Biological prior for network inference with Gaussian graphical models.

Application to Estrogen Receptor Status in Breast Cancer.

Marine Jeanmougin
Problem

Microarray data

\( n \approx 10s/100s \) of microarrays

\( p \approx 1000s \) of genes

\( O(g^2) \) parameters (edges)!

Gene regulatory network

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
Problem

High dimensional setting

- “large p, small n”
  - 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 ⇔ absence of edge)
- Sparse estimation via L1-regularization

A challenging issue

A vast space of possible network structures

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

1 Method
   a) Biological prior definition: differential and pathway analysis
   b) Network inference: regularized GGM, multitask strategy

2 Application
   a) Context: ER status in Breast Cancer
   b) Results and interpretation

3 Conclusion
Method

Biological prior definition
Differential analysis

\( X_{ig}^{(c)} \): expression level of the \( i \)th 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:

\[
\begin{align*}
H_0 & : \mu_g^{(1)} = \mu_g^{(2)}, \\
H_1 & : \mu_g^{(1)} \neq \mu_g^{(2)}.
\end{align*}
\]

Limma t-statistic (Smyth 2004)

\[
t_{\text{limma}}^{ig} = \frac{\bar{x}^{(1)} - \bar{x}^{(2)}}{S_{\text{limma}}^{ig} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},
\]

- \( S_{\text{limma}}^{ig} \): Bayesian estimator of the variance
- Stabilize the estimation of gene variances
Method - Summary

Differential analysis

Signature

Microarray data
How to interpret gene signatures in biologically meaningful terms?

⇝ by determining whether the signature is enriched in pathway* key actors.

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

Figure: Group testing for pathway analysis
Under the null hypothesis of no over-representation

\[ P(Y \geq y) = 1 - P(Y \leq y) \]

\[ = 1 - \sum_{i=0}^{y} \frac{(s)_i (p-s)_{t-i}}{(p)_t}. \]

\( P(Y \geq y) \) probability of observing at least \( y \) genes of a pathway of size \( t \) in the signature
**Method - Biological prior definition**

In practice...

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Genes in pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>HER-2 Signaling in Breast Cancer</td>
<td>CCNE1, CDK6, PARD6B, ERBB3, EGFR</td>
</tr>
<tr>
<td>Glioblastoma Multiforme Signaling</td>
<td>CCNE1, RHOB, IGF1R, CDK6, EGFR</td>
</tr>
<tr>
<td>Estrogen-Dependent Breast Cancer Signaling</td>
<td>IGF1R, ESR1, EGFR</td>
</tr>
<tr>
<td>Small Cell Lung Cancer Signaling</td>
<td>CCNE1, CDK6, BCL2</td>
</tr>
<tr>
<td>Aryl Hydrocarbon Receptor Signaling</td>
<td>CCNE1, TFF1, CDK6, ESR1</td>
</tr>
</tbody>
</table>

**Table:** Results of pathway analysis

- Pathways do not clearly represent distinct entities!
- We need to summarize the set of pathways found significant!
Method - Biological prior definition

<table>
<thead>
<tr>
<th>Pathway Name</th>
<th>Genes in pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>HER-2 Signaling in Breast Cancer</td>
<td>CCNE1, CDK6, PARD6B, ERBB3, EGFR</td>
</tr>
<tr>
<td>Glioblastoma Multiforme Signaling</td>
<td>CCNE1, RHOB, IGFR1, CDK6, EGFR</td>
</tr>
<tr>
<td>Estrogen-Dependent Breast Cancer Signaling</td>
<td>IGF1R, ER, ER, EGFR</td>
</tr>
<tr>
<td>Small Cell Lung Cancer Signaling</td>
<td>CCNE1, CDK6, BCL2</td>
</tr>
<tr>
<td>Aryl Hydrocarbon Receptor Signaling</td>
<td>CCNE1, TFF1, CDK6, ER, ER</td>
</tr>
</tbody>
</table>

- Binary matrix
- Jaccard distance
- Distance matrix
- Ward's criterion
- Core pathways
Method - Summary

Differential analysis

Pathway analysis

Microarray data

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

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

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

After a simple rescaling $\Theta$ can be interpreted as the adjacency matrix conditional dependency or non null partial correlation between $i$ and $j$. 

i \quad if and only if \quad j

if and only if

conditional dependency or non null partial correlation between
Method - Network Inference

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

- $S^{-1}$ is not defined for $n < p$.
- If $n < p$, neither $\Theta$ 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 $\ell_1$-norm)
Take into account the core-pathways information as an \textit{a-priori} knowledge:

\[ \leadsto \text{Edges between two genes of the same core-pathway are less penalized} \]

**Statistical approach**

Use adaptive penalty parameters for different coefficients

- Let $Z$ be the set of indicator variable for nodes

\[ \hat{\Theta}_\lambda = \arg \max_{\Theta} \mathcal{L}(\Theta; \text{data}) - \lambda \| P_Z \ast \Theta \|_{\ell_1}, \]

where $P_Z$ is a matrix of weights depending on the core-pathway membership $Z$. 
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**

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

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**

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

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

$$\rho_{Z_i Z_j} = \begin{cases} 
\sum_{q, \ell \in \mathcal{Q}} Z_{iq} Z_{j\ell} \frac{1}{\lambda_{in}}, & \text{if } i \neq j, \text{ and } q = \ell, \\
\sum_{q, \ell \in \mathcal{Q}} Z_{iq} Z_{j\ell} \frac{1}{\lambda_{out}}, & \text{if } i \neq j, \text{ and } q \neq \ell, \\
1, & \text{otherwise.} 
\end{cases} \quad (2)$$
Method - Summary

Microarray data 

Differential analysis 

Pathway analysis 

Network inference 

Signature 

Core pathways 

Regulation network
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.

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

▶ Comparison of regulation patterns
Cellular growth & proliferation

Apoptosis

Aryl Hydrocarbon Receptor Signaling
P53 Pathway
Small Cell Lung Cancer
Molecular Mechanisms of Cancer
Sphingolipid Metabolism
Valine Leucine and Isoleucine Degradation
Progesterone Mediated Oocyte Maturation
Oocyte Meiosis
CDK5 Signaling
Endocytosis
HER-2 Signaling in Breast Cancer

Cell death

Estrogen-Dependent Breast Cancer Signaling
Glioblastoma Multiforme Signaling
Glioma
Melanoma
Pathways in Cancer
Prostate Cancer
Bad Pathway
Tel Pathway
ERBB Signaling Pathway
Calcium Signaling Pathway
HER-2 Signaling in Breast Cancer
Endocytosis

Protein trafficking

Small molecules biochemistry

Figure: Core pathways
ER status in breast cancer

**Figure:** Sub-network inferred from the ER status signature
ER status in breast cancer

Anti-apoptotic mechanisms

Common regulations

Estrogen receptor (ESR1) - BCL2 (Peterson et 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
Discussion

Summary

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

Perpectives: 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

Pharnext
Mickaël Guedj
Serguei Nabirotchkin
Ilya Chumakov

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

Jan, Caroline, Fabrice, Mickaël, Matthieu.