Sparse Gaussian graphical models for biological network inference

Julien Chiquet

Statistique et Génome, UMR CNRS 8071 - Université d'Évry Val-d'Essonne

Invited researcher at Statistique et Génome, UMR INRA 518 - AgroParisTech

NetBio – September the 12th, 2013

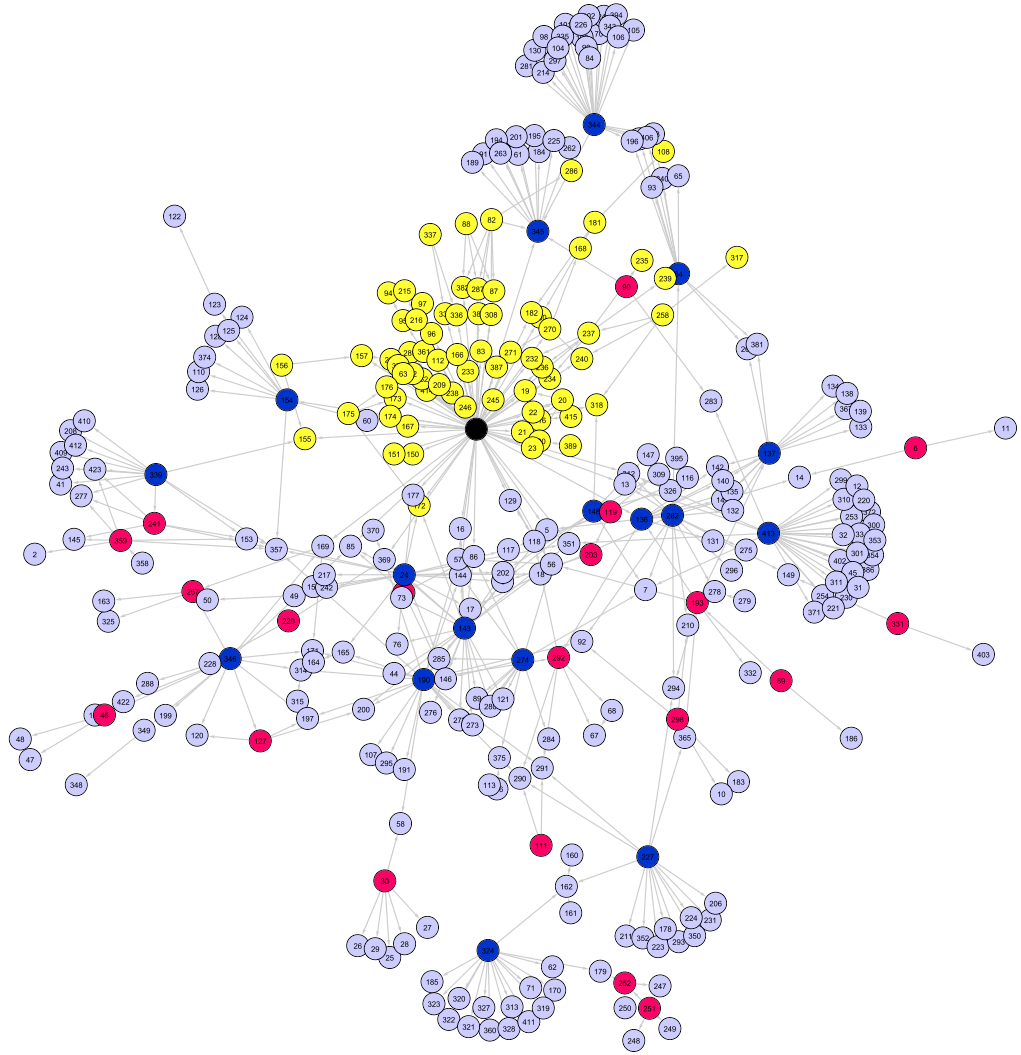
Outline

What the reconstructed networks are expected to be [1] (1)

Regulatory networks

E. coli regulatory network

- relationships between gene and their products
- inhibition/activation
- impossible to recover at large scale
- always incomplete



1

[1] and are presumably *wrongly* assumed to be

What the reconstructed networks are expected to be (2)

Regulatory networks

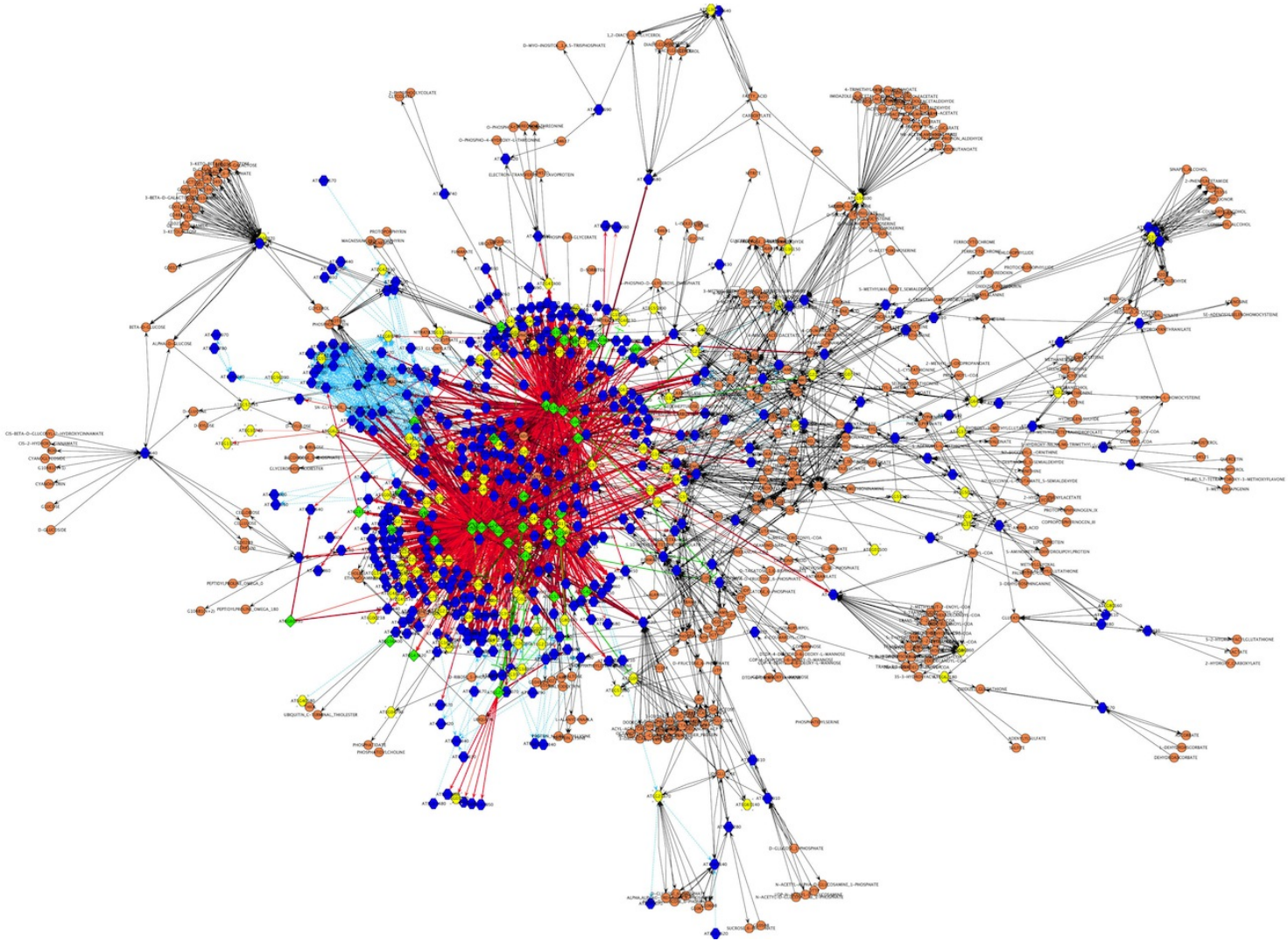


Figure: Regulatory network identified in mammalian cells: highly structured

What the reconstructed networks are expected to be (3)

Protein-Protein interaction networks

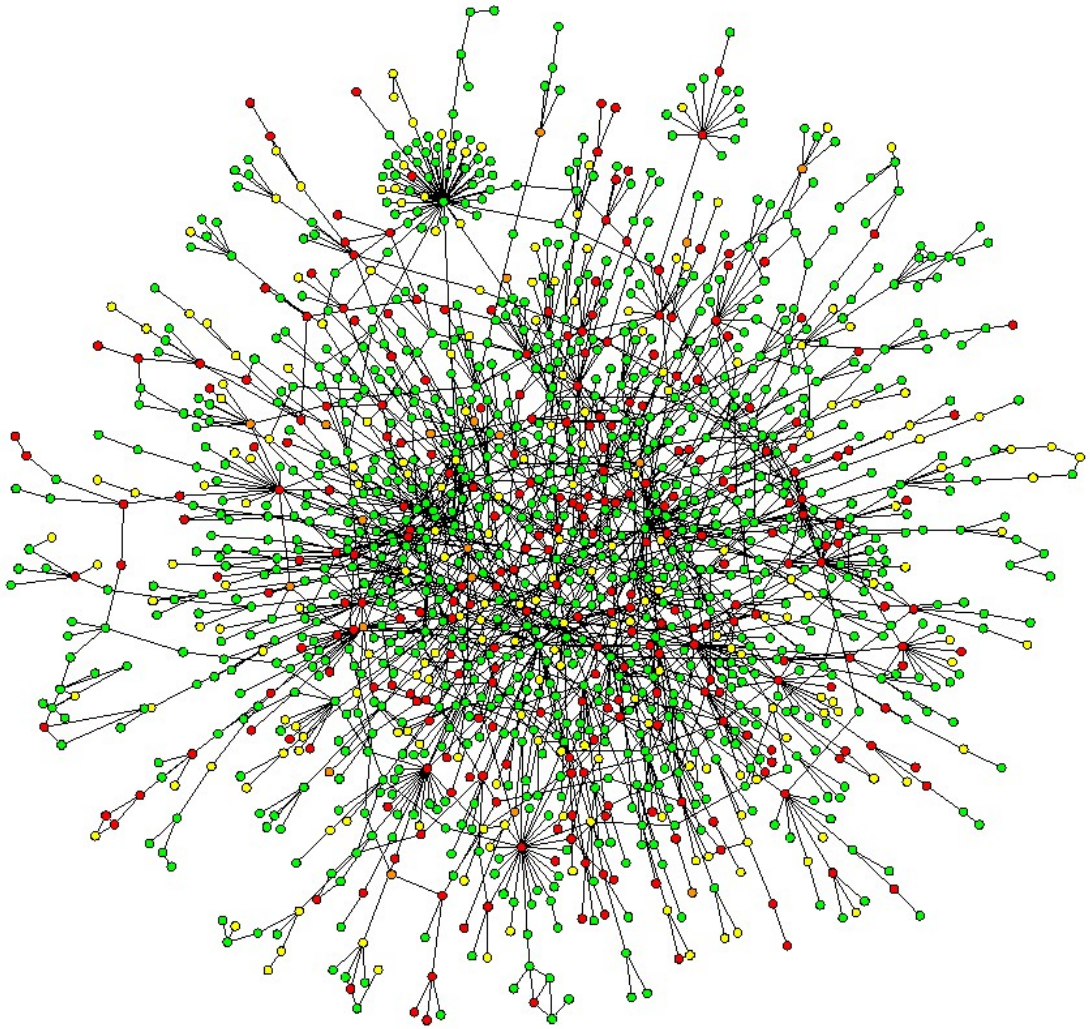


Figure: Yeast PPI network : do not be mislead by the representation, trust stat !

What the reconstructed networks are expected to be (3)

Protein-Protein interaction networks

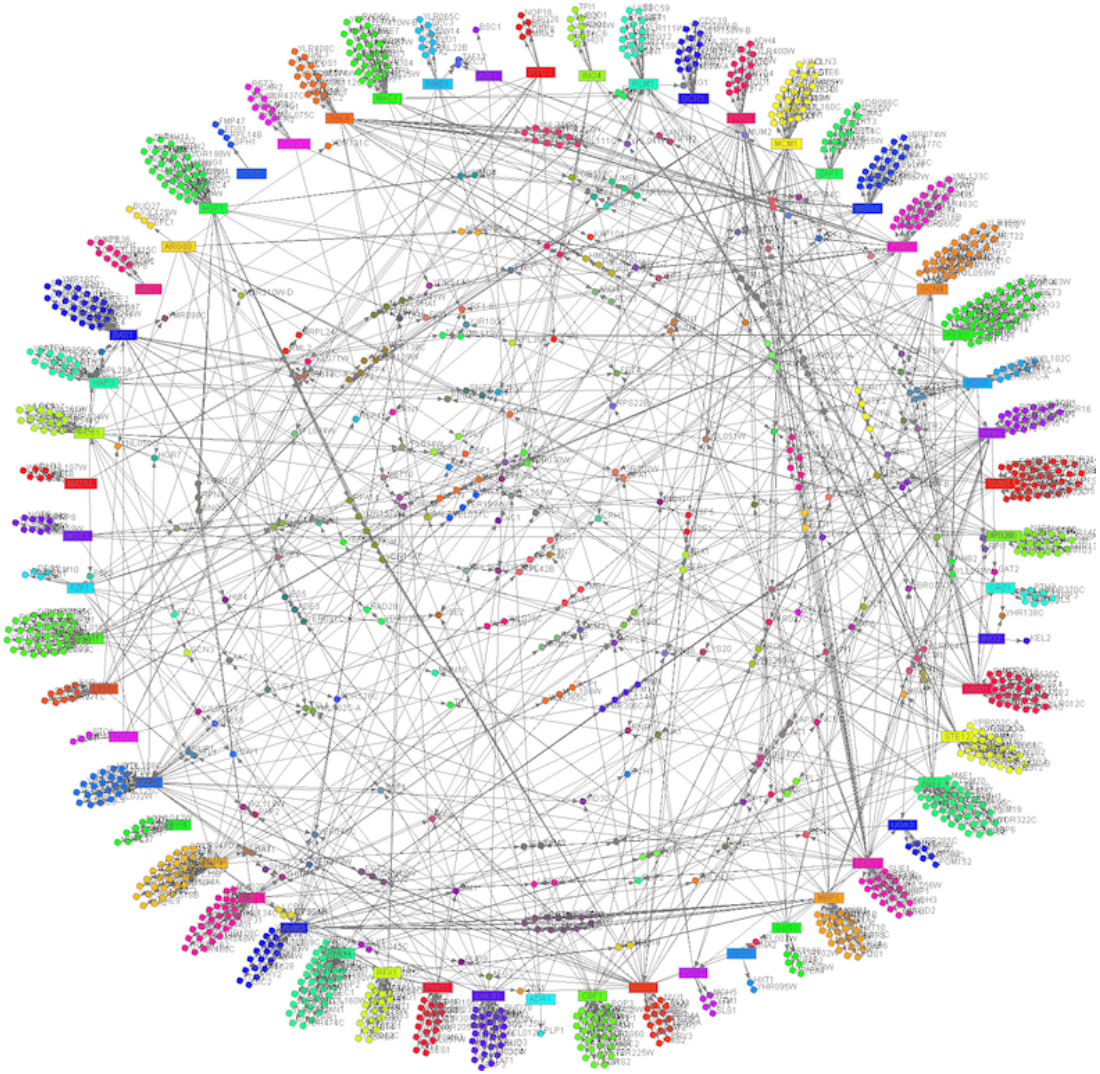


Figure: Yeast PPI network : do not be mislead by the representation, trust stat !

What the reconstructed networks are expected to be (3)

Protein-Protein interaction networks

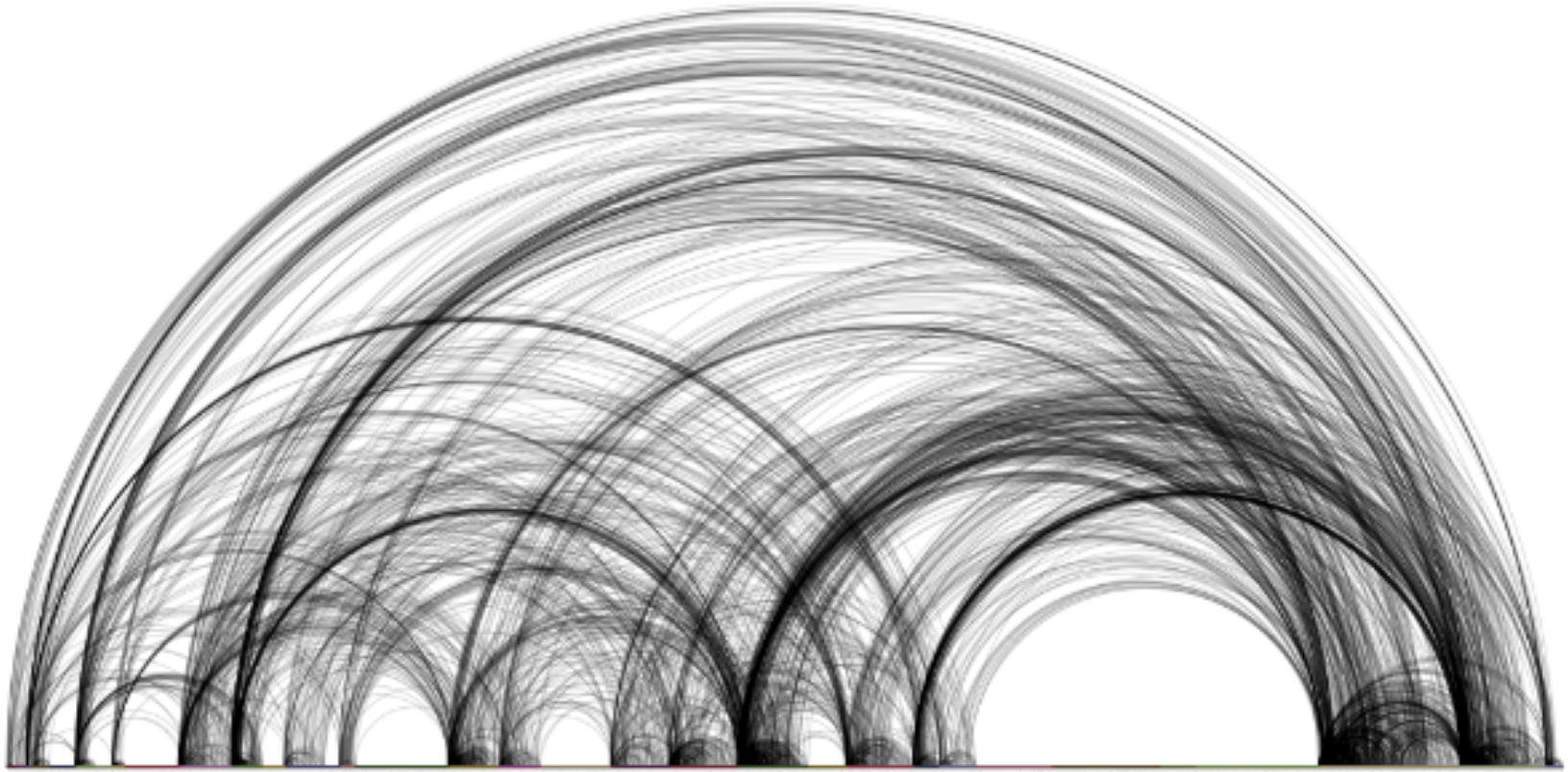


Figure: Yeast PPI network : do not be mislead by the representation, trust stat !

Why caring about network inference?

Unraveling significant interactions at large scale is impossible "manually".

Exploratory research

Assist the biologist by

- pointing important molecules/pathways in a organism,
- giving further insight about the regulatory mechanisms,
- elucidation of gene/protein functions,

$\rightsquigarrow$ It helps at formulating a hypothesis for further wet lab experiment.

Why caring about network inference?

Unraveling significant interactions at large scale is impossible "manually".

May plausibly

help to understand the mechanisms of complex diseases or treatments.

- pointing important molecules/pathways in a organism,
- giving further insight about the regulatory mechanisms

Does not (and I do not think it will in close future)

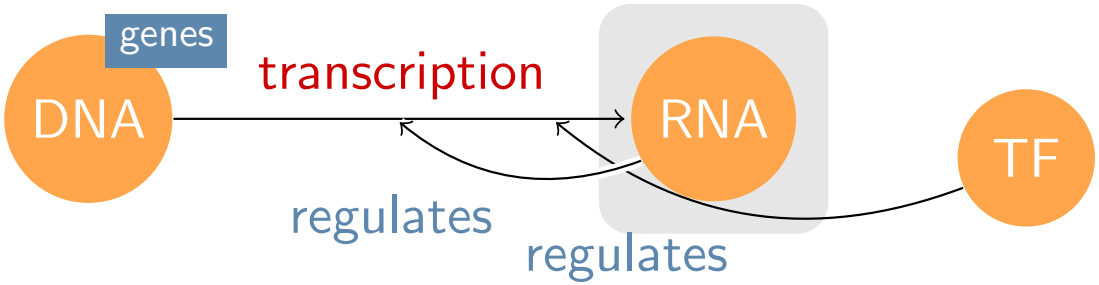
reconstruct a trustful regulatory network at large scale.

$\rightsquigarrow$ It helps at formulating a hypothesis for further wet lab experiment.

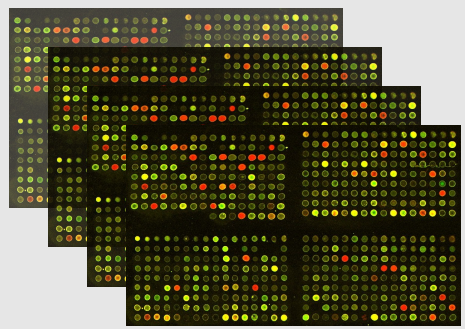
How is this measured?

Microarray technology: parallel measurement of many biological features

Focus e.g. on *transcription*, looking toward *gene regulatory networks*



signal processing



Matrix of features $n \ll p$

Expression levels of p probes are simultaneously monitored for n individuals

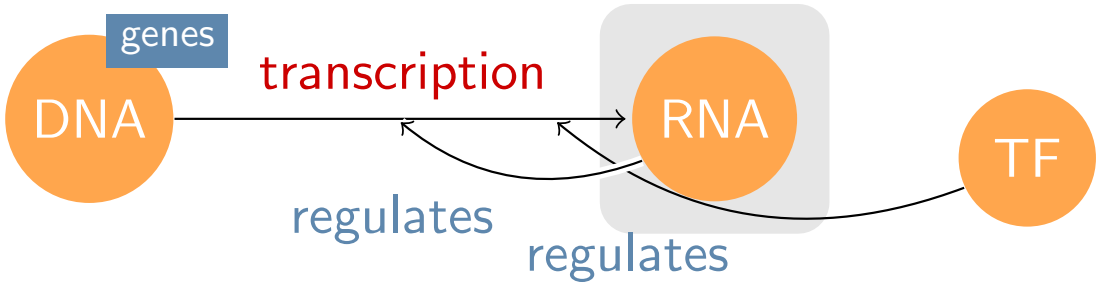
pretreatment

$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & x_1^3 & \dots & x_1^p \\ \vdots & & & & \\ x_n^1 & x_n^2 & x_1^2 & \dots & x_n^p \end{pmatrix}$$

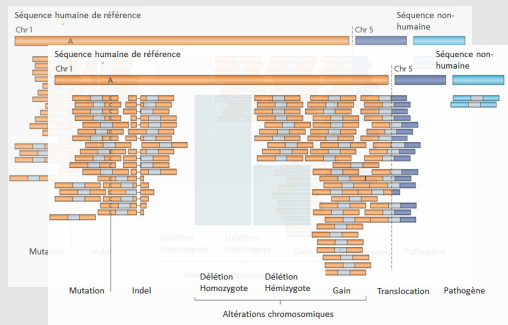
How is this measured?

Next Generation Sequencing: parallel measurement of even many more biological features

Focus e.g. on *transcription*, looking toward *gene regulatory networks*



assembling



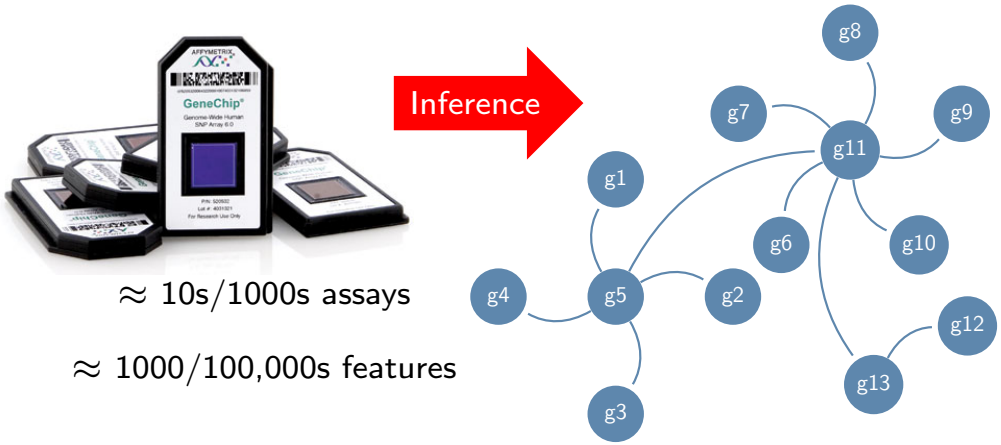
pretreatment

Matrix of features $n \lll p$

Expression counts are extracted from small repeated sequences and monitored for n individuals

$$\mathbf{X} = \begin{pmatrix} k_1^1 & k_1^2 & k_1^3 & \dots & k_1^p \\ \vdots & & & & \\ k_n^1 & k_n^2 & k_1^2 & \dots & k_n^p \end{pmatrix}$$

Summary of the problem at hand



≈ 10s/1000s assays

≈ 1000/100,000s features

1. Nodes (genes) are fixed
   - restricted to a set of interest (e.g., TF/target or via DA)

   Q: *what if we missed some relevant actors?*

2. Edges (regulations) are inferred
   - based upon statistical concepts

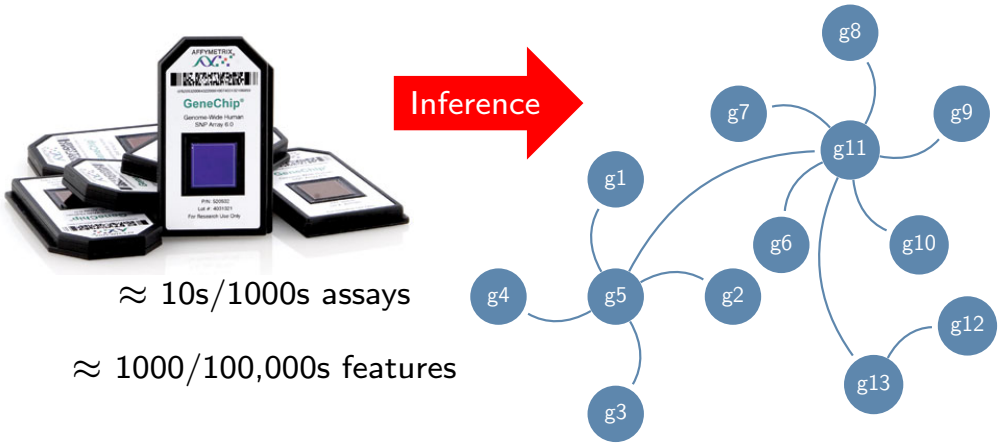
   Q: *biological relevance?*

Main statistical challenges

1. Ultra high dimensionality ($n \lll p$),
2. Heterogeneity of the data (noise, many techniques/signals/scales).

$\rightsquigarrow$ Omic data is hopefully structured in many ways.

Summary of the problem at hand



$\approx$ 10s/1000s assays

$\approx$ 1000/100,000s features

1. Nodes (genes) are fixed
   - restricted to a set of interest (e.g., TF/target or via DA)

   Q: *what if we missed some relevant actors?*
2. Edges (regulations) are inferred
   - based upon statistical concepts

   Q: *biological relevance?*

Main statistical challenges

1. ~~Ultra~~ high dimensionality ($n < p$),
2. Heterogeneity of the data (noise, many techniques/signals/scales).

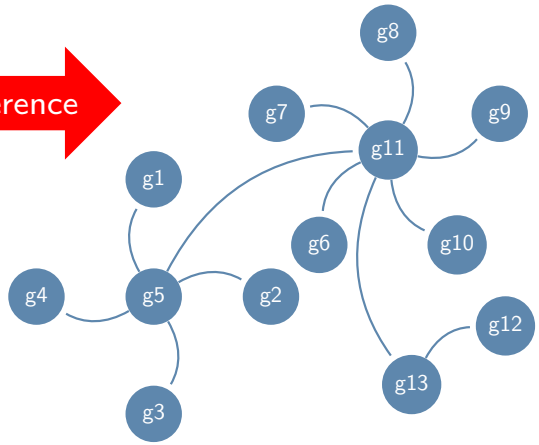
$\rightsquigarrow$ Omic data is hopefully structured in many ways.

Summary of the problem at hand

≈ 10s/1000s assays

≈ 1000/100,000s features

Inference



1. Nodes (genes) are fixed
   - restricted to a set of interest (e.g., TF/target or via DA)

   Q: *what if we missed some relevant actors?*
2. Edges (regulations) are inferred
   - based upon statistical concepts

   Q: *biological relevance?*

Main statistical challenges

1. high dimensionality ,
2. Heterogeneity of the data (noise, many techniques/signals/scales).

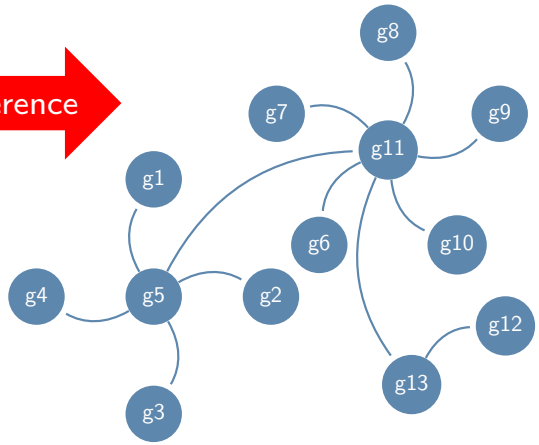
$\rightsquigarrow$ Omic data is hopefully structured in many ways.

Summary of the problem at hand

≈ 10s/1000s assays

≈ 1000/100,000s features

Inference



1. Nodes (genes) are fixed
   - restricted to a set of interest (e.g., TF/target or via DA)

   Q: *what if we missed some relevant actors?*
2. Edges (regulations) are inferred
   - based upon statistical concepts

   Q: *biological relevance?*

Main statistical challenges

1. high dimensionality ,
2. Heterogeneity of the data (noise, many techniques/signals/scales).

$\rightsquigarrow$ Omic data is hopefully structured in many ways.

Outline

Gaussian Graphical Model: canonical settings

Microarrays in comparable Gaussian conditions

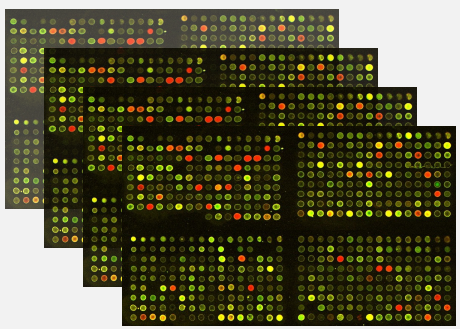
Profiles of a set $\mathcal{P} = \{1, \ldots, p\}$ of genes is described by $X \in \mathbb{R}^p$ such as

1. $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ the precision matrix.
2. a sample $(X^1, \ldots, X^n)$ of chips stacked in an $n \times p$ data matrix $\mathbf{X}$.

Conditional independence structure

The data

Stacking $(X^1, \ldots, X^n)$, we met the usual individual/variable table $\mathbf{X}$



stacked in

$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & x_1^3 & \ldots & x_1^p \\ \vdots & & & & \\ x_n^1 & x_n^2 & x_1^2 & \ldots & x_n^p \end{pmatrix}$$

$\rightsquigarrow$ "Covariance" selection

Gaussian Graphical Model: canonical settings

Microarrays in comparable Gaussian conditions

Profiles of a set $\mathcal{P} = \{1, \ldots, p\}$ of genes is described by $X \in \mathbb{R}^p$ such as

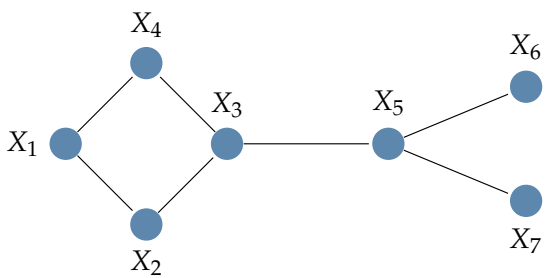
1. $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ the precision matrix.
2. a sample $(X^1, \ldots, X^n)$ of chips stacked in an $n \times p$ data matrix $\mathbf{X}$.

Conditional independence structure

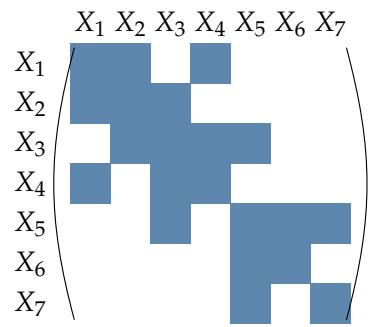
$$(i,j) \notin \mathcal{E} \Leftrightarrow X_i \perp\!\!\!\perp X_j | X_{\setminus\{i,j\}} \Leftrightarrow \Theta_{ij} = 0.$$

Graphical interpretation

$$\mathcal{G} = (\mathcal{P}, \mathcal{E})$$



$\boldsymbol{\Theta}$



$\rightsquigarrow$ "Covariance" selection

Gaussian Graphical Model and Linear Regression

Linear regression viewpoint

Gene expression X_i is linearly explained by the other genes':

$$X_i | X_{\setminus i} = -\sum_{j \neq i} \frac{\Theta_{ij}}{\Theta_{ii}} X_j + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i), \quad \varepsilon_i \perp X$$

Conditional on its neighborhood, other profiles do not give additional insights

$$X_i | X_{\setminus i} = \sum_{j \in \text{neighbors}(i)} \beta_j X_j + \varepsilon_i \quad \text{with } \beta_j = -\frac{\Theta_{ij}}{\Theta_{ii}}.$$

Gaussian Graphical Model and Linear Regression

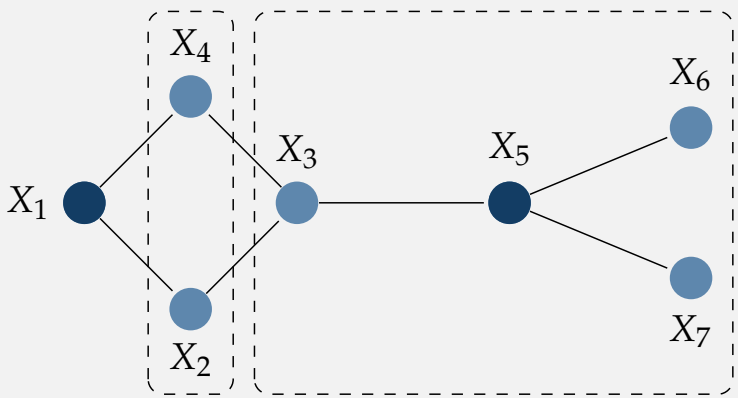
Linear regression viewpoint

Gene expression X_i is linearly explained by the other genes':

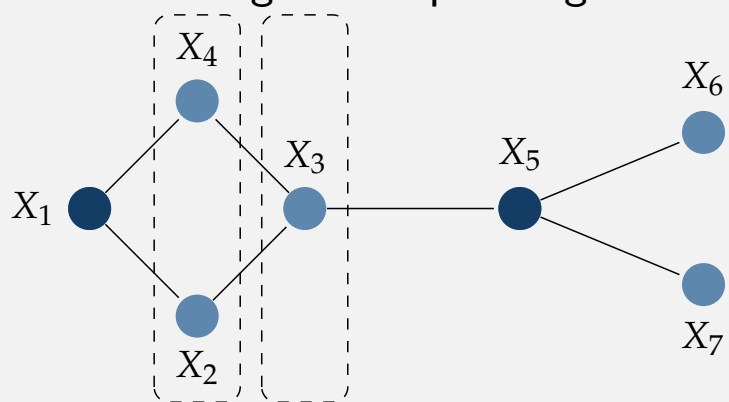
$$X_i | X_{\setminus i} = -\sum_{j \neq i} \frac{\Theta_{ij}}{\Theta_{ii}} X_j + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i), \quad \varepsilon_i \perp X$$

Graphical Interpretation

Local Markov property
conditioning on the neighborhood



Global Markov property
conditioning on a separating node



Gold standard penalized approaches

Use ℓ_1 for both regularizing and promoting *sparsity*

Penalized likelihood (Banerjee *et al.*, Yuan and Lin, 2008)

$$\hat{\mathbf{\Theta}}_\lambda = \arg\max_{\mathbf{\Theta}\in\mathbb{S}_+} \ell(\mathbf{\Theta};\mathbf{X}) - \lambda\|\mathbf{\Theta}\|_1$$

+ symmetric, positive-definite
– solved by the "Graphical-Lasso" ($\mathcal{O}(p^3)$, *Friedman et al, 2007*).

Neighborhood Selection (Meinshausen & Bülhman, 2006)

$$\widehat{\boldsymbol{\beta}}^{(i)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \frac{1}{n}\left\|\mathbf{X}_i - \mathbf{X}_{\setminus i}\boldsymbol{\beta}\right\|_2^2 + \lambda\left\|\boldsymbol{\beta}\right\|_1$$

CLIME – Pseudo-likelihood (Cai et al., 2011; Yuan, 2010)

$$\widehat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \|\mathbf{\Theta}\|_1 \text{ subjected to } \left\|n^{-1}\mathbf{X}^t\mathbf{X}\mathbf{\Theta} - \mathbf{I}\right\|_\infty \leq \lambda$$

Gold standard penalized approaches

Use ℓ_1 for both regularizing and promoting *sparsity*

Penalized likelihood (Banerjee *et al.*, Yuan and Lin, 2008)

$$\hat{\mathbf{\Theta}}_\lambda = \arg\max_{\mathbf{\Theta}\in\mathbb{S}_+} \ell(\mathbf{\Theta};\mathbf{X}) - \lambda\|\mathbf{\Theta}\|_1$$

Neighborhood Selection (Meinshausen & Bülhman, 2006)

$$\widehat{\boldsymbol{\beta}}^{(i)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \frac{1}{n}\left\|\mathbf{X}_i - \mathbf{X}_{\setminus i}\,\boldsymbol{\beta}\right\|_2^2 + \lambda\,\|\boldsymbol{\beta}\|_1$$

- not symmetric, not positive-definite
+ p Lasso solved with Lars-like algorithms ($\mathcal{O}(npd)$ for d neighbors).

CLIME – Pseudo-likelihood (Cai et al., 2011; Yuan, 2010)

$$\widehat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \|\mathbf{\Theta}\|_1 \text{ subjected to } \left\|n^{-1}\mathbf{X}^t\mathbf{X}\mathbf{\Theta} - \mathbf{I}\right\|_\infty \leq \lambda$$

Gold standard penalized approaches

Use ℓ_1 for both regularizing and promoting *sparsity*

Penalized likelihood (Banerjee *et al.*, Yuan and Lin, 2008)

$$\hat{\mathbf{\Theta}}_\lambda = \arg\max_{\mathbf{\Theta}\in\mathbb{S}_+} \ell(\mathbf{\Theta};\mathbf{X}) - \lambda\|\mathbf{\Theta}\|_1$$

Neighborhood Selection (Meinshausen & Bülhman, 2006)

$$\widehat{\boldsymbol{\beta}}^{(i)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \frac{1}{n}\left\|\mathbf{X}_i - \mathbf{X}_{\setminus i}\,\boldsymbol{\beta}\right\|_2^2 + \lambda\left\|\boldsymbol{\beta}\right\|_1$$

CLIME – Pseudo-likelihood (Cai et al., 2011; Yuan, 2010)

$$\widehat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \|\mathbf{\Theta}\|_1 \text{ subjected to } \left\|n^{-1}\mathbf{X}^t\mathbf{X}\mathbf{\Theta} - \mathbf{I}\right\|_\infty \leq \lambda$$

– not positive-definite

\+ p linear programs easily distributed ($\mathcal{O}(p^2 d)$ for d neighbors).

Gold standard penalized approaches

Use ℓ_1 for both regularizing and promoting *sparsity*

Penalized likelihood (Banerjee *et al.*, Yuan and Lin, 2008)

$$\hat{\mathbf{\Theta}}_\lambda = \underset{\mathbf{\Theta} \in \mathbb{S}_+}{\arg\max}\, \ell(\mathbf{\Theta}; \mathbf{X}) - \lambda \|\mathbf{\Theta}\|_1$$

Variants and recent improvements

'13 NIPS submissions

- Use square-root Lasso in place of Lasso for tuning insensitive property package
- Solve CLIME for $p = 10^6$ (on 400 cores).

See `R` package `huge`, `fastclime`, `flare`, `QUIC`.

Practical implications of theoretical results

Selection consistency (Ravikumar, Wainwright, 2009-2012)

Denote $d = \max_{j \in \mathcal{P}}(\text{degree}_j)$. Consistency for an appropriate λ and

- $n \approx \mathcal{O}(d^2 \log(p))$ for the graphical Lasso and Clime.
- $n \approx \mathcal{O}(d \log(p))$ for neighborhood selection (sharp).

(Irrepresentability) conditions are not strictly comparable...

Ultra high-dimension phenomenon (Verzelen, 2011)

Minimax risk for sparse regression with d-sparse models: useless when

$$\frac{d \log(p/d)}{n} \geq 1/2, \qquad (\text{e.g., } n = 50, p = 200, d \geq 8).$$

Good news! when n is small, we don't need to solve huge problems because they can't but fail.

Practical implications of theoretical results

Selection consistency (Ravikumar, Wainwright, 2009-2012)

Denote $d = \max_{j \in \mathcal{P}}(\text{degree}_\text{j})$. Consistency for an appropriate λ and

- $n \approx \mathcal{O}(d^2 \log(p))$ for the graphical Lasso and Clime.
- $n \approx \mathcal{O}(d \log(p))$ for neighborhood selection (sharp).

(Irrepresentability) conditions are not strictly comparable...

Ultra high-dimension phenomenon (Verzelen, 2011)

Minimax risk for sparse regression with d-sparse models: useless when

$$\frac{d \log(p/d)}{n} \geq 1/2, \qquad (\text{e.g.}, n = 50, p = 200, d \geq 8).$$

Good news! when n is small, we don't need to solve huge problems because they can't but fail.

Model selection

Cross-validation

Optimal in terms of prediction, not in terms of selection

Information based criteria

Since, $\mathrm{df}(\hat{\beta}_\lambda^{\text{lasso}}) = \left\|\hat{\beta}_\lambda^{\text{lasso}}\right\|_0$ (Zou, Hastie, 2008)

- Straightforward application of BIC/AIC
- Adaptation for the sparse high dimensionalproblem (eBIC, AICc,. . .),
- GGMSelect (Girault *et al*, '12) selects among a family of candidates.

Stability selection (Meinshausen and Bühlman, 2010, Bach 2008)

Keep edges frequently selected on an range of λ after sub-samplings

+ Selecting "the" right λ is not a problem anymore
+ Works well for network inference (see Haury et al. 2012).

Limitations towards biological network inference

- Sparse GGM
  - \+ very solid statistical and computational framework
  - \+ extend to non strictly normal distribution (NGS)

- DREAM 5 benchmark, 2012.
  - \+ competitive to other inference methods
  - − performances remain questionable on real data, as for other methods

Idea: try to take into account biological/data features

Three tentatives follow to strengthen the inference by handling with

1. structure of the network (organization of biological mechanisms)
2. sample heterogeneity (patient heterogeneity)
3. horizontal integration (use multiple data and platforms)

$\rightsquigarrow$ Illustration on cancer data sets.

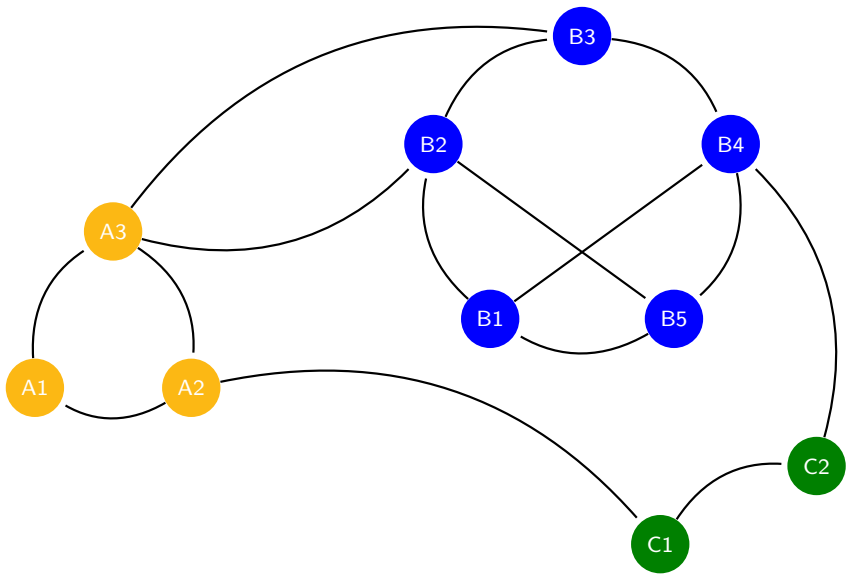
Outline

Handling with the data structure and scarcity

By introducing some prior

Priors should be biologically grounded

1. no too many genes effectively interact: sparsity,
2. networks are organized: latent clustering.



Structured regularization

SIMoNe: Statistical Inference for MOdular NEtworks

$$\arg\max_{\mathbf{\Theta},\mathbf{Z}} \ell(\mathbf{\Theta};\mathbf{X}) - \lambda \, \|\mathbf{P}_{\mathbf{Z}} \star \mathbf{\Theta}\|_{\ell_1},$$

where $\mathbf{P}_{\mathbf{Z}}$ is a matrix of weights depending on a underlying latent structure $\mathbf{Z}$ (depicted through a stochastic block model).

$\rightsquigarrow$ Cluster-driven inference via an EM-like strategy.


Ambroise, Chiquet, Matias. Inferring sparse GGM with latent structure, EJS, 2009.


Charbonnier, Chiquet, Ambroise. Weighted-Lasso for Structured Network Inference from Time Course Data, SAGMB, 2010.


Chiquet et al., SIMoNe R-package *(needs updates...)*, Note Bioinformatics, 2009.

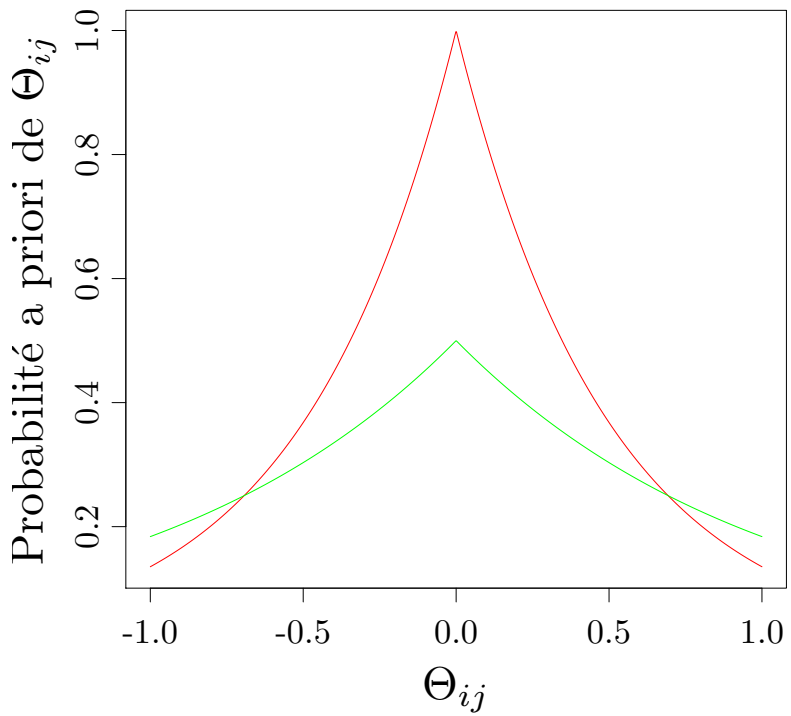
Structured regularization

"Bayesian" interpretation of ℓ_1 regularization

Laplacian prior on $\mathbf{\Theta}$ depends on the clustering $\mathbf{Z}$

$$\mathbb{P}(\mathbf{\Theta}|\mathbf{Z}) \propto \prod_{i,j} \exp\left\{-\lambda \cdot \mathbf{P}^{\mathbf{Z}}_{ij} \cdot |\mathbf{\Theta}_{ij}|\right\}.$$

$\mathbf{P_Z}$ summarizes prior information on the position of edges



How to come up with a latent clustering?

Biological expertise

- Build $\mathbf{Z}$ from prior biological information
  - transcription factors vs. regulatees,
  - number of potential binding sites,
  - KEGG pathways, ...
- Build the weight matrix from $\mathbf{Z}$.

Inference: Erdös-Rényi **Mix**ture for **Net**works (Daudin et al., 2008; Latouche et al., 2011)

- Equivalent to the Stochastic Bloc Model (SBM);
- Spread the nodes into $\mathcal{Q}$ classes;
- Connexion probabilities depend upon node classes:

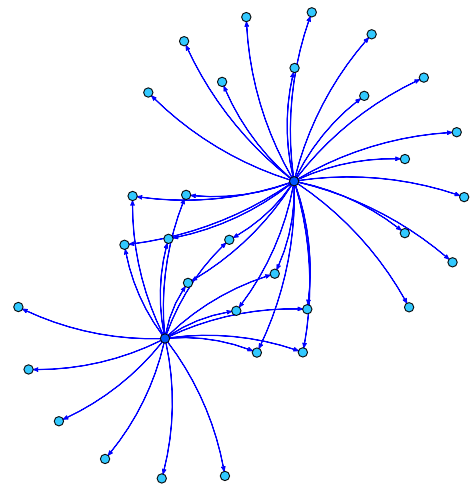
$$\mathbb{P}(i \leftrightarrow j | i \in \text{class } q, j \in \text{class } \ell) = \pi_{q\ell}.$$

- Build $P_{\mathbf{Z}} \propto 1 - \pi_{q\ell}$.

Learning scheme

Suppose you want to recover a clustered network:

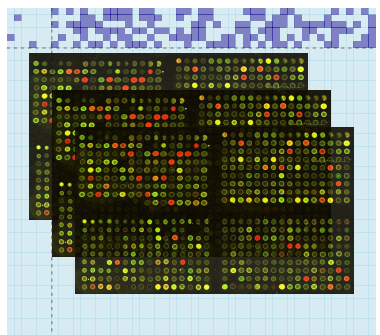
Target Adjacency Matrix



Target Network

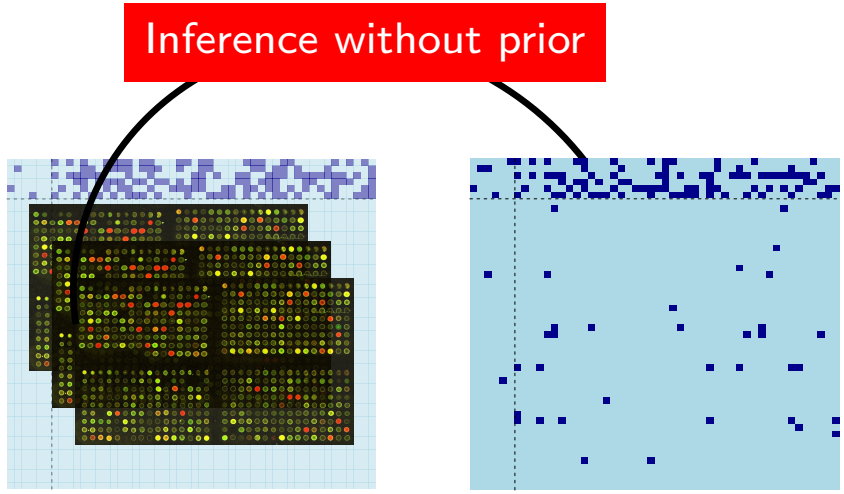
Learning scheme

Start with microarray data

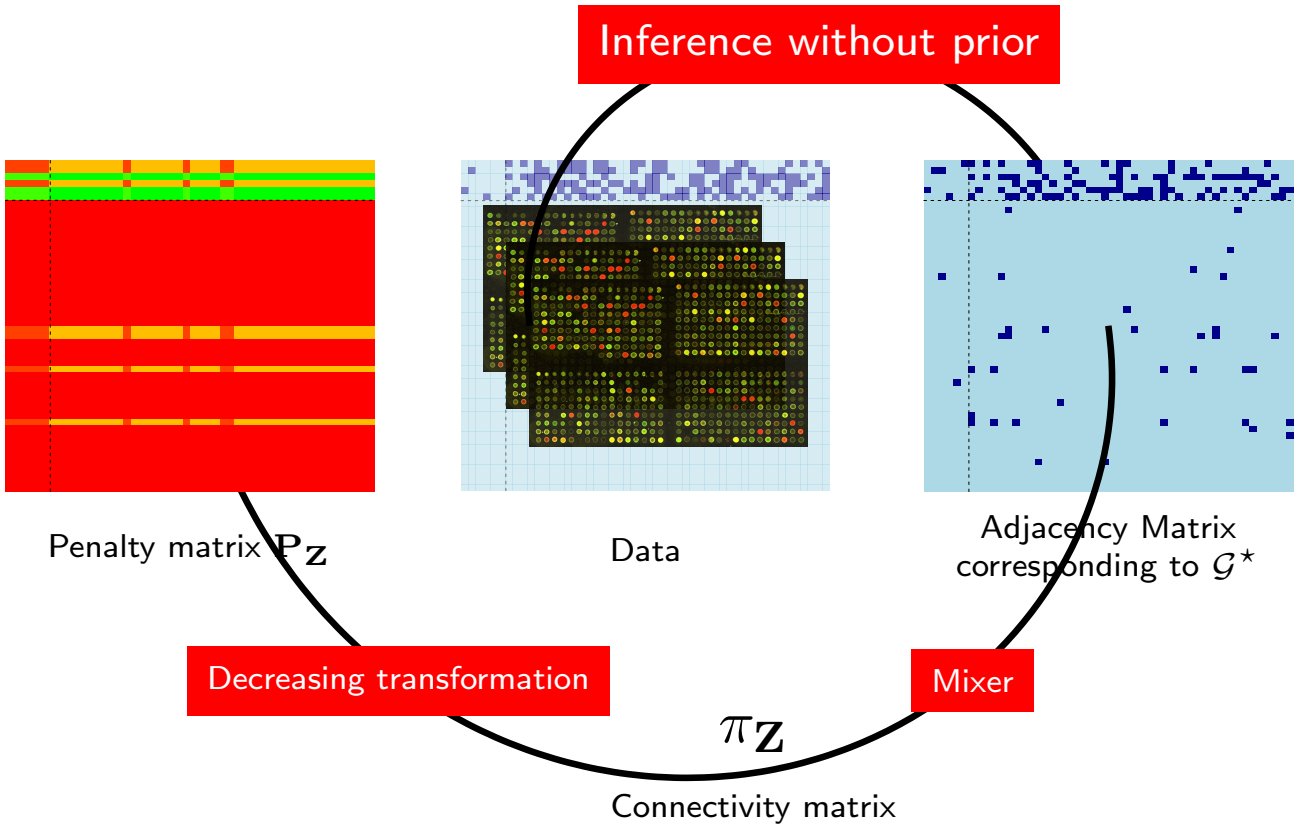


Data

Learning scheme



Learning scheme



Learning scheme

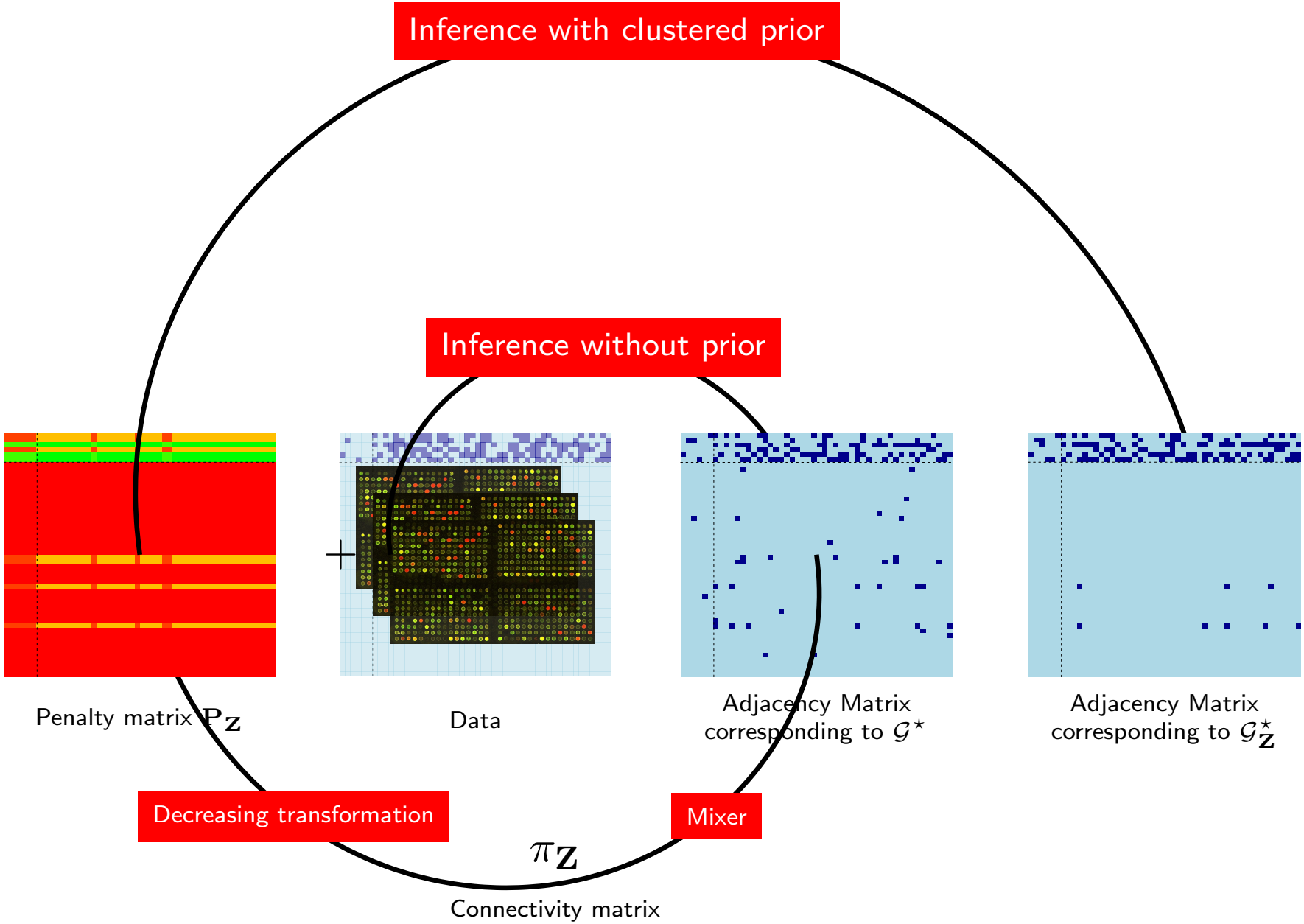


Illustration on breast Cancer

Prediction of the outcome of preoperative chemotherapy

Hess *et al.*
Journal. of Clinical Oncology, 2006.

Data set

133 patients classified as

1. pathologic complete response,
2. residual disease,

according to a signature of 26 genes (small network).

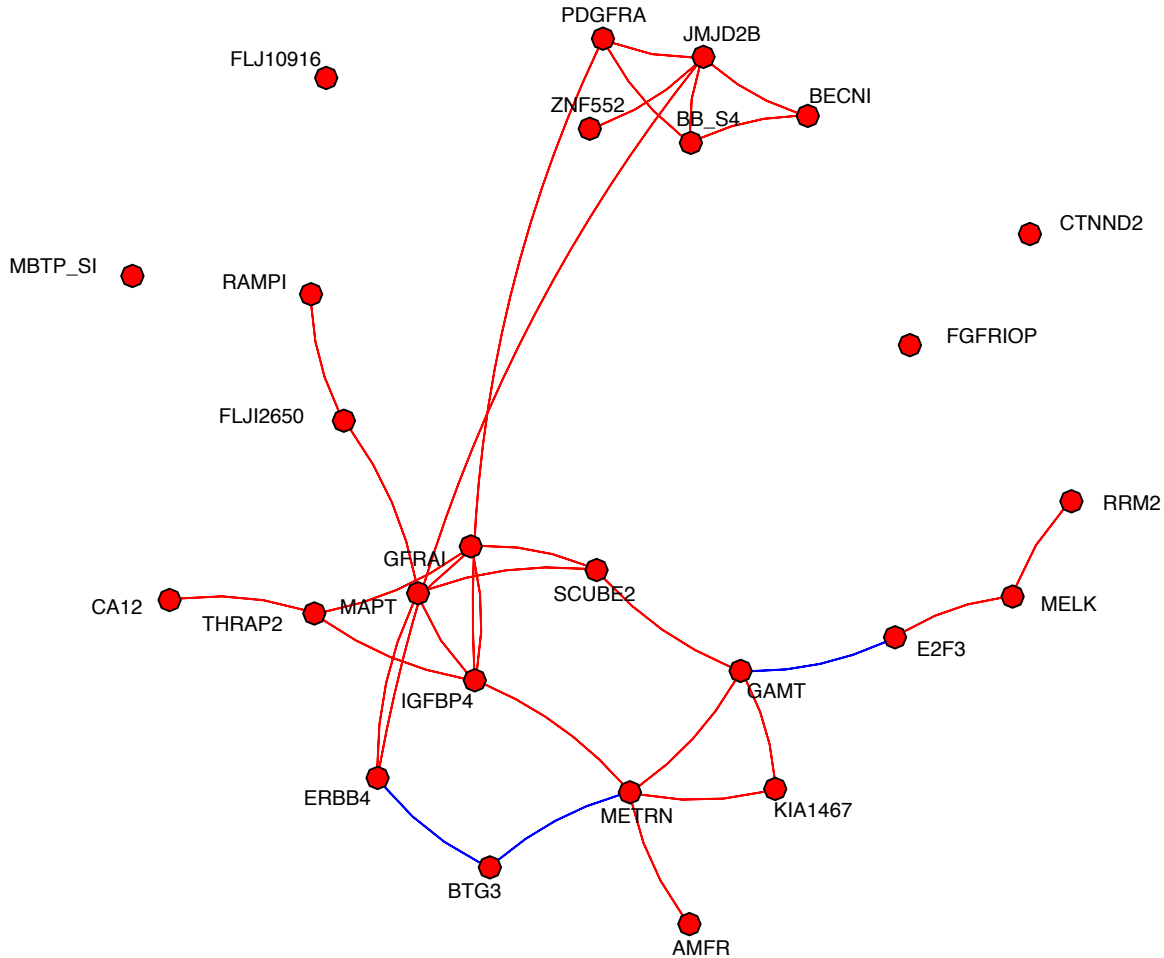


Figure: Pooling the data, Neighborhood Selection

Illustration on breast Cancer

Prediction of the outcome of preoperative chemotherapy

Hess *et al.*
Journal. of Clinical Oncology, 2006.

Data set

133 patients classified as

1. pathologic complete response,
2. residual disease,

according to a signature of 26 genes (small network).

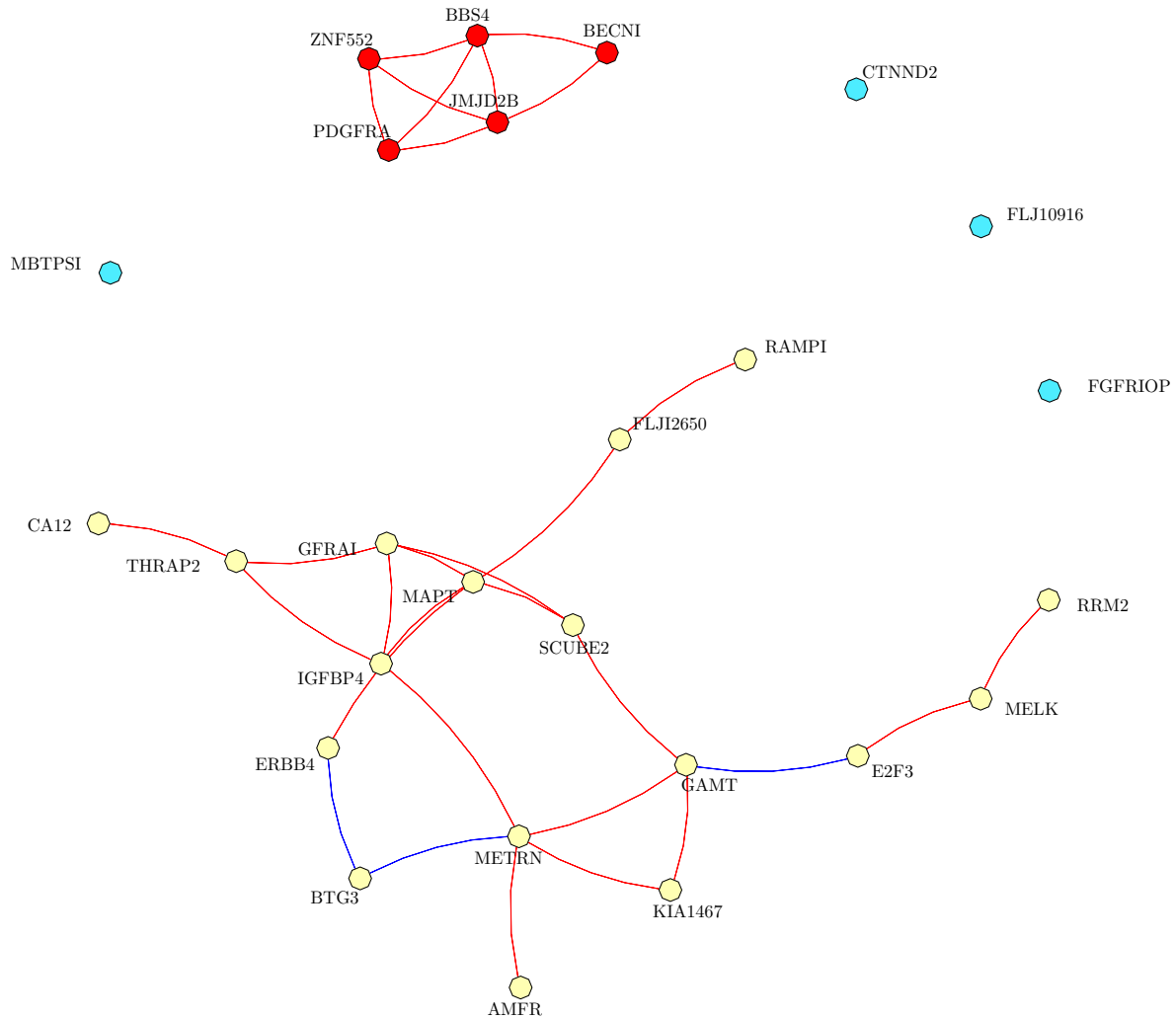


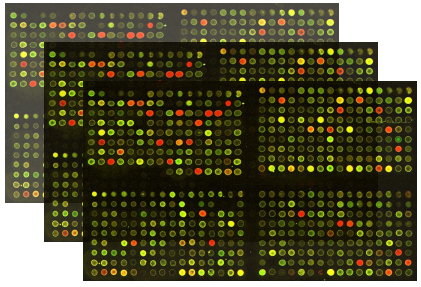
Figure: Pooling the data, SIMoNE with clustering

Outline

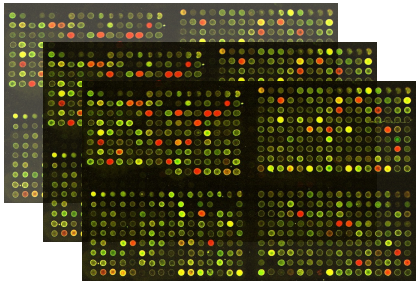
Handling the scarcity of data

Merge several experimental conditions

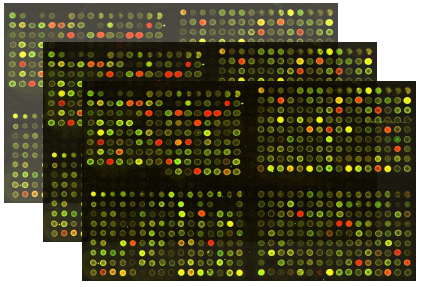
condition 1



condition 2



condition 3



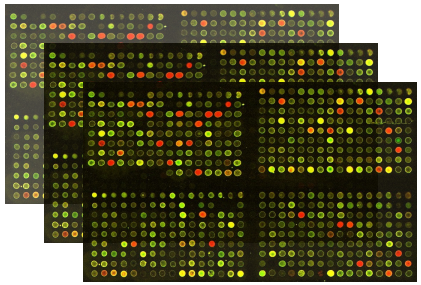
Multiple inference of GGM

$$\arg\max_{\mathbf{\Theta}^{(c)},c=1\ldots,C} \sum_{c=1}^{C} \ell(\mathbf{\Theta}^{(c)};\mathbf{S}^{(c)}) - \lambda\, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

Handling the scarcity of data

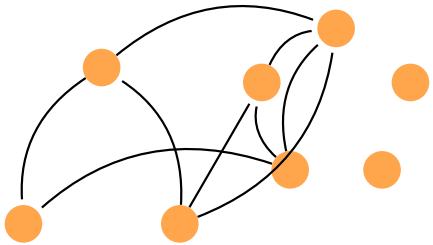
Inferring each graph independently does not help

condition 1

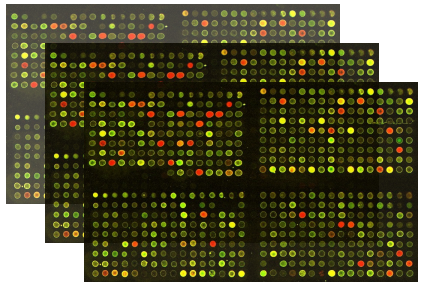


$(X_1^{(1)}, \ldots, X_{n_1}^{(1)})$

inference

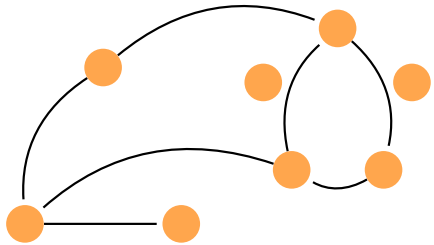


condition 2

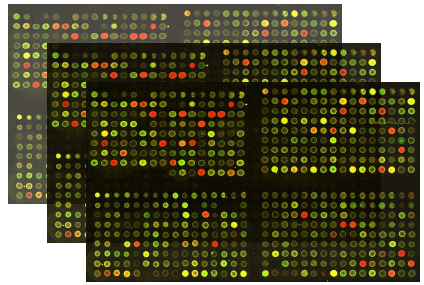


$(X_1^{(2)}, \ldots, X_{n_2}^{(2)})$

inference

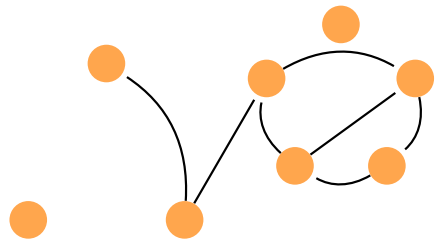


condition 3



$(X_1^{(3)}, \ldots, X_{n_3}^{(3)})$

inference

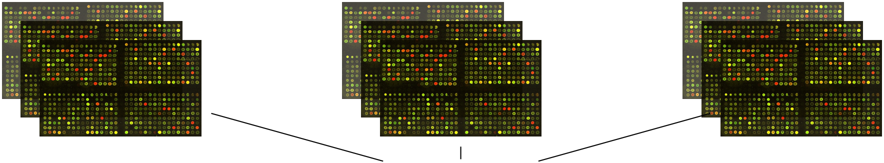


Multiple inference of GGM

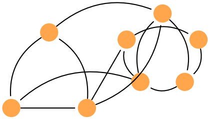
Handling the scarcity of data

By pooling all the available data (like we just have with Hess' data set)

condition 1 condition 2 condition 3



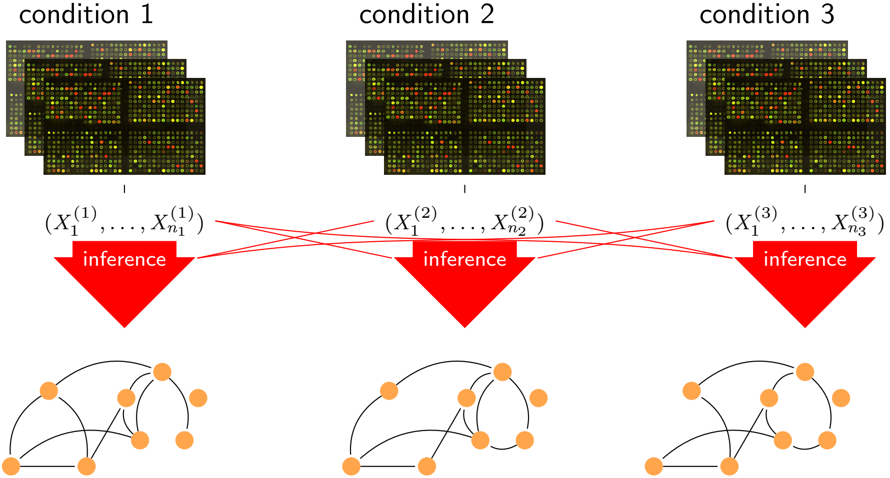
$$(X_1, \ldots, X_n),\ n = n_1 + n_2 + n_3.$$



Multiple inference of GGM

Handling the scarcity of data

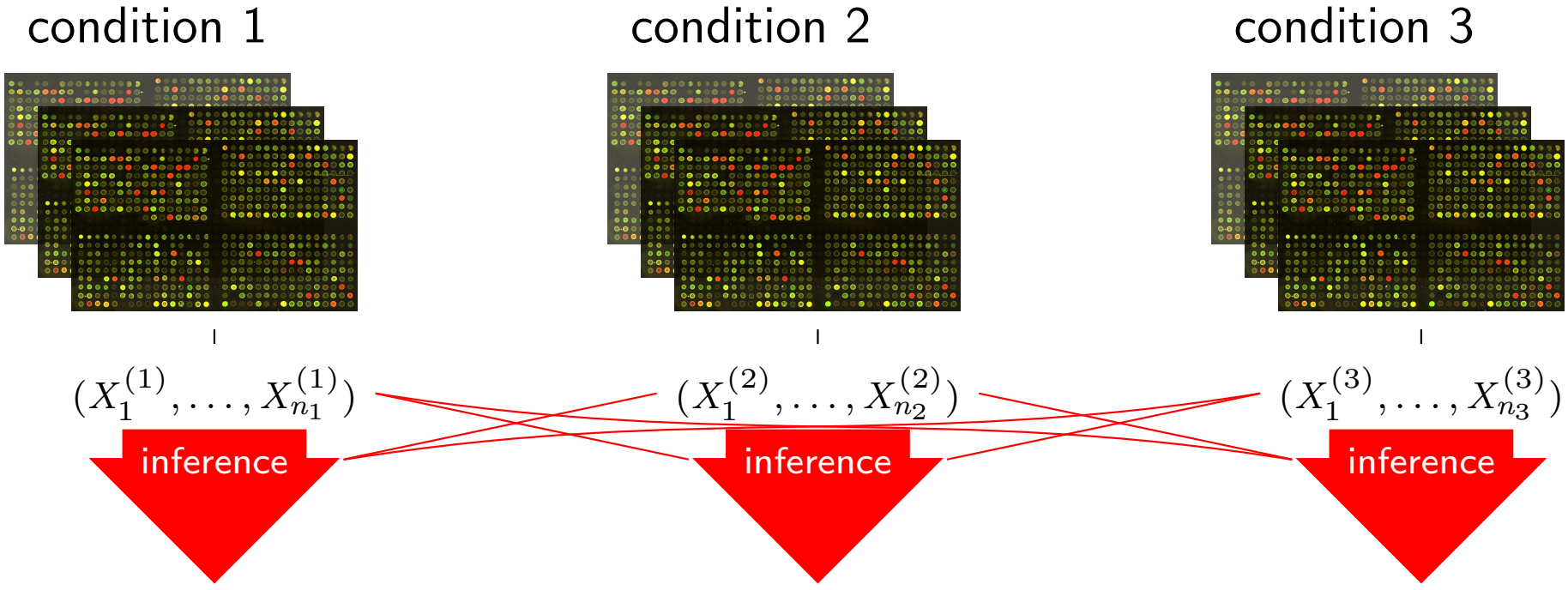
By breaking the separability



Multiple inference of GGM

Handling the scarcity of data

By breaking the separability



Multiple inference of GGM

$$\underset{\mathbf{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \ell(\mathbf{\Theta}^{(c)}; \mathbf{S}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

A multitask approach

Chiquet, Grandvalet, Ambroise, Statistics and Computing 2010/11

Break the separability

Joint the optimization problem by either modifying

$$\underset{\mathbf{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \tilde{\ell}(\mathbf{\Theta}^{(c)}; \tilde{\mathbf{S}}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

1. the fitting term
2. the regularization term

A multitask approach

Chiquet, Grandvalet, Ambroise, Statistics and Computing 2010/11

Break the separability

Joint the optimization problem by either modifying

$$\underset{\mathbf{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \tilde{\ell}(\mathbf{\Theta}^{(c)}; \tilde{\mathbf{S}}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

1. the fitting term
2. the regularization term

Intertwined-Lasso

- $\overline{\mathbf{S}} = \frac{1}{n} \sum_{t=1}^{T} n_t \mathbf{S}^{(t)}$ is the "pooled-tasks" covariance matrix.
- $\widetilde{\mathbf{S}}^{(t)} = \alpha \mathbf{S}^{(t)} + (1-\alpha)\overline{\mathbf{S}}$ is a mixture between specific and pooled covariance matrices.

A multitask approach

Chiquet, Grandvalet, Ambroise, Statistics and Computing 2010/11

Break the separability

Joint the optimization problem by either modifying

$$\underset{\mathbf{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \tilde{\ell}(\mathbf{\Theta}^{(c)}; \tilde{\mathbf{S}}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

1. the fitting term
2. the regularization term

Sparsity with grouping effect

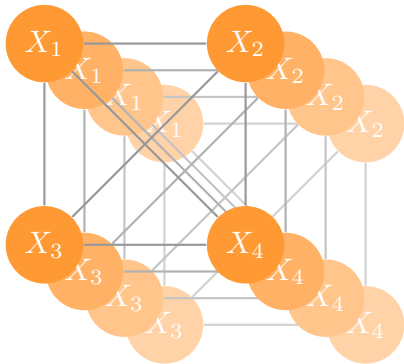
- Group-Lasso (Yuan and Lin 2006, Grandvalet and Canu, 1998),
- Cooperative-Lasso (Chiquet et al, AoAS, 2012),

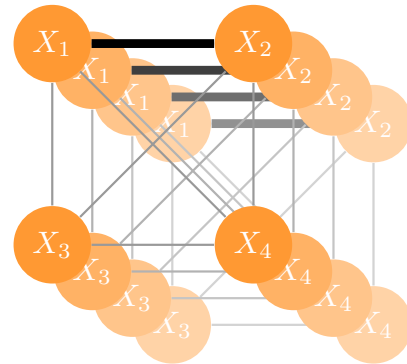
Grouping effects induced

Potential groups

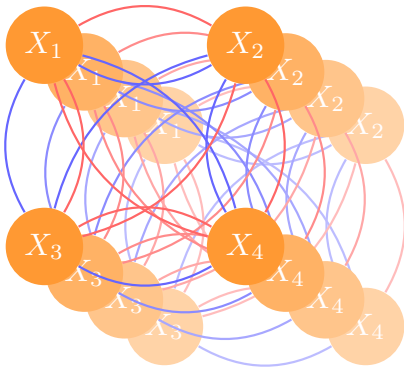
Group(s) induced by edges $(1, 2)$

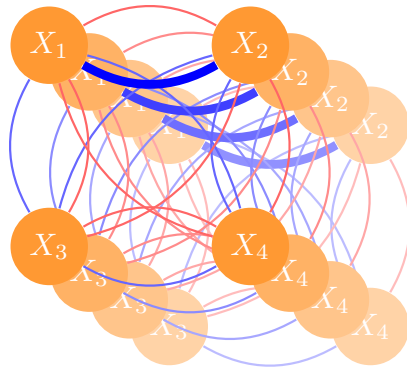
Group-Lasso

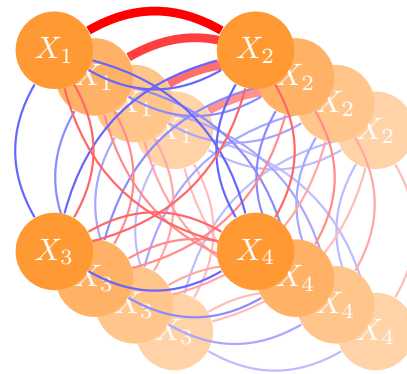




Cooperative-Lasso







Grouping effects induced

Recent works

- Use Fused-Lasso, sparse group-Lasso
- Adapted several time to the Graphical Lasso framework
  - See, e.g. D. Witten's team works.
  - The multitask/neighborhood selection's approach remains competitive.
- Promising manuscript (Mohan et al. arXiv, 2013)
  - Networks differences are only due to perturbations at the node level.
  - For instance, a hub is encouraged to be shared across tasks.

Revisiting the Hess *et al.* data set

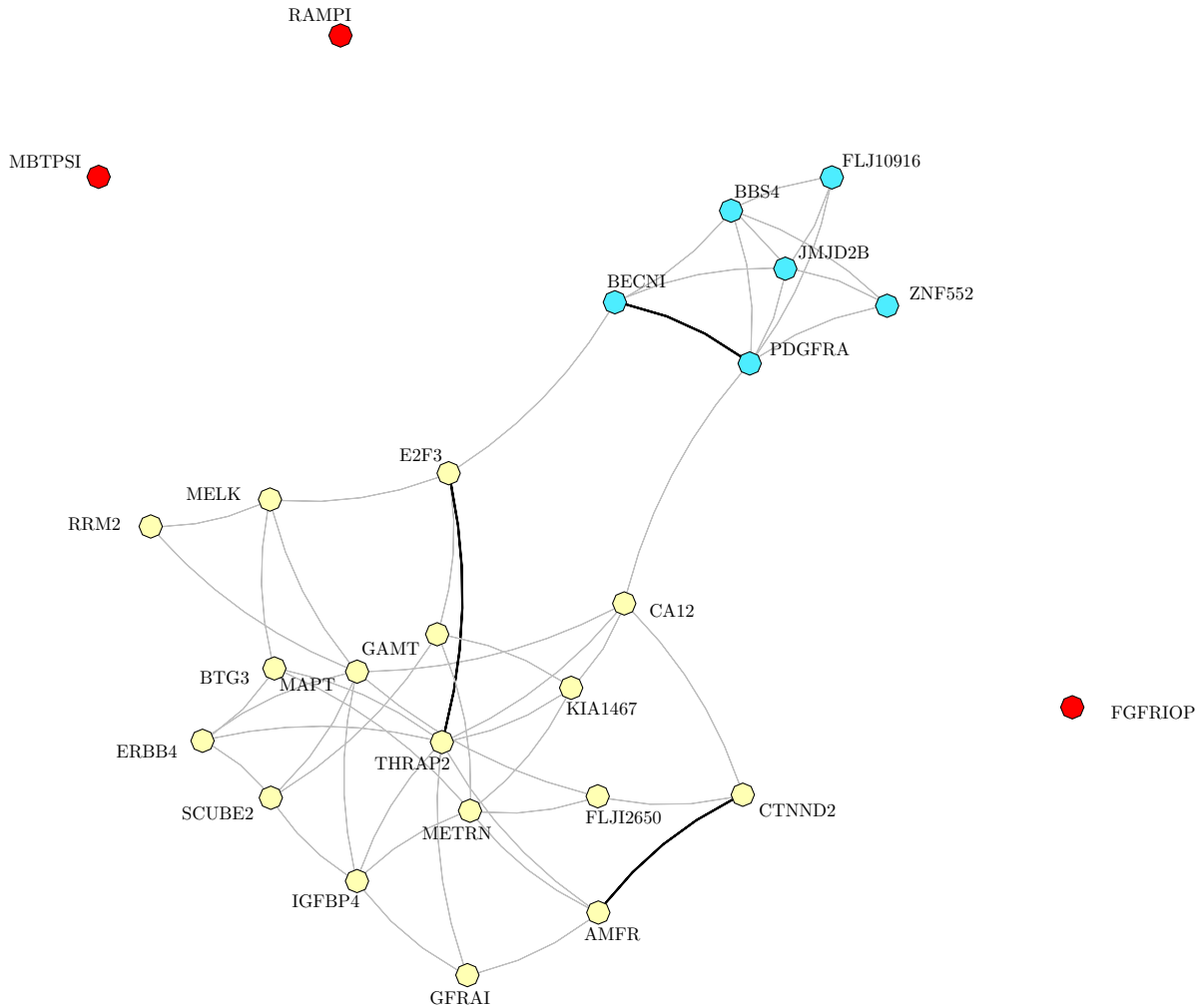


Figure: Cooperative-Lasso applied on the two sets of patients (PCR/noPCR). Bold edges are different in the finally selection graph.

Application: ER status in Breast cancer

Dataset: 466 patients with breast cancer

provided by Guedj *et al.*,
A refined molecular taxonomy of breast cancer, *Oncogene*, 2011.

Objective: identify changes in regulatory mechanisms

- ER^+/ER^-: breast cancer growth stimulated by estrogen hormones,
- ER^+ tackled with anti-hormonal therapies,
- ER^- found clinically more aggressive.

Jeanmougin, Charbonnier, Guedj and Chiquet, Network inference in breast cancer with Gaussian graphical models and extensions.

Probabilistic graphical models for genetics, Oxford University Press, to appear.

Application: ER status in Breast cancer

Network inference with cooperative-Lasso on 200 candidate genes (partial view)

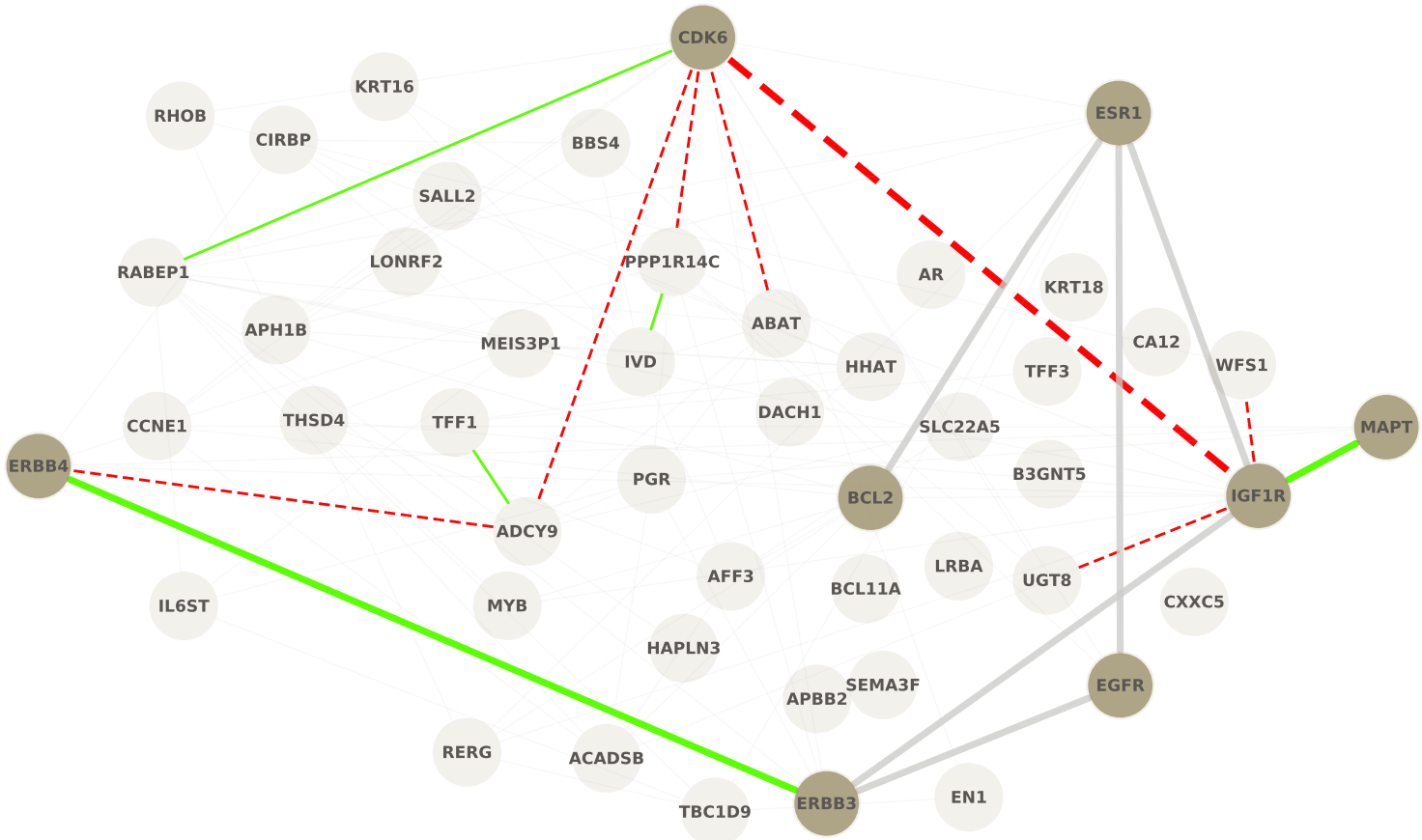


Figure: The dashed black edges are inferred only under the ER- condition and the solid black edges are only predicted under the ER+ condition. Gray are common to both conditions

Application: ER status in Breast cancer

Network inference with the cooperative-Lasso fits known anti-apoptotic mechanisms

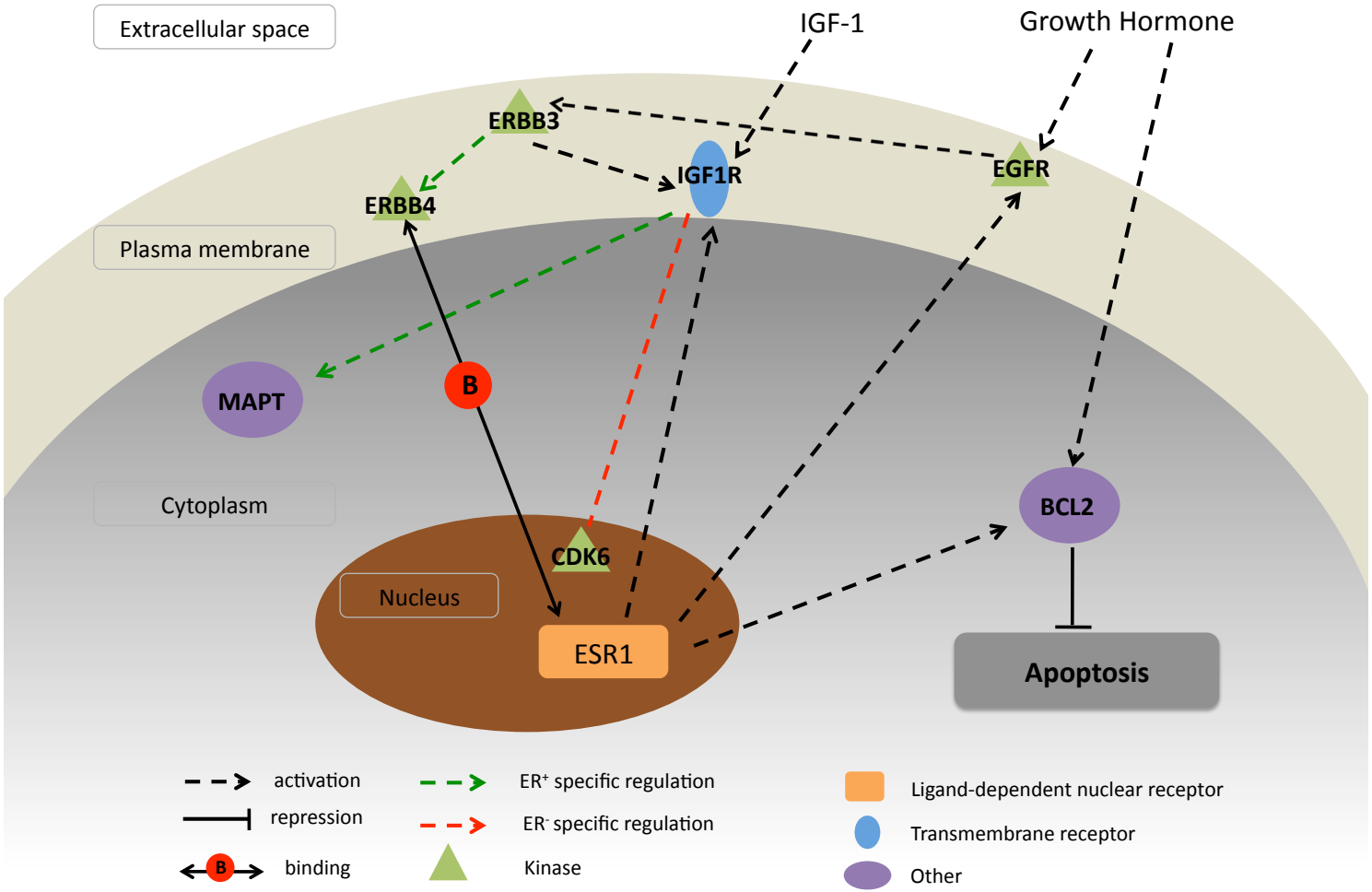
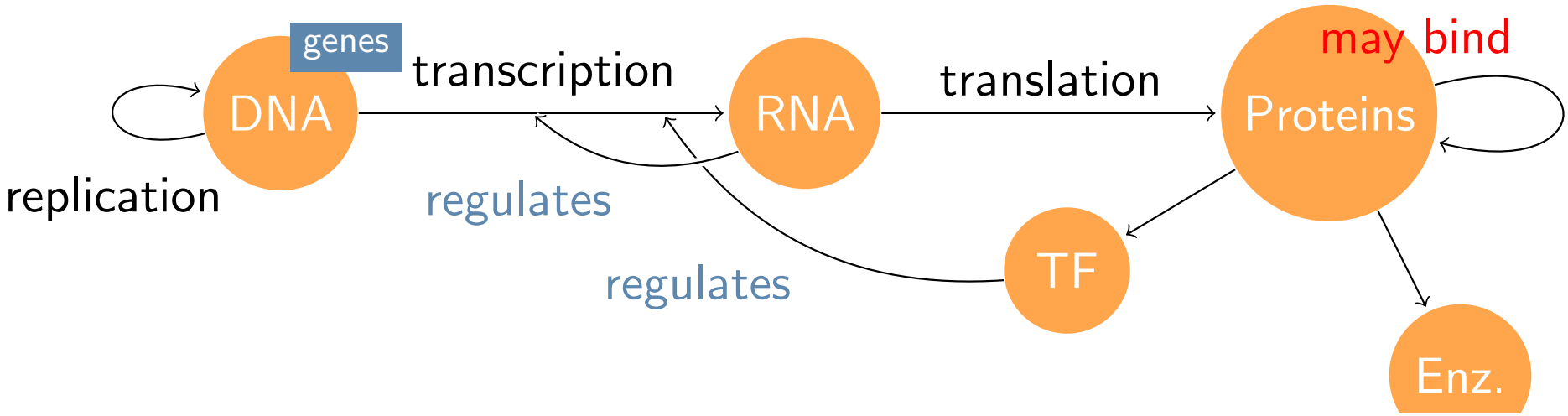


Figure: Most edges are supported by the literature (except two)

Outline

Why Multi-attribute Networks?

Joint work with E. Kolaczyk (Boston) and C. Ambroise (Évry)



Data integration

- Omic technologies can profile cells at different levels: DNA, RNA, protein, chromosomal, and functional.
- multiple molecular profiles combined on the same set of biological samples can be *synergistic*.

Remark: *a close independent work of Kolar and Xing appeared late 2012...*

Multiattribute GGM

Consider e.g. some p genes of interest and the $K = 2$ omic experiments

1. X_{i1} is the expression profile of gene i (transcriptomic data),
2. X_{i2} is the corresponding protein concentration (proteomic data).

Define a block-wise precision matrix

- $X = (X_1, \ldots, X_p)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ in $\mathbb{R}^{pK}$,
- $X_i = (X_{i1}, \ldots, X_{iK})^\intercal \in \mathbb{R}^K$.

$$\mathbf{\Theta} = \mathbf{\Sigma}^{-1} = \begin{bmatrix} \mathbf{\Theta}_{11} & & \mathbf{\Theta}_{1p} \\ & \ddots & \\ \mathbf{\Theta}_{p1} & & \mathbf{\Theta}_{pp} \end{bmatrix}, \qquad \mathbf{\Theta}_{ij} \in \mathcal{M}_{K,K}, \ \forall (i,j) \in \mathcal{P}^2.$$

Graphical Interpretation

Define $\mathcal{G} = (\mathcal{P}, \mathcal{E})$ as the multivariate analogue of the *conditional graph*:

$$(i,j) \in \mathcal{E} \Leftrightarrow \mathbf{\Theta}_{ij} \neq \mathbf{0}_{KK}.$$

Multiattribute GGM as Multivariate regression

Multivariate analysis view point

Straightforward algebra and we have

$$X_i \,|\, X_{\setminus i} = x \sim \mathcal{N}(-\mathbf{\Theta}_{ii}^{-1}\mathbf{\Theta}_{i\setminus i}x, \mathbf{\Theta}_{ii}^{-1}) \ .$$

or equivalently, letting $\mathbf{B}_i^T = -\mathbf{\Theta}_{ii}^{-1}\mathbf{\Theta}_{i\setminus i}$,

$$X_i \,|\, X_{\setminus i} = \mathbf{B}_i^T X_{\setminus i} + \boldsymbol{\varepsilon}_i \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Theta}_{ii}^{-1}), \quad \boldsymbol{\varepsilon}_i \perp X.$$

Remembering the univariate case?

$$X_i | X_{\setminus i} = -\sum_{j \in \text{neighbors}(i)} \frac{\Theta_{ij}}{\Theta_{ii}} X_j + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \Sigma_{ii}), \quad \varepsilon_i \perp X.$$

So once the data set as been carefully reshaped...

The Data

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^1 \\ \hline \vdots \\ \hline \mathbf{x}^N \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \dots & \mathbf{X}_p \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^1 & \dots & \mathbf{x}_p^1 \\ \hline \vdots & & \\ \hline \mathbf{x}_1^N & \dots & \mathbf{x}_p^N \end{bmatrix}$$

$$= \left[\begin{array}{cc|c|ccc} x_{11}^1 & x_{1K}^1 & \dots & x_{p1}^1 & \dots & x_{pK}^1 \\ \hline \vdots & \vdots & \dots & & & \\ \hline x_{11}^N & x_{1K}^N & \dots & x_{1K}^N & \dots & x_{pK}^N \end{array}\right],$$

- $\mathbf{x}^n$, is a pK-size row vector containing the data related to the nth individual.
- $\mathbf{X}_i \in \mathcal{M}_{N,K}$ is $N \times K$ bloc matrix containing the data related to the ith gene.

Multivariate Neighborhood selection

The penalized multivariate regression approach

For each node /gene, recover its neighborhood by solving

$$\arg \min_{\mathbf{B}_i \in \mathcal{M}_{(p-1)K,K}} \frac{1}{2N} \left\| \mathbf{X}_i - \mathbf{X}_{\setminus i} \mathbf{B}_i \right\|_F^2 + \lambda \Omega(\mathbf{B}_i),$$

Choice of Penalty

Group-based penalty to activate the set of attributes simultaneously on a given link:

$$\Omega(\mathbf{B}_i) = \sum_{j \in \mathcal{P} \setminus i} \|\mathbf{B}_{ij}\| \quad , \quad \mathbf{B}_{ij} \in \mathcal{M}_{KK}$$

- $\|M\| = \|M\|_F = \left(\sum_{i,j} M_{ij}^2 \right)^{1/2}$, the Frobenius norm,
- $\|M\| = \|M\|_\infty = \max_{i,j} |M_{ij}|$, the sup norm (shared magnitude),
- $\|M\| = \|M\|_\star = \sum \text{eig}(M)$, the nuclear norm (rank penalty).

Simulation Study Design

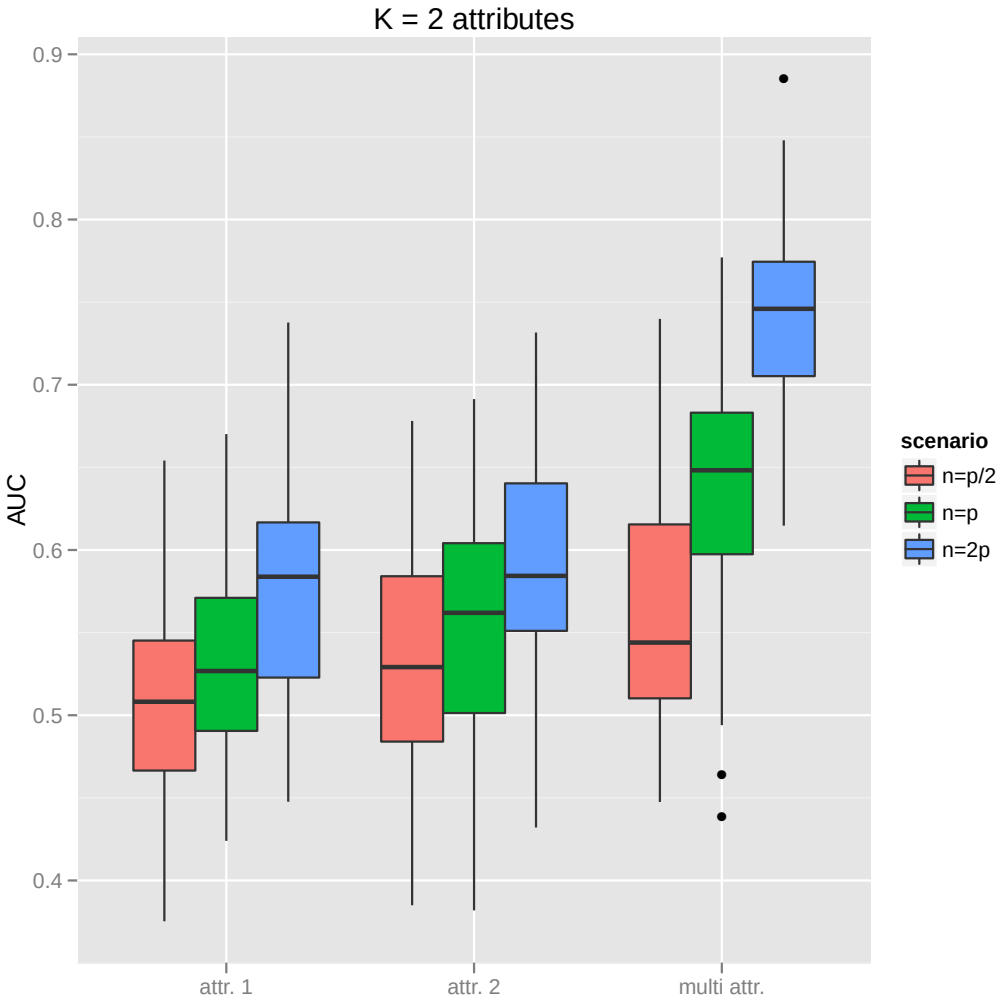
Small study, set up as follows.

1. Simulation of an Erdös-Renyi graph;
2. Expand the adjacency matrix to multivariate space

$$\mathbf{A} = (\mathbf{A} + I) \otimes \mathbb{I}_{K \times K};$$

3. Compute $\mathbf{\Theta}$ a positive definite approximation of $\mathbf{A}$ by replacing null and negative eigenvalues by a small constant
4. $\mathbf{\Theta} = \mathbf{\Theta} + \gamma I$ with γ a parameter controlling the difficulty of the problem;
5. Draw an i.i.d. sample $\mathbf{X}$ of $X \sim \mathcal{N}(0, \mathbf{\Sigma})$.

Simulation Results

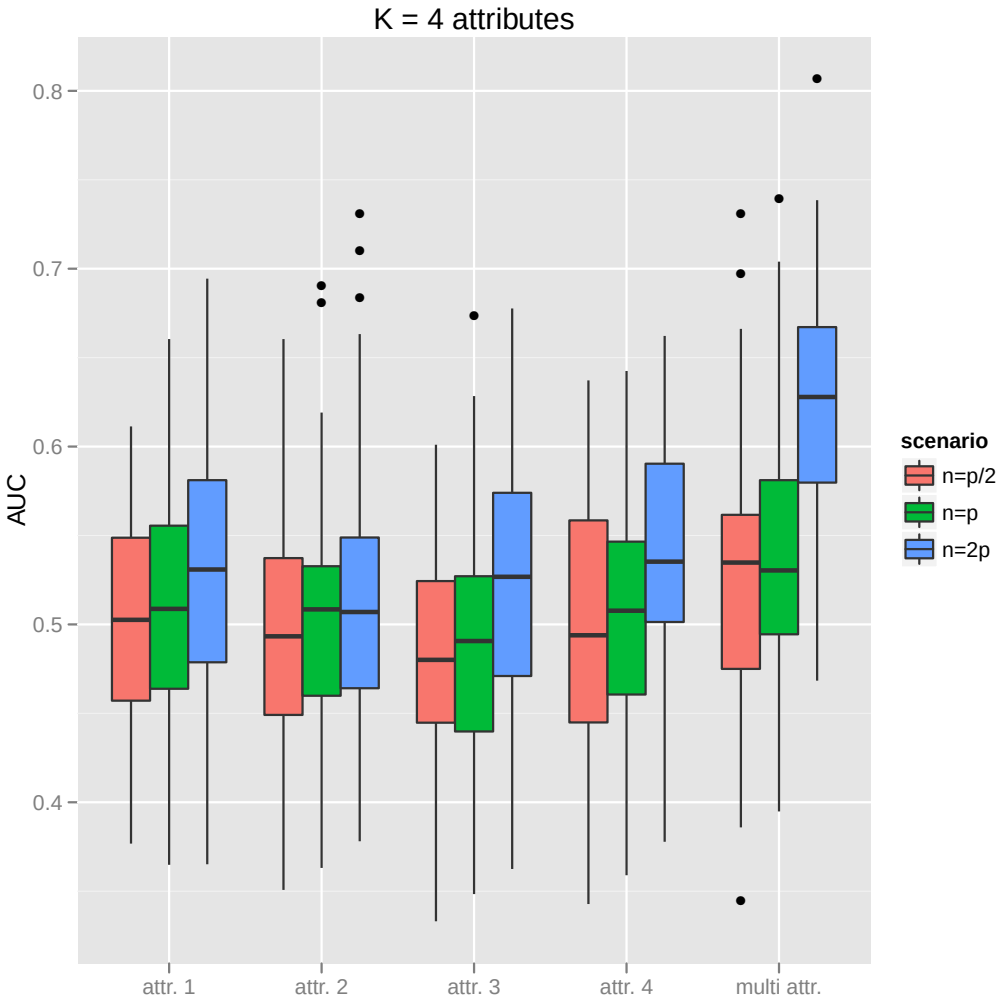


Settings

- K=2 attributes
- $p = 20$ (small networks),
- 20 edges on average,
- vary n from $p/2$ to $2p$,
- AUC averaged over 50 runs.

Aggregation improves upon single-attribute methods for learning networks

Simulation Results



Settings

- K=4 attributes
- $p = 20$ (small networks),
- 20 edges on average,
- vary n from $p/2$ to $2p$,
- AUC averaged over 50 runs.

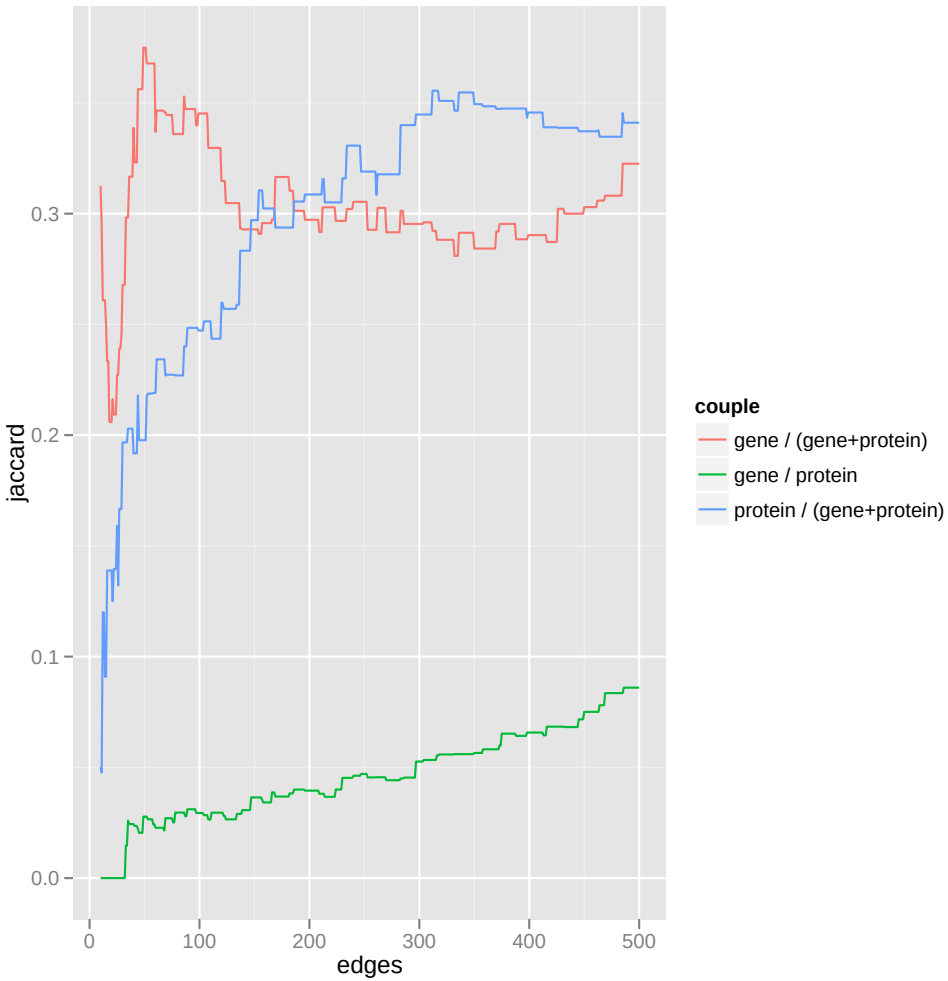
Aggregation improves upon single-attribute methods for learning networks

Illustration on the NCI-60 data set

Molecular profile data on a panel of 60 diverse human cancer cell lines

1. Protein: reverse-phase lysate arrays (RPLA) for 92 antibodies;
2. Gene : Human Genome U95 affymetrix ($\sim$ 9,000 genes).

$\rightsquigarrow$ consensus set with 91 protein and corresponding gene profiles.

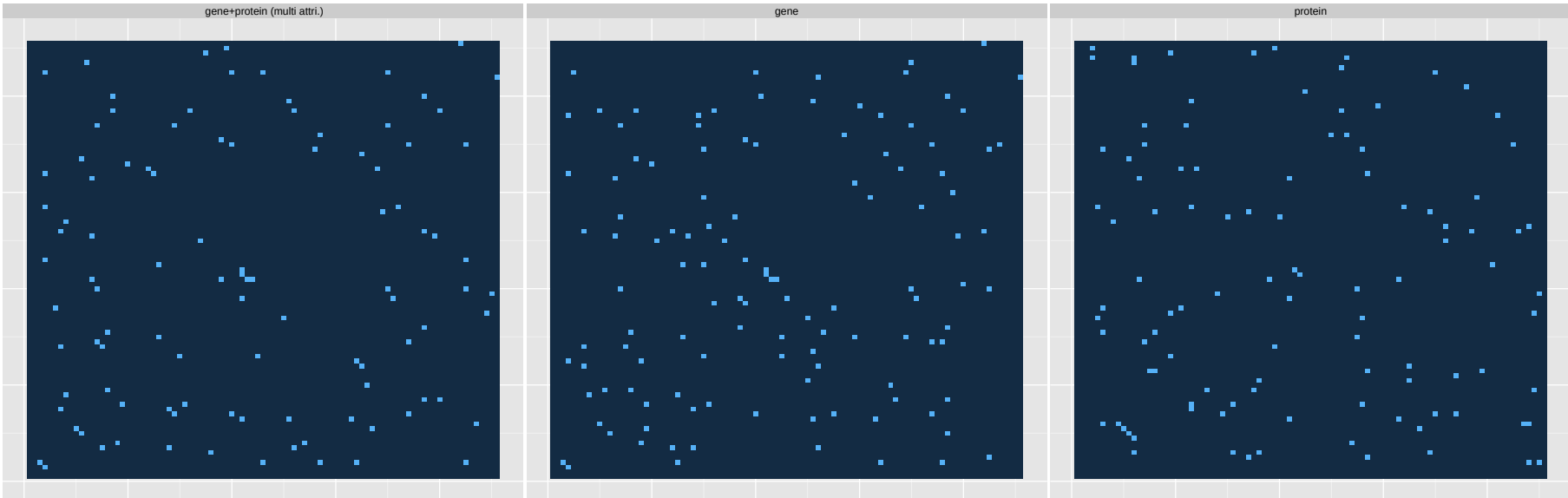


Jaccard's similarity index

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$\rightsquigarrow$ multiattribute network shares a high Jaccard index with both uni attribute networks.

Illustration: Three Types of Regulatory Networks

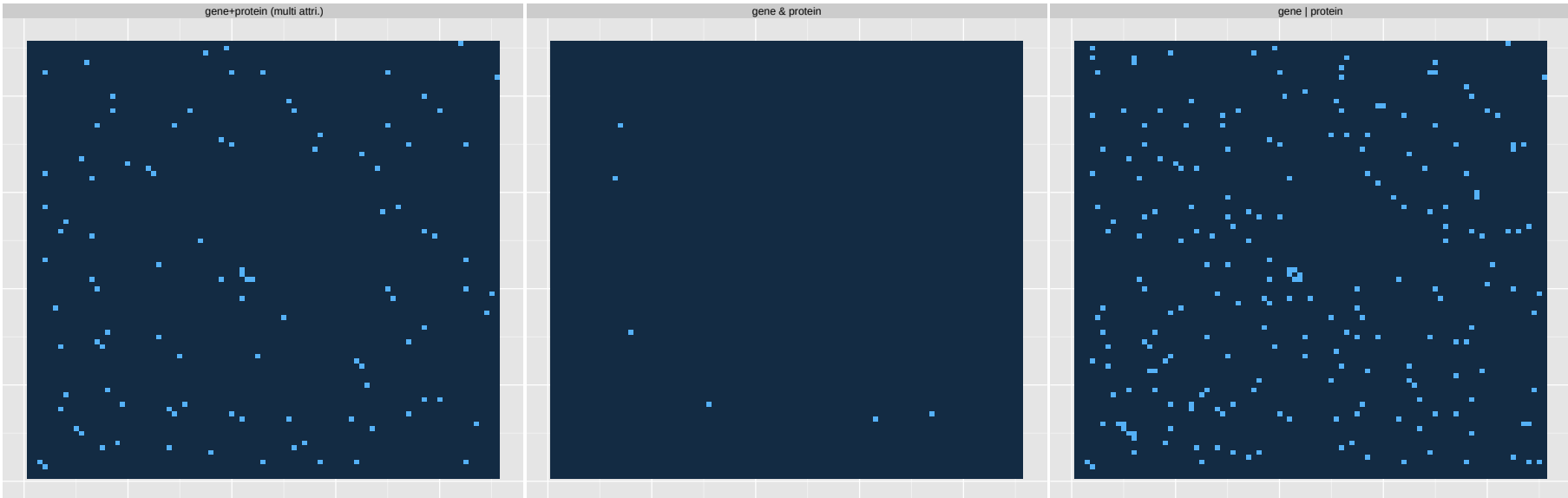


multi attribute
gene + protein

attribute 1
gene

attribute 2
protein

Illustration: Three Types of Regulatory Networks



multi attribute
gene + protein

attribute 1 & 2
gene AND protein

attribute 1 | 2
gene OR protein

Conclusion

Sparse Gaussian Graphical Model

Well established framework with a vast, growing literature

1. Nice modeling tool (conditional dependencies),
2. Good theoretical framework (which I have not much talked about),
3. Powerful algorithms (screening, first-order, homotopy)
   - that scale the dimension (large p large n)
   - that allow resampling/parallelization (for robustness)

$\rightsquigarrow$ Great tool for covariance estimation/selection in a reasonably high dimensional settings.

Still...

- phenomena are quite complex: a biological interaction is not even well defined
- more data is coming...

$\rightsquigarrow$ Need for methods with data integration and to solve couple problems

Perspectives/Ongoing work

Joint network inference to the estimation of a related biological feature

Enhance network reconstruction by simultaneously identifying TF

Knowledge of TF is crucial to achieve good network reconstruction

 Haury *et al.*, BMC Bioinformatics, 2012.

$\rightsquigarrow$ With S. Robin, we are working on TF elucidation at large scale from transcriptomic data through penalized multivariate regression.

Couple differential analysis (DA) to network inference

Introducing network knowledge is of great benefit for DA

 F. Rapaport *et al.*, BMC Bioinfo, 2007 / L. Jacob *et al.*,Ann. Appl. Stat., 2012.

$\rightsquigarrow$ With P. Gutierrez and G. Rigaill we proposed fused-ANOVA, a penalized model for differential analysis

$\rightsquigarrow$ A unifying convex method is planed to be part of Trung Ha's PhD Thesis (with Guillem and M-L Martin-Magnette).

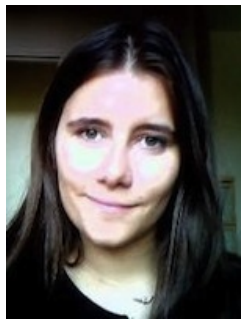
Thanks

To you for your patience and for listening...

Co-authors

C. Ambroise
PU, Évry

C. Charbonnier
PhD, Inserm



M. Jeanmougin
PhD, Curie

Y. Grandvalet
DR, Compiègne

C. Matias
DR, Évry

Co-Workers

S. Robin
DR, AgroParis

E. Kolaczyk
PU, Boston

P. Gutierrez
M2, PhD?

G. Rigaill
MCF, Évry