LABORATOIRE MAP5

UNIVERSITÉ PARIS DESCARTES

Identification of deregulated transcription factors in bladder cancer

Magali Champion, E. Birmelé, J. Chiquet and P. Neuvial

Colloque Apprentissage de Réseaux : de la Théorie aux Applications en Biologie et Ecologie

October 14th 2016

Motivations : *Bladder cancer, a critical disease*

- One of the most widespread cancers in North America and Europe

New cases	Number of deaths	
76,960	16,390	(US, 2015)

- Four times more common in men than in women
- Major risks include smoking and age

Objectives (*LIONS project*)

Create mechanistic models of cancers to

- Understand how gene expression is influenced by genomic events,
- Identify deregulated transcription factors and their targets.

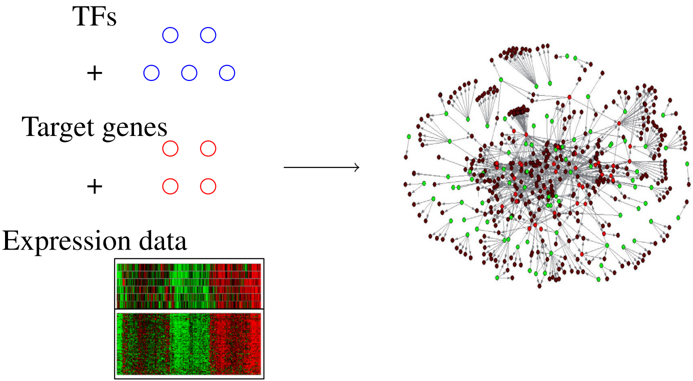
Gene expression clustering is commonly used to identify subtypes of cancer. Specific studies of these subtypes are done to :

- *gain new insights into the molecular heterogeneity of tumors,*
- *improve the management of cancer patients.*

Model : *Networks for representing gene regulations*

We used data from the Carte d'Identité des Tumeurs French program :

- 182 patients : 3 healthy + 179 cancerous,
- 1,704 TFs + 16,967 targets = 18,671 genes.



TFs

+

Target genes

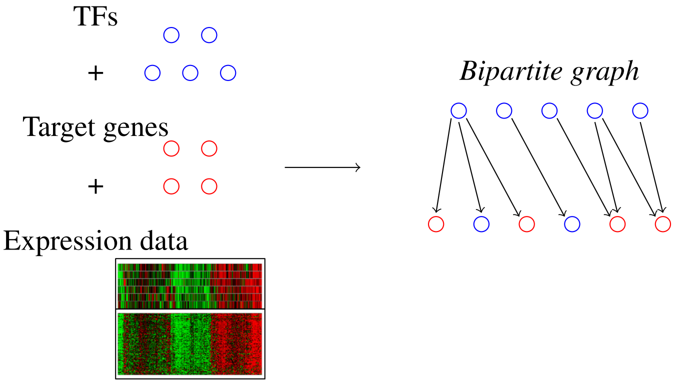
+

Expression data

Model : *Networks for representing gene regulations*

We used data from the Carte d'Identité des Tumeurs French program :

- 182 patients : 3 healthy + 179 cancerous,
- 1,704 TFs + 16,967 targets = 18,671 genes.



TFs

+

Target genes

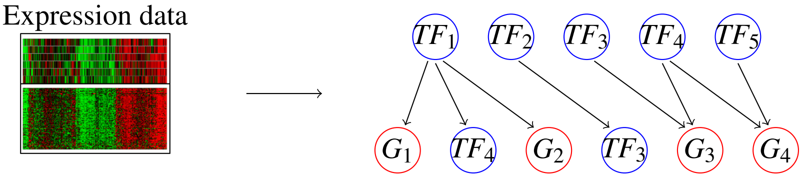
+

Expression data

Bipartite graph

Methods : *Overview*

1. Inferring the Gene Regulatory Network of reference
2. Computing a deregulation score for target genes
3. Finding the TFs that best explain the deregulated target genes

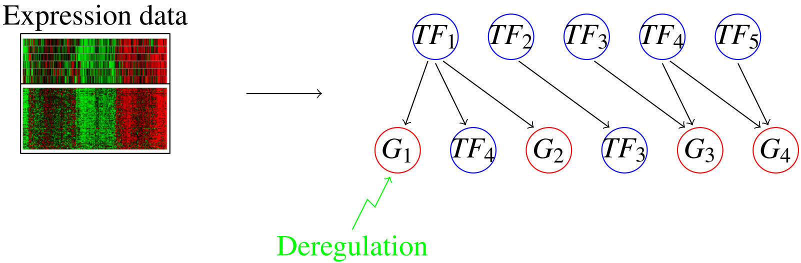


Inferring the GRN of reference

LICORN model

Methods : *Overview*

1. Inferring the Gene Regulatory Network of reference
2. Computing a deregulation score for target genes
3. Finding the TFs that best explain the deregulated target genes

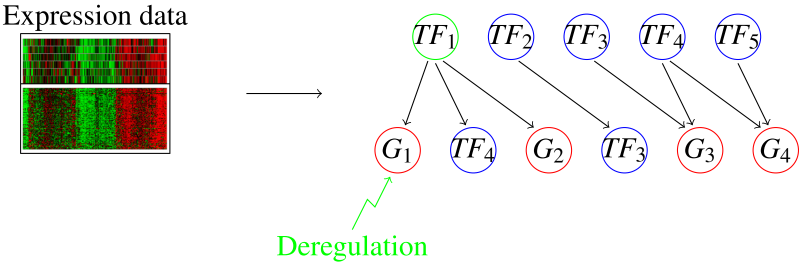


Computing a deregulation score

EM-strategy

Methods : *Overview*

1. Inferring the Gene Regulatory Network of reference
2. Computing a deregulation score for target genes
3. Finding the TFs that best explain the deregulated target genes



Identifying the deregulated TFs

Linear model

Method I : Inferring a Gene Regulatory Network

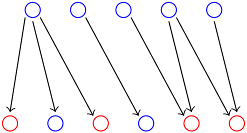
Main goal

Given a set of target genes $\mathcal{G}$, a set of TFs, their expression matrices $M_{\mathcal{G}}$ and M_{TF}, we aim at finding for each target gene the set of regulators that best explains the level of expression.

Classical methods of inference include :

- linear independent regressions,
- Bayesian networks modelling...

$\longrightarrow$ LICORN (*LearnIng Cooperative Regulation Networks from gene expression data*), which aims at finding cooperative regulations between co-regulated TFs and target genes.



Elati et al (Bioinformatics, 2007)

Method I : Inferring a Gene Regulatory Network

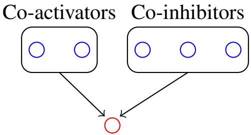
Main goal

Given a set of target genes $\mathcal{G}$, a set of TFs, their expression matrices $M_{\mathcal{G}}$ and M_{TF}, we aim at finding for each target gene the set of regulators that best explains the level of expression.

Classical methods of inference include :

- linear independent regressions,
- Bayesian networks modelling...

$\longrightarrow$ LICORN (*LearnIng Cooperative Regulation Networks from gene expression data*), which aims at finding cooperative regulations between co-regulated TFs and target genes.



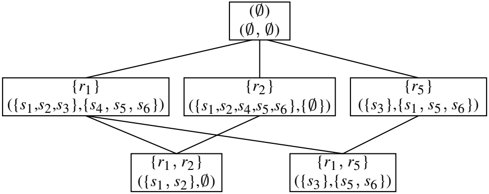
Elati et al (Bioinformatics, 2007)

Method I : Inferring a GRN using LICORN

Step 1 : Mining global candidate of co-regulator sets

- Computing frequent itemsets from discrete data : adaptation of the Apriori algorithm (Agrawal, 1993) to the ternary case,
- Creating the candidate co-regulator sets.

M_{TF}	r_1	r_2	r_3	r_4	r_5
s_1	1	1	0	0	-1
s_2	1	1	0	0	0
s_3	1	0	0	0	1
s_4	-1	1	1	0	0
s_5	-1	1	-1	0	-1
s_6	-1	1	0	1	-1



$(\emptyset)$
$(\emptyset, \emptyset)$

$\{r_1\}$
$(\{s_1,s_2,s_3\},\{s_4, s_5, s_6\})$

$\{r_2\}$
$(\{s_1,s_2,s_4,s_5,s_6\},\{\emptyset\})$

$\{r_5\}$
$(\{s_3\},\{s_1, s_5, s_6\})$

$\{r_1, r_2\}$
$(\{s_1, s_2\},\emptyset)$

$\{r_1, r_5\}$
$(\{s_3\},\{s_5, s_6\})$

Threshold : $\max\left(|\mathcal{S}^1|, |\mathcal{S}^{-1}|\right) \geq 20\% \times 6$

Method I : Inferring a GRN using Licorn

Step 2 : Searching for candidate GRNs

- Co-regulator status : $\mathcal{S}_X = \begin{cases} -1 & \text{if } \forall x \in X, x = -1 \\ 1 & \text{if } \forall x \in X, x = 1 \\ 0 & \text{otherwise} \end{cases}$
- Co-regulator constraint : X is a co-regulator set for g if

$$\exists x, y \in \{-1, 1\}, \;\; \mathrm{Coreg}_X(x, y) = |\mathcal{S}_X^x \cap \mathcal{S}_g^y| / |\mathcal{S}_g^y| \geq 60\%.$$

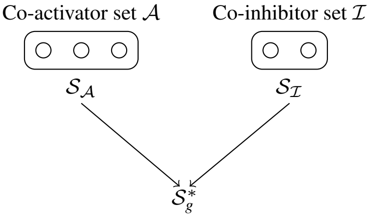
	r_1	r_2	Co-reg	g
s_1	1	1	1	1
s_2	1	1	1	-1
s_3	1	0	0	-1
s_4	-1	1	0	0
s_5	-1	1	0	0
s_6	-1	1	0	0

$\{r_1, r_2\}$
co-activates g

Method I : Inferring a GRN using Licorn

Step 3 : Scoring GRNs

- Define a regulatory program



Co-activator set $\mathcal{A}$

Co-inhibitor set $\mathcal{I}$

$\mathcal{S}_\mathcal{A}$

$\mathcal{S}_\mathcal{I}$

$\mathcal{S}_g^*$

$$\begin{array}{cc|c|c|c|} & & & \mathcal{S}_\mathcal{A} & \\ & & -1 & 0 & 1 \\ \hline & -1 & 0 & 1 & 1 \\ \hline \mathcal{S}_\mathcal{I} & 0 & -1 & 0 & 1 \\ \hline & 1 & -1 & -1 & -1 \\ \hline \end{array}$$

- Rank candidate networks (($\mathcal{A}, \mathcal{I}$) pairs) for the regulation of gene g in terms of Mean Absolute Error (MAE) :

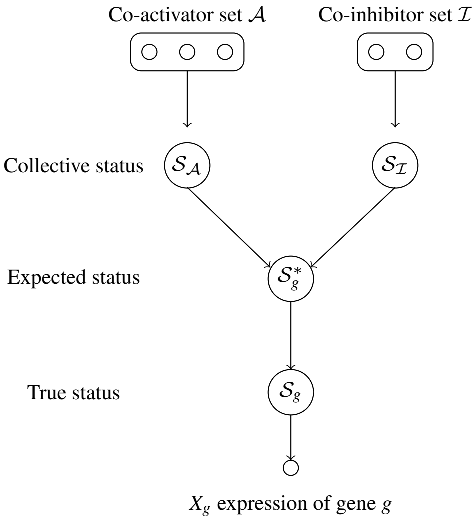
$$MAE_g(\mathcal{A}, \mathcal{I}) = \sum_{s=1}^{n} |\mathcal{S}_g^*(s) - \mathcal{S}_g(s)|.$$

- Select the best networks :

$$\text{GRN}^*(g) = \text{argmin}_{(\mathcal{A},\mathcal{I})} MAE_g(\mathcal{A}, \mathcal{I}).$$

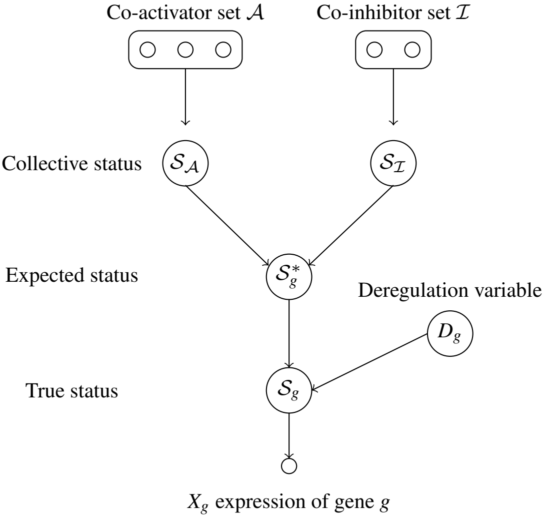
Method II : Computing a deregulation score

- *Model* :


Co-activator set $\mathcal{A}$
Co-inhibitor set $\mathcal{I}$
Collective status
$\mathcal{S}_{\mathcal{A}}$
$\mathcal{S}_{\mathcal{I}}$
Expected status
$\mathcal{S}_g^*$
True status
$\mathcal{S}_g$
X_g expression of gene g

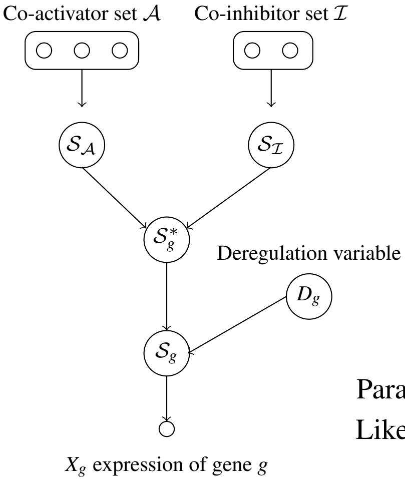
Method II : Computing a deregulation score

- *Model* :


Co-activator set $\mathcal{A}$
Co-inhibitor set $\mathcal{I}$
Collective status
$\mathcal{S}_\mathcal{A}$
$\mathcal{S}_\mathcal{I}$
Expected status
$\mathcal{S}_g^*$
Deregulation variable
D_g
True status
$\mathcal{S}_g$
X_g expression of gene g

Method II : Computing a deregulation score

- *Model* :



Co-activator set $\mathcal{A}$
Co-inhibitor set $\mathcal{I}$
$\mathcal{S}_{\mathcal{A}}$
$\mathcal{S}_{\mathcal{I}}$
$\mathcal{S}_g^*$
Deregulation variable
D_g
$\mathcal{S}_g$
X_g expression of gene g

- $\begin{cases} \mathcal{S}_g = \mathcal{S}_g^* & \text{if } D_g = 0 \\ \forall x \neq \mathcal{S}_g^*, \ \ \mathbb{P}(\mathcal{S}_g = x) = 1/2 & \text{if } D_g = 1 \end{cases}$
- $D_g = 1$ with probability E
- $X_g | \mathcal{S}_g = x \sim \mathcal{N}(\mu_x, \sigma_x)$
- $\mathcal{S}_g$ multinomial distribution with parameters $\alpha = (\alpha_-, \alpha_0, \alpha_+)$

Parameters : $\theta = (\mu_x, \sigma_x, \alpha, E)$

Likelihood : $p(X, Z|\theta) = \underbrace{\ldots\ldots\ldots\ldots}_{\text{intractable}}$

Method II : Computing a deregulation score

EM-strategy :

- initial guess θ^0 of the model parameters,
- E-step : fix θ and compute the conditional probability distribution of the hidden variables given the observed expression values :

$$q(Z) = \mathbb{P}(Z|X, .),$$

- M-step : fix q and find θ that maximizes

$$\sum q(Z) \log \mathbb{P}(X, Z|.).$$

Picchetti et al., BMC Systems Biology (2015)

Method III : Identifying deregulated TFs

Observations :

- $G \in \mathcal{M}_{(p-q)\times q}(\mathbb{R})$ the adjacency matrix associated to the GRN of reference (q TFs and $p-q$ target genes)
- $Y \in \mathcal{M}_{(p-q)\times n}(\mathbb{R})$ the matrix of deregulation score

Linear model :

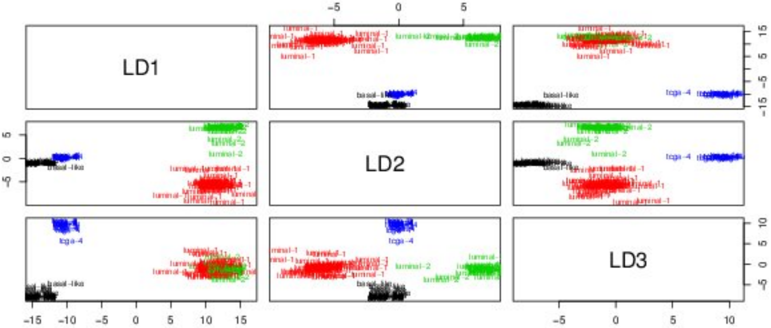
$$Y = G\beta + \varepsilon,$$

where β of size $q \times n$ measures the effect of TF deregulation across all patients.

- Least-squares estimation of β
- Linear Discriminant Analysis (LDA) to make $\hat{\beta}$ sparse

Dimensionality reduction technique for classification that projects a dataset onto a lower dimensional space with the aim of maximizing the separation between multiple classes.

Method III : Identifying deregulated TFs



4 subtypes : luminal 1 luminal 2 basal tcga 4

FIGURE – LDA visualization obtained on the bladder cancer data set.

Biological results

Subtypes	Deregulated TFs
Luminal 1	SPOCD1 (33%), ASXL1 (28%), ZNF295 (28%), FOXM1 (22%), HOXB3 (22%)
Luminal 2	PRDM12 (20%), ASXL1 (20%)
Basal	ZSCAN16 (50%), TBX2 (36%), CEBPB (32%), MNDA (27%), TOP2B (23%), ZNF540 (23%), MYCL1 (23%), ANKRA2 (23%), HOXB3 (23%), PRRX1 (23%)
Tcga 4	MNDA (59%), NFYA (53%), TNFAIP3 (35%), NR3C1 (35%), ZNF440 (29%), TRIM25 (29%), PRRX1 (29%), TRIM32 (24%), NFIA (24%), RUNX3 (24%), HOXD1 (24%), ZNF469 (24%)

TABLE – Top deregulated TFs (number of patients for which the TFs are involved in the deregulation of target genes between brackets)

Conclusion

- Development of a 3-steps strategy for the identification of deregulated genes in sick patients
- Identification of deregulated TFs involved in specific subtypes of bladder cancer

To be done...

- *discuss with biologists to validate these results and check for discoveries ?*
- *find gene sets that characterize specific subtypes of cancer*
- *integrate multi-omics data (copy number of variations, methylation, mutation,...)*

Thank you for your attention !