Biological prior for network inference with Gaussian graphical models.

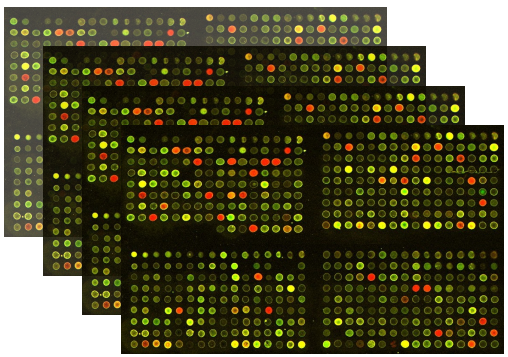
Application to Estrogen Receptor Status in Breast Cancer .

Marine Jeanmougin

MIA - Network Inference, Paris – February 8, 2012

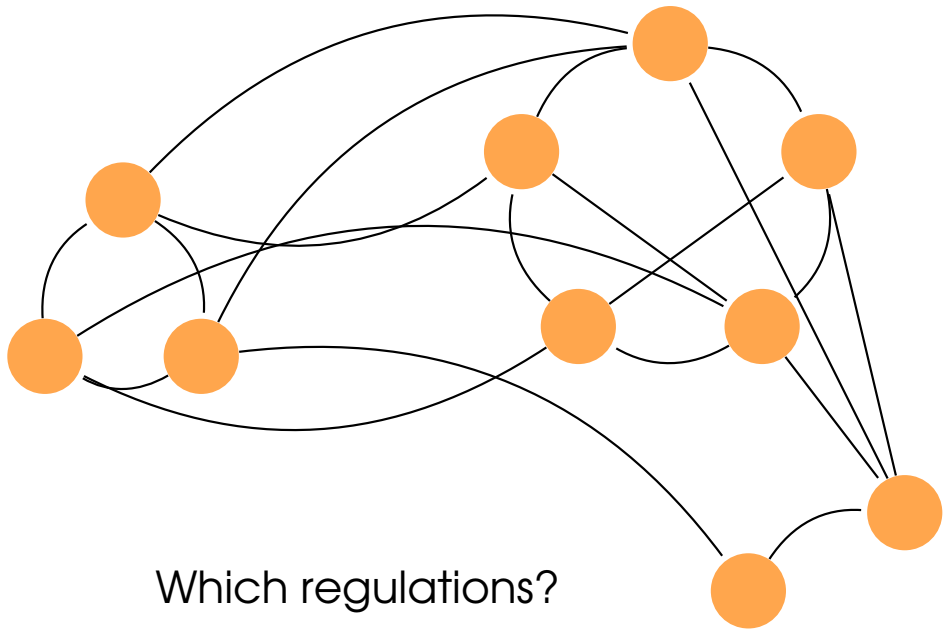
Problem

Microarray data



Inference

Gene regulatory network



$n \approx$ 10s/100s of microarrays

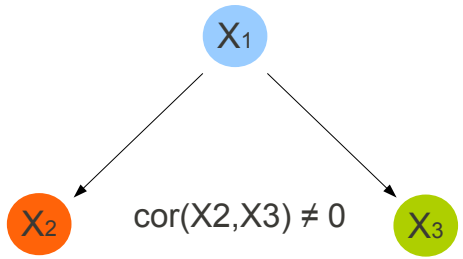
$p \approx$ 1000s of genes

$\mathcal{O}(g^2)$ parameters (edges) !

Which regulations?

Which measure to use ?

- **Correlation**
  - Tends to group genes with close expression profiles



  - Do not provide any clue on how the chain of information goes from gene to gene

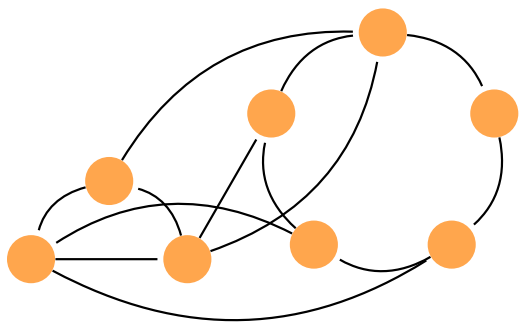
- **Partial Correlation**
  - Quantify the correlation between two genes after excluding the effects of other genes

High dimensional setting

- "large p, small n"
  $\rightsquigarrow$ number of random variables (p) is much larger than the number of individuals (n)
- $p(p-1)/2$ possible interactions

Handling the scarcity of data

- Sparsity:



  Among all possible interactions only a few actually take place.
- Coefficient matrix with mostly zero-valued entries

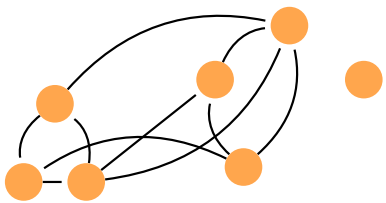
Regularized Gaussian graphical model

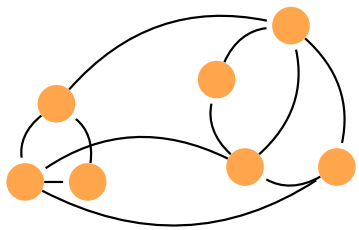
- GGM: a well-studied framework to spot those direct relationships
- Dependency pattern described by the covariance matrix (independency between variables $\Leftrightarrow$ absence of edge)
- Sparse estimation via L1-regularization

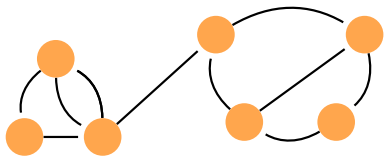
Banerjee, O. and El Ghaoui L. and d'Aspremont A.) Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *JMLR* - 2008

A challenging issue

A vaste space of possible network structures





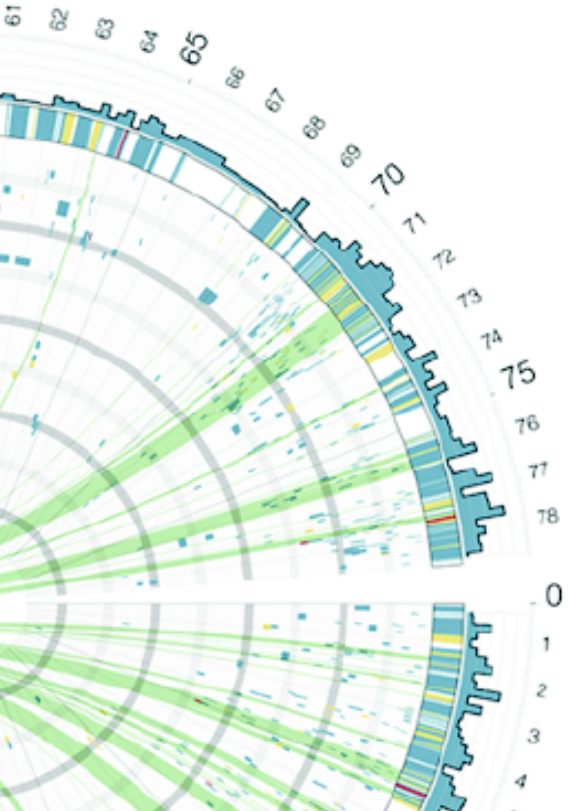


Biological prior knowledge could be used to limit the set of candidate networks

Outline

Method

Biological prior definition



Differential analysis

$X_{ig}^{(c)}$: expression level of the ith sample for gene g under condition c

$$\mathbb{E}(X_{ig}^{(c)}) = \mu_g^{(c)} \quad \text{and} \quad \mathbb{V}(X_{ig}^{(c)}) = \sigma_g^2,$$

Null hypothesis to test:

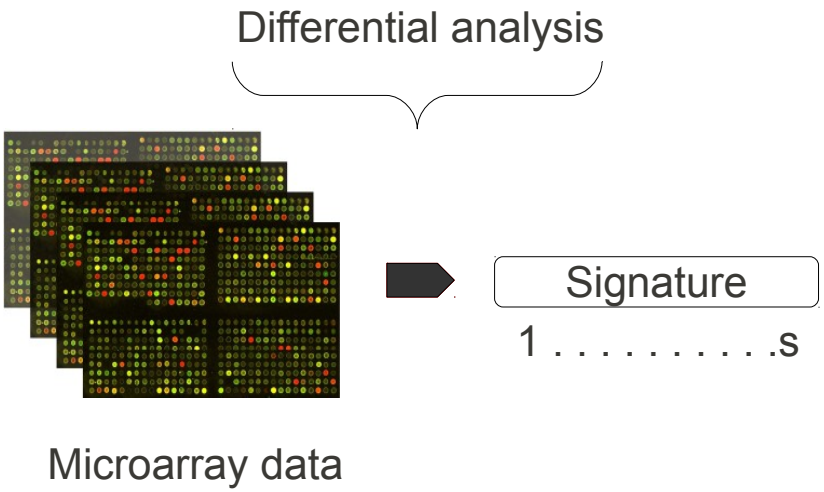
$$\begin{cases} H_0 : & \mu_g^{(1)} = \mu_g^{(2)}, \\ H_1 : & \mu_g^{(1)} \neq \mu_g^{(2)}. \end{cases}$$

Limma t-statistic (Smyth 2004)

$$t_g^{\text{limma}} = \frac{\bar{x}_{.g}^{(1)} - \bar{x}_{.g}^{(2)}}{s_g^{\text{limma}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

- s_g^{limma}: Bayesian estimator of the variance
- Stabilize the estimation of gene variances

Method - Summary



Method - Biological prior definition

How to interpret gene signatures in biologically meaningful terms ?

$\rightsquigarrow$ by determining whether the signature is enriched in pathway* key actors.

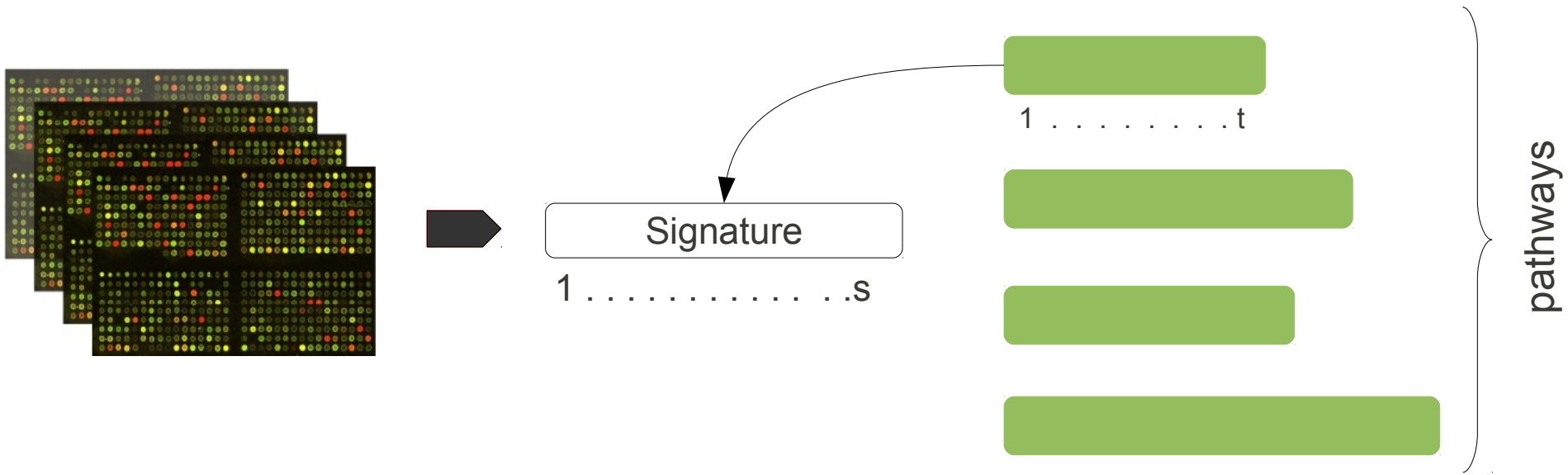
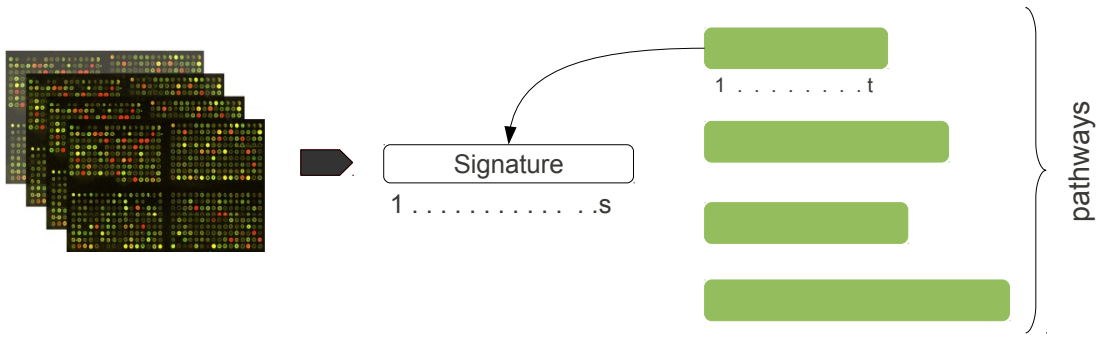


Figure: Group testing for pathway analysis

* Pathway: set of gene interacting in order to achieve a specific cellular function

Method - Biological prior definition



Under the null hypothesis of no over-representation

$$\begin{aligned}\mathbb{P}(Y \geq y) &= 1 - \mathbb{P}(Y \leq y) \\ &= 1 - \sum_{i=0}^{y} \frac{\binom{s}{i}\binom{p-s}{t-i}}{\binom{p}{t}}.\end{aligned}$$

$\mathbb{P}(Y \geq y)$ probability of **observing at least *y* genes** of a **pathway of size *t*** in the signature

Method - Biological prior definition

In practice...

Pathway Name	Genes in pathway
HER-2 Signaling in Breast Cancer	**CCNE1**,**CDK6**,PARD6B,ERBB3,**EGFR**
Glioblastoma Multiforme Signaling	**CCNE1**,RHOB,IGF1R,**CDK6**,**EGFR**
Estrogen-Dependent Breast Cancer Signaling	IGF1R,ESR1,**EGFR**
Small Cell Lung Cancer Signaling	**CCNE1**,**CDK6**,BCL2
Aryl Hydrocarbon Receptor Signaling	**CCNE1**,TFF1,**CDK6**,ESR1

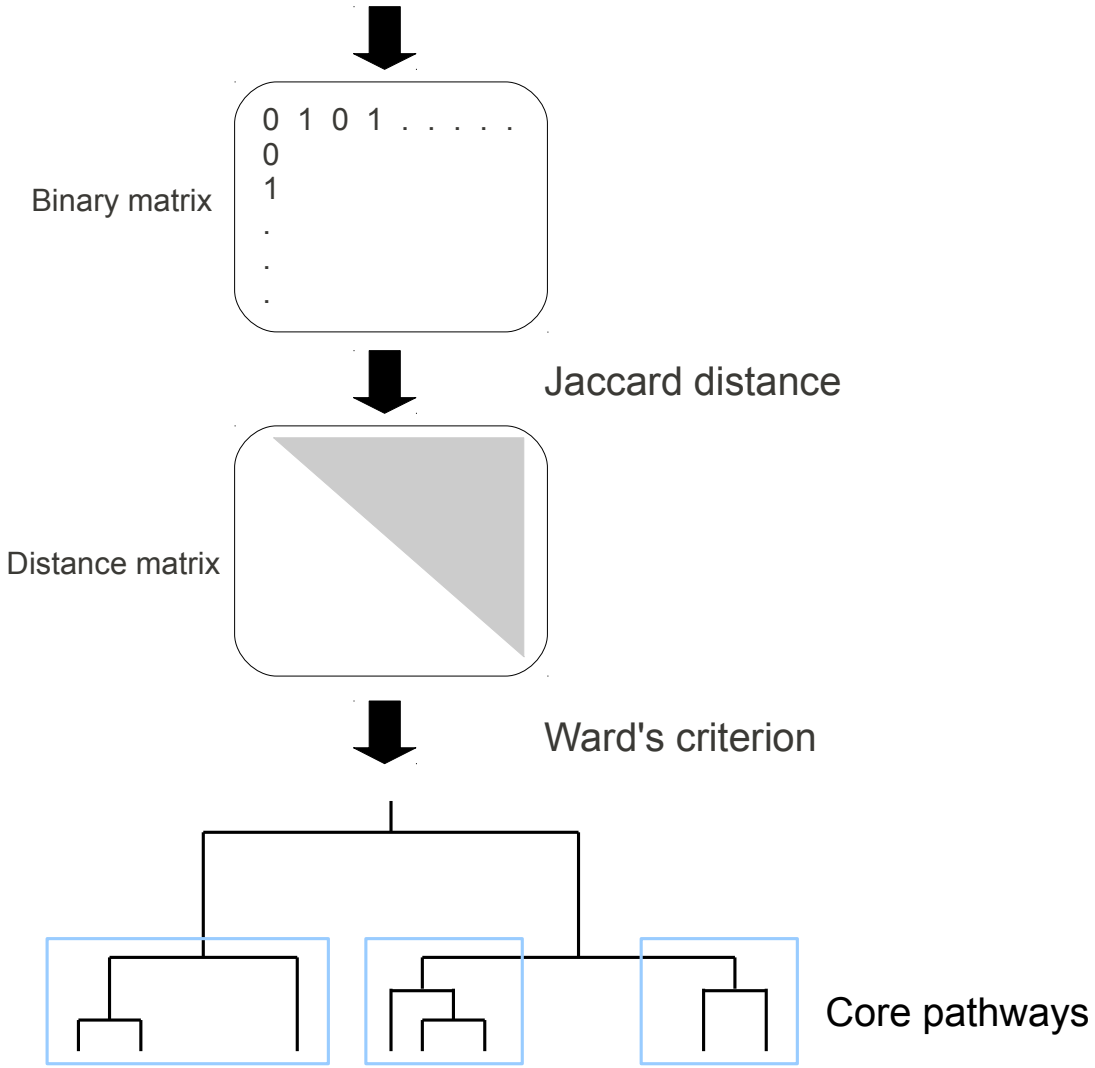
Table: Results of pathway analysis

- Pathways do not clearly represent distinct entities !

$\rightsquigarrow$ we need to summarize the set of pathways found significant

Method - Biological prior definition

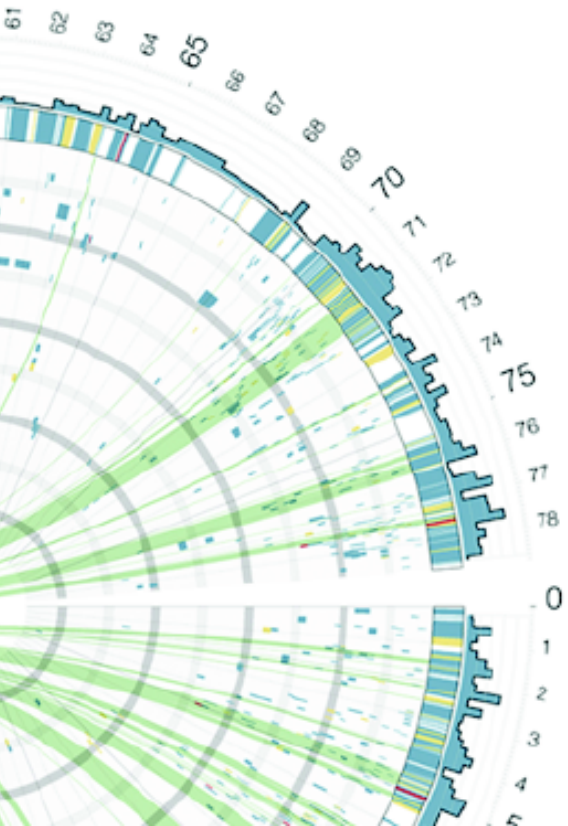
Pathway Name	Genes in pathway
HER-2 Signaling in Breast Cancer	**CCNE1**,**CDK6**,PARD6B,ERBB3,**EGFR**
Glioblastoma Multiforme Signaling	**CCNE1**,RHOB,IGF1R,**CDK6**,**EGFR**
Estrogen-Dependent Breast Cancer Signaling	IGF1R,ESR1,**EGFR**
Small Cell Lung Cancer Signaling	**CCNE1**,**CDK6**,BCL2
Aryl Hydrocarbon Receptor Signaling	**CCNE1**,TFF1,**CDK6**,ESR1



Method - Summary

Method

Network Inference



R package SIMoNe : general settings

- Enables inference of undirected networks:
  - In a Gaussian graphical models (GGM) framework
  - Multitask inference strategy: joint estimation of the graphs by coupling the estimation problems

- Based on partial correlation coefficients

Chiquet et al. 2010,
Inferring Multiple Graphical Models.
Statistics and Computing

Graphical model

Def.: *Probabilistic model for which a graph denotes the conditional independence structure between random variables.*

Gaussian model for an i.i.d. sample

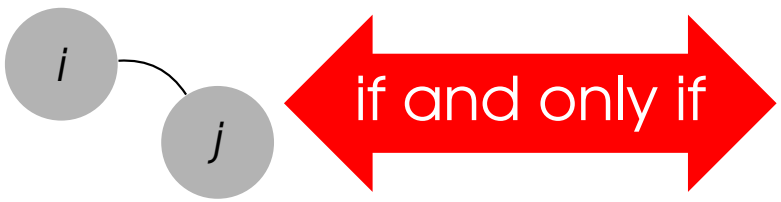
- Let $\mathcal{P} = \{1, ..., p\}$ be a set of nodes (i.e. genes)
- $X = (X_1, ...X_p)^T$ is the signal over this set (i.e. the gene expression levels), such as: $X \sim \mathcal{N}(\mathbf{0}_p, \Sigma)$
- Let Θ be the parameter to be inferred (i.e. the edges)
  - $\Theta = (\theta_{ij})_{i,j \in \mathcal{P}} \triangleq \mathbf{\Sigma}^{-1}$ is the concentration matrix.
  - $\mathrm{cor}_{ij|\mathcal{P}\setminus\{i,j\}} = -\theta_{ij}/\sqrt{\theta_{ii}\theta_{jj}}$ for $i \neq j$

Interpretation

If 2 nodes *i* and *j* are partially uncorrelated, no edge is inferred:

$$X_i \perp\!\!\!\perp X_j | X(\mathcal{P} \backslash \{i,j\}) \Leftrightarrow \theta_{ij} = 0$$

After a simple rescaling Θ can be interpreted as the adjacency matrix



conditional dependency
or
non null partial correlation between

Method - Network Inference

Let $\mathbf{S} = n^{-1}\mathbf{X}^T\mathbf{X}$ be the empirical variance-covariance matrix.

- $\mathbf{S}^{-1}$ is not defined for $n < p$.
- If $n < p$, neither Θ nor its support can be estimated
- The need for regularization is huge

Estimation: a penalized likelihood approach

$$\hat{\Theta}_\lambda = \arg\max_{\Theta} \mathcal{L}(\Theta; \text{data}) - \lambda \, \text{pen}_{\ell_1}(\Theta),$$

- $\mathcal{L}$ is the model log-likelihood,
- $\text{pen}_{\ell_1} = ||\Theta||_{\ell_1}$ is a penalty function tuned by $\lambda > 0$.

It performs:

1. regularization (needed when n $\ll$ p),
2. selection (sparsity induced by the ℓ_1-norm)

Method - Network Inference

Take into account the core-pathways information as an *a-priori* knowledge:

$\rightsquigarrow$ Edges between two genes of the same core-pathway are less penalized

Statistical approach

Use adaptive penalty parameters for different coefficients

- Let Z be the set of indicator variable for nodes

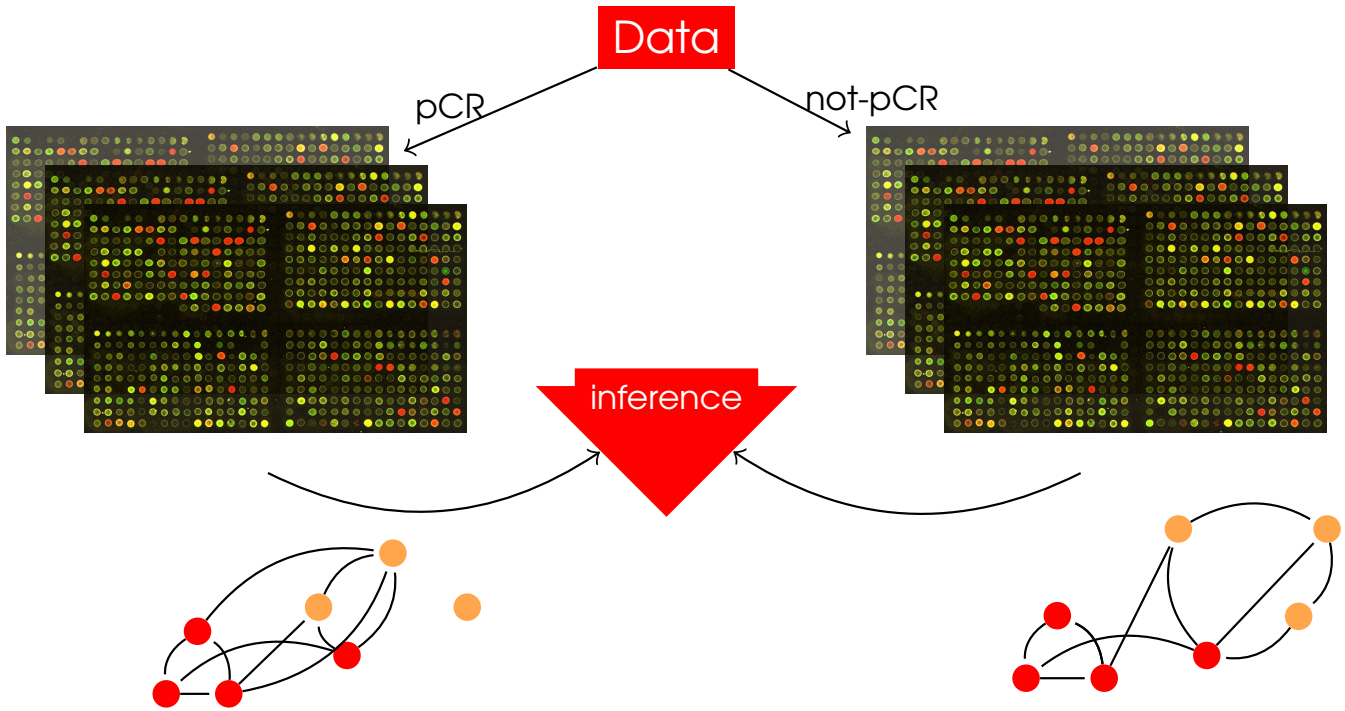
$$\hat{\Theta}_\lambda = \arg\max_{\Theta} \mathcal{L}(\Theta; \text{data}) - \lambda \|\mathbf{P_Z} \star \Theta\|_{\ell_1},$$

where $\mathbf{P_Z}$ is a matrix of weights depending on the core-pathway membership $\mathbf{Z}$.

Multitask inference

⇝ How to deal with various conditions ?

- Assumption: strong relationship between both networks
- Approach: joint estimation of the graphs by coupling the estimation problems



Chiquet et al. 2010, Inferring Multiple Graphical Models
Statistics and Computing

Consider C conditions where the same p genes are measured

Graphical coop-LASSO

$$\max_{\boldsymbol{\Theta}^{(c)}} \sum_{c=1}^{C} \mathcal{L}\left(\Theta^{(c)}; \text{data}\right) - \lambda \sum_{\substack{i,j \in \mathcal{P} \\ i \neq j}} \left\{ \left(\sum_{c=1}^{C} \left[\theta_{ij}^{(c)}\right]_+^2 \right)^{1/2} + \left(\sum_{c=1}^{C} \left[\theta_{ij}^{(c)}\right]_-^2 \right)^{1/2} \right\},$$

where $[u]_+ = \max(0, u)$ and $[u]_- = \min(0, u)$.

- Group-lasso like penalty
- Disconnect the activation of up and down regulation

- $\mathcal{Q} = \{1, \ldots, Q\}$ of given overlapping core-pathways
- $Z_{iq} = 1$ if $i \in q$ and 0 otherwise

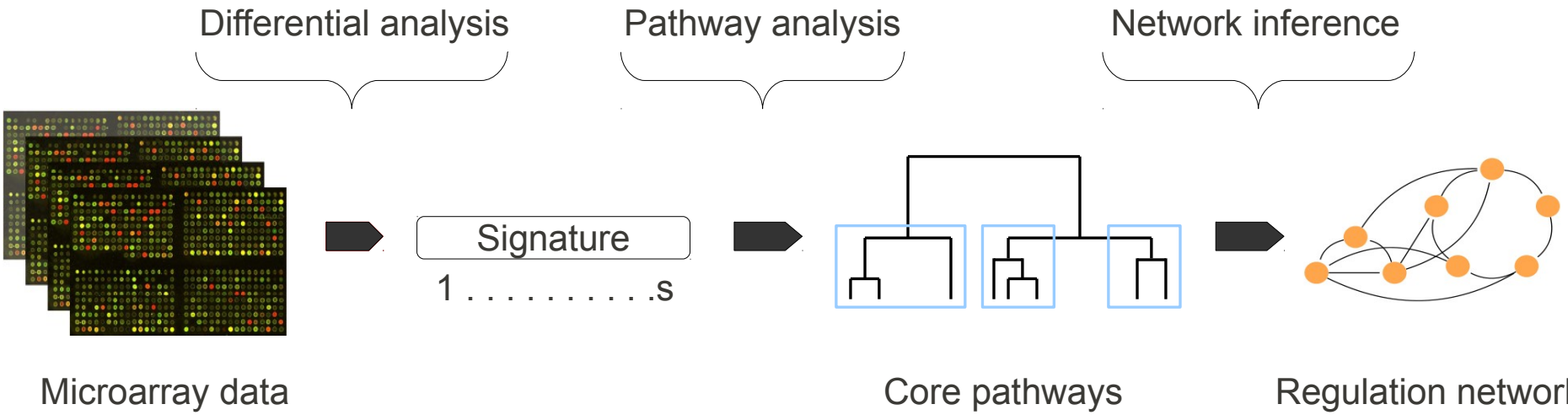
Maximisation Problem

$$\max_{\theta^{(c)}} \sum_{c=1}^{C} \mathcal{L}\left(\Theta^{(c)}; \text{data}\right) - \lambda \sum_{\substack{i,j \in \mathcal{P} \\ i \neq j}} \rho_{\mathbf{z}_i \mathbf{z}_j} \left\{ \left(\sum_{c=1}^{C} \left[\theta_{ij}^{(c)}\right]_+^2 \right)^{1/2} + \left(\sum_{c=1}^{C} \left[\theta_{ij}^{(c)}\right]_-^2 \right)^{1/2} \right\}, \quad (1)$$

where $[u]_+ = \max(0, u)$ and $[u]_- = \min(0, u)$ and the coefficients of the penalty are defined as:

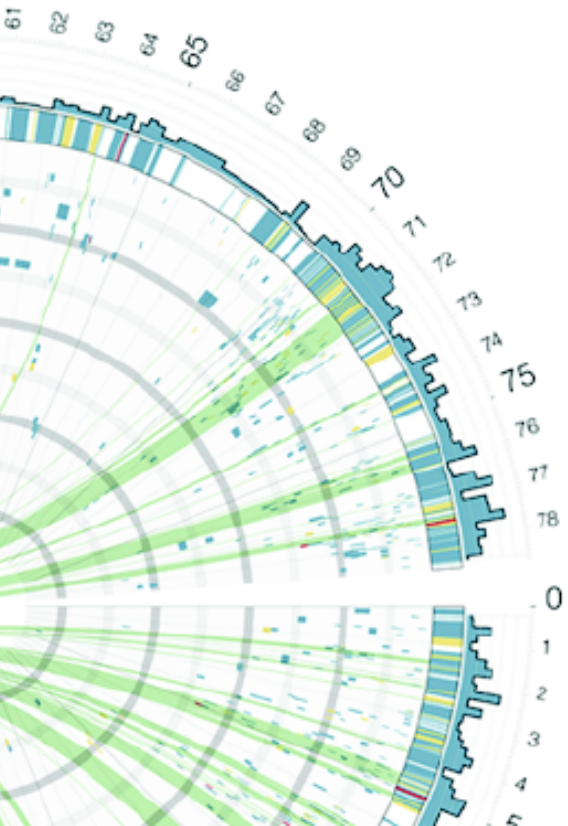
$$\rho_{\mathbf{z}_i \mathbf{z}_j} = \begin{cases} \displaystyle\sum_{q,\ell \in \mathcal{Q}} Z_{iq} Z_{j\ell} \frac{1}{\lambda_{\text{in}}}, & \text{if } i \neq j, \text{ and } q = \ell, \\ \displaystyle\sum_{q,\ell \in \mathcal{Q}} Z_{iq} Z_{j\ell} \frac{1}{\lambda_{\text{out}}}, & \text{if } i \neq j, \text{ and } q \neq \ell, \\ 1, & \text{otherwise.} \end{cases} \quad (2)$$

Method - Summary



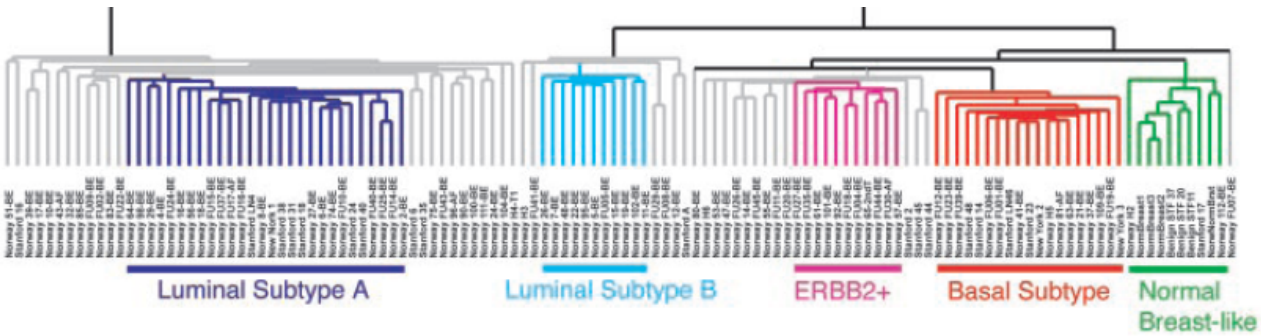
Application

ER status in breast cancer



Breast cancer in a few words

- An heterogeneous disease (5 subtypes)

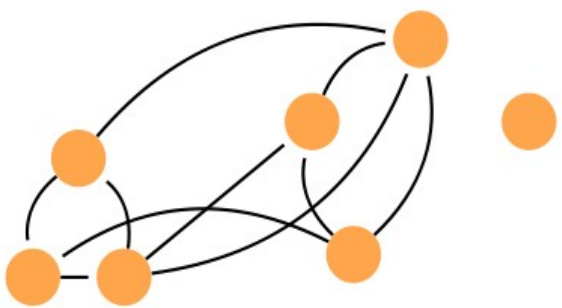


- Presence (ER+)/absence (ER-) of estrogen receptors: an essential parameter of tumor characterization.

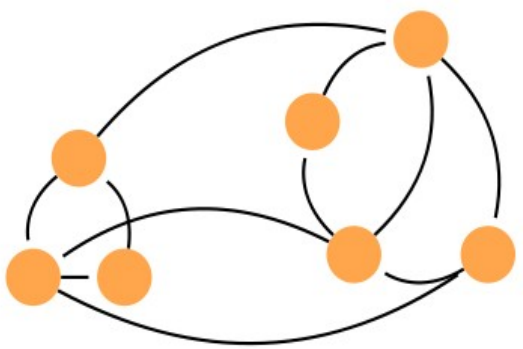
$\rightsquigarrow$ Understanding the molecular mechanism of ER status: a key issue for treatment and prognosis

ER status in breast cancer

Inference of regulation networks under ER+ and ER- conditions



ER +



ER -

- Comparison of regulation patterns

ER status in breast cancer

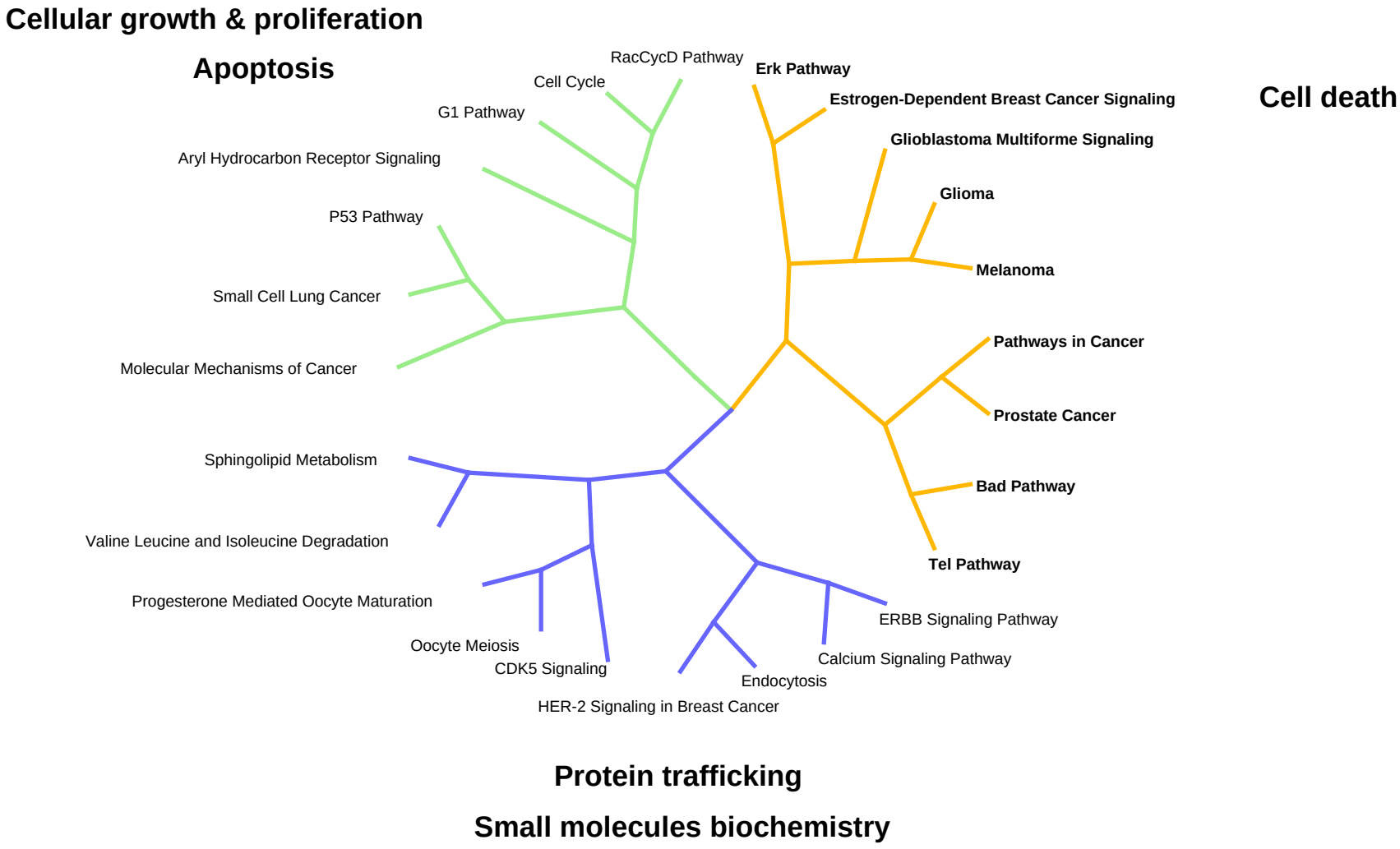


Figure: Core pathways

ER status in breast cancer

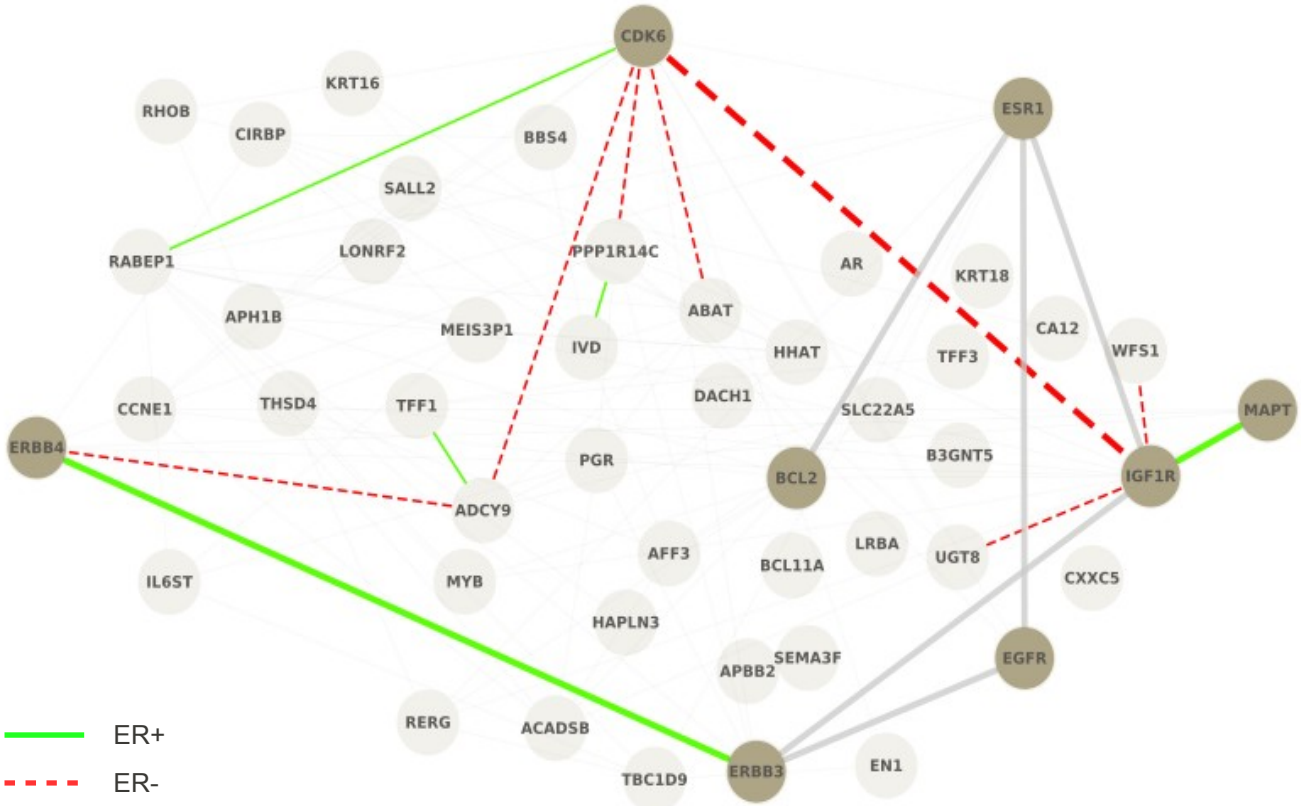
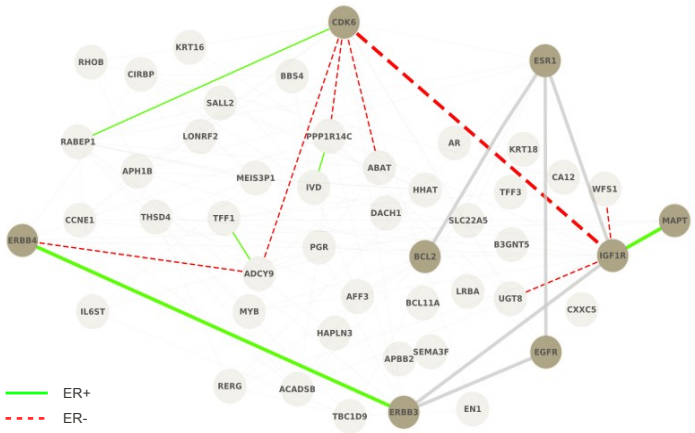


Figure: Sub-network inferred from the ER status signature

Anti-apoptotic mechanisms



Common regulations

Estrogen receptor (ESR1) - BCL2 (Peterson at *al.* 2007)

ESR1 - EGFR/IGF1R (Salvatori et *al.* 2000, Oesterreich et *al.* 2001)

Specific regulations

EGF receptor family: ERBB3 - ERBB4 (Lee et *al.* 2001)

CDK6 - IGF1R

ER status in breast cancer

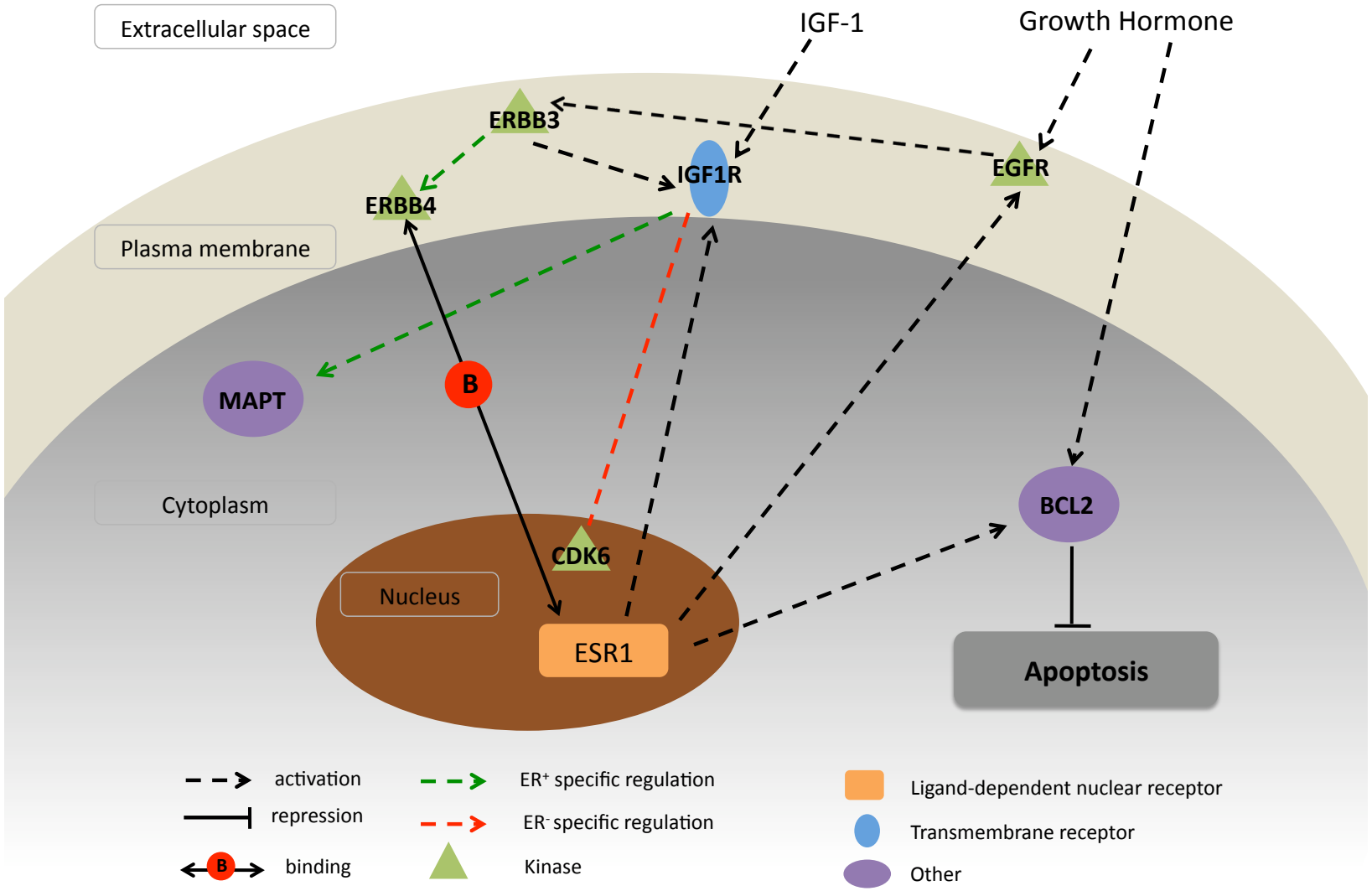


Figure: Anti-apoptotic mechanisms

Summary

- Very challenging issue
- Introducing biological priors reduce the space of possible networks
- Promising application on Breast cancer dataset
- Importance of missing covariates

$\rightsquigarrow$ Persepectives: need for integration of heterogeneous omics data.

Acknowledgments

J. Chiquet

C. Charbonnier

M. Guedj

C. Ambroise

Acknowledgements

Statistic & Genome Laboratory

Christophe Ambroise

Julien Chiquet

Bernard Prum

Michèle, Carène, Claudine, Catherine, Camille, Etienne, Pierre, Gilles, Cecile, Maurice, Marie-Luce, Anne-Sophie, Cyril, Justin, Van-Hanh, Yolande, Sarah, Marius.

Jan, Caroline, Fabrice, Mickaël, Matthieu.

Pharnext

Mickaël Guedj

Serguei Nabirotchkin

Ilya Chumakov