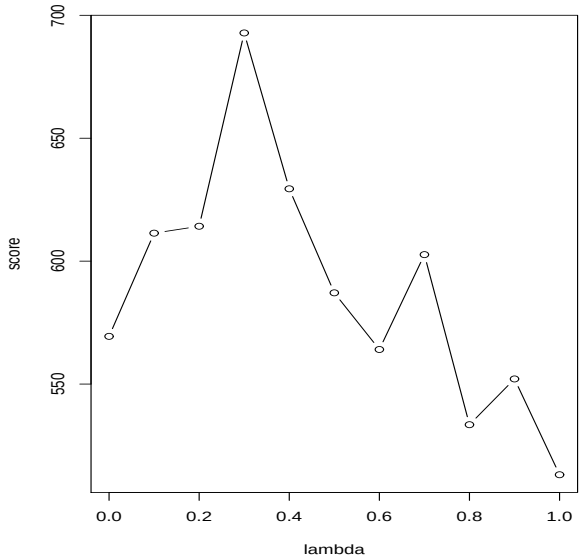
The data set

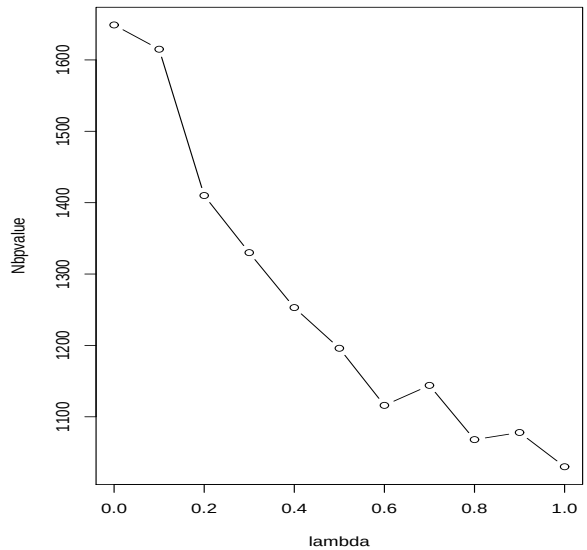
- Gasch data set[2]: 173 stress conditions for 6152 genes.
- Set of 237 known and putative transcription factors.
- Applying Licorn: 2041 significant GRNs
- 167 genes not appearing in the GO data for yeast

[2]Gasch et al., Genomic expression programs in the response of yeast cells to environmentel changes, 2000

Mapping to GO terms (1)

- Best score obtained for $\lambda = .3$





- 59 clusters, among which one giant (1457 genes) and 20 to small ones ($\leq 4 genes$)

Mapping to GO terms (2)

GOBPID	Pvalue	Count	Size	clusterId	clusterSize	Term
0000278	4.63e-08	8	43	5	28	mitotic cell cycle
0022402	6.04e-07	9	81	5	28	cell cycle process
0007049	8.31e-07	9	84	5	28	cell cycle
0022403	1.49e-06	8	66	5	28	cell cycle phase
0006511	1.83e-06	6	45	13	19	ubiquitin-dependent protein catabolic process
0019941	1.83e-06	6	45	13	19	modification-dependent protein catabolic process
0051603	2.09e-06	6	46	13	19	proteolysis involved in cellular protein catabolic pr
0043632	3.47e-06	6	50	13	19	modification-dependent macromolecule catabolic p
0044257	4.40e-06	6	52	13	19	cellular protein catabolic process
0000279	4.83e-06	7	54	5	28	M phase

Perspectives

- Function prediction
- Human data
- Supervised classification

Thanks

- Monique Bolotin-Fukuhara
- Pierre Neuvial, Emmanuel Barillot
- François Radvanyi