Transcriptomic data analysis: from variable selection to network inference.

Application to the study of treatment response in basal breast cancer.

Marine Jeanmougin
Christophe Ambroise, Mickaël Guedj

Ecole Thématique “Genomique et Modélisation”
How to translate gene lists into a better understanding of the biological phenomena?

→ Two methods hold great promises:
  - Pathway Analysis
  - Network Inference
How to translate gene lists into a better understanding of the biological phenomena?

Two methods hold great promises:

- Pathway Analysis
- Network Inference
How to translate gene lists into a better understanding of the biological phenomena?

Two methods hold great promises:

- Pathway Analysis
- Network Inference
Context

How to translate gene lists into a better understanding of the biological phenomena?

⇝ Two methods hold great promises:

▶ Pathway Analysis

▶ Network Inference
Pathways: sets of gene products interacting in order to achieve a specific cellular function.

Pathway Analysis

- **Biological purpose:** Is this pathway targeted by the set of differentially expressed genes?

- **Methodology:** Testing whether the set of differentially expressed genes is “enriched” by a given pathway or cellular function.

Cell Cycle
Context
Network Inference

Identification of biomarkers

select key genes
Differentially expressed
between 2 (or more) conditions

Various statistical frameworks
Boolean modelisation
Differential equations
Graphical Models (GGM)
Identification of biomarkers

Select key genes
Differentially expressed
between 2 (or more) conditions

Various statistical frameworks
Boolean modelisation
Differential equations
Graphical Models (GGM)
A challenging issue

A vast space of possible network structures

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

- Pathway Analysis results: an informative prior to drive Network Inference
A challenging issue

A vast space of possible network structures

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

- Pathway Analysis results: an informative prior to drive Network Inference
A challenging issue

A vast space of possible network structures

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

- Pathway Analysis results: an informative prior to drive Network Inference
Analysis process: outlines

Normalized data
Analysis process: outlines

Step 1
Identification of biomarkers

Normalized data
Analysis process: outlines

Step 1
Identification of biomarkers

Step 2
Pathway Analysis

Normalized data
Analysis process: outlines

Normalization of data

Step 1: Identification of biomarkers

Step 2: Pathway Analysis

Step 3: Pathway analysis results + Network inference
Step 1: Identification of biomarkers

1 - Differential analysis

- Identify genes associated with a phenotype of interest

→ R package **Limma**: moderate t-test approach (Smyth, G.K. 2004 - SAGMB)

Let $X_{cr}^i$ be the level of expression observed for gene $i$, replicate $r$, under condition $c$ such as:

$$\mathbb{E}(X_{cr}^i) = \mu_c^i \quad \text{and} \quad \text{Var}(X_{cr}^i) = (\sigma_c^i)^2.$$  

The limma statistic is defined as:

$$t_{\text{limma}}^i = \frac{\bar{x}^i_1 - \bar{x}^i_2}{S_{\text{limma}}^i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$  

$S_{\text{limma}}^i$: combinaison of an estimate obtained from a prior distribution and the pooled variance.

---


Jeanmougin et al. 2010, Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PLoS ONE*
Step 1: Identification of biomarkers

1 - Differential analysis

Identify genes associated with a phenotype of interest

R package **Limma**: moderate t-test approach (Smyth, G.K. 2004 - SAGMB)

Let $X_{cr}^i$ be the level of expression observed for gene $i$, replicate $r$, under condition $c$ such as:

$$E(X_{cr}^i) = \mu_c^i \quad \text{and} \quad Var(X_{cr}^i) = (\sigma_c^i)^2.$$  

The limma statistic is defined as:

$$t_{\text{limma}}^i = \frac{\bar{x}_{1}^i - \bar{x}_{2}^i}{S_{\text{limma}}^i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

$S_{\text{limma}}^i$: combinaison of an estimate obtained from a prior distribution and the pooled variance.

---


Jeanmougin et al. 2010, Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. PLoS ONE
Step 1: Identification of biomarkers

Differential analysis limitation

Signature is **not a unique set** (Ein-Dor et al. 2005):

- Lack of agreement between studies
- Possible explanations
  - outliers measures,
  - sampling variation for moderate sample size,
  - genetic heterogeneity,
  - ...

Ein-Dor et al. 2005, Outcome signature genes in breast cancer: is there a unique set? Bioinformatics
Step 1: Identification of biomarkers

2 - Cleanup of the gene signature

- Improve the overall **homogeneity** of the signature

~~ Random Forest (Breiman L. 2001 - *Machine Learning*)

1. Identify the most informative genes
2. Remove the potential outliers

![Before running Random Forest](chart1.png)

![After running Random Forest](chart2.png)
Step 1: Identification of biomarkers

2 - Cleanup of the gene signature

- Improve the overall homogeneity of the signature

⇒ Random Forest (Breiman L. 2001 - *Machine Learning*)

1. Identify the most informative genes
2. Remove the potential outliers

**Before running Random Forest**

**After running Random Forest**
Step 1: Identification of biomarkers

3- Functionnal partners identification

“Genes causing the same phenotype are likely to be functionally related” (Gandhi et al. 2006):

- They form some kind of module: for instance, a multi-protein complex.

![Diagram showing interactions between biomarkers]

⇒ Adding strongly connected first partners using PPI networks: String Software
3- Functionnal partners identification

“Genes causing the same phenotype are likely to be functionally related” (Gandhi et al. 2006):

- They form some kind of module: for instance, a multi-protein complex.

Adding strongly connected first partners using PPI networks: String Software
Analysis process: outlines

Step 1: Identification of biomarkers

Step 2: Pathway Analysis

Step 3: Pathway analysis results + Network inference

Normalized data

Signature

Differentially expressed genes
First partners
Step 2: Pathway analysis

**Enrichment analysis**
- Describe the gene signature in a **biological meaningful** way
  - Analysis for enrichment of KEGG pathway membership
    - Fisher’s exact test

**Hierarchical clustering of pathways**
- Increase the interpretability of enrichment analysis results
  - Identifying subsets of pathways that share common information (biological function, mechanisms...): **core-pathways**
    - Hierarchical clustering: binary distance & Ward’s algorithm (linkage criteria)
Step 2: Pathway analysis

Enrichment analysis

▶ Describe the gene signature in a biological meaningful way

▶ Analysis for enrichment of KEGG pathway membership

⇝ Fisher’s exact test

Hierarchical clustering of pathways

▶ Increase the interpretability of enrichment analysis results

▶ Identifying subsets of pathways that share common information (biological function, mechanisms...): core-pathways

⇝ Hierarchical clustering: binary distance & Ward’s algorithm (linkage criteria)
Step 2: Pathway analysis

Enrichment analysis
► Describe the gene signature in a biological meaningful way
  ► Analysis for enrichment of KEGG pathway membership
    ⇝ Fisher’s exact test

Hierarchical clustering of pathways
► Increase the interpretability of enrichment analysis results
  ► Identifying subsets of pathways that share common information (biological function, mechanisms...): core-pathways
    ⇝ Hierarchical clustering: binary distance & Ward’s algorithm (linkage criteria)
Step 2: Pathway analysis

**Enrichment analysis**
- Describe the gene signature in a biological meaningful way
  - Analysis for enrichment of KEGG pathway membership
    - Fisher’s exact test

**Hierarchical clustering of pathways**
- Increase the interpretability of enrichment analysis results
  - Identifying subsets of pathways that share common information (biological function, mechanisms...): **core-pathways**
    - Hierarchical clustering: binary distance & Ward’s algorithm (linkage criteria)
Analysis process: outlines

- Step 1: Identification of biomarkers
- Step 2: Pathway Analysis
- Step 3: Pathway analysis results + Network inference

Normalized data

Signature

Core-pathways 1

Core-pathways 2
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
Step 3: Network Inference
Inferring sparse Gaussian graphical models

Identification of biomarkers

- Select $p$ key genes $\mathcal{P}$
  - $p$ “reasonable” compared to $n$
  - Typically, $n \in [p/5; 5p]$

- The learning dataset
  - $n$ size-$p$ vectors of expression
  - $(X_1, \ldots, X_n)$ with $X_i \in \mathbb{R}^p$
The 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.
  - \( \text{cor}_{ij|P\backslash\{i,j\}} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}} \) for \( i \neq j \)

Interpretation

If 2 nodes \( i \) and \( j \) are partially uncorrelated, no edge is inferred:

\[
X_i \perp \perp X_j | X(P \backslash \{i, j\}) \iff \theta_{ij} = 0
\]

After a simple rescaling \( \Theta \) can be interpreted as the adjacency matrix.
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)
Step 3: Network Inference

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

$\Rightarrow$ 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

$$\hat{\Theta}_\lambda = \arg \max_{\Theta} \mathcal{L}(\Theta; \text{data}) - \lambda \| P_Z \ast \Theta \|_1,$$

where $P_Z$ is a matrix of weights depending on the core-pathway membership $Z$. 
Analysis process: outlines

Step 1
Identification of biomarkers

Step 2
Pathway Analysis

Step 3
Pathway analysis results + Network inference

Normalized data

Signature
Core-pathways 1
Core-pathways 2
Core-pathways

Semi-supervised network
Application to treatment response in basal breast tumors.
Breast cancer: an heterogeneous disease

Sorlie et al. 2003.
Repeated observation of breast tumor subtypes in independent gene expression data sets. *PNAS*

**pathologic Complete Response (pCR)**
*Def.*: No residual invasive cancer in the breast or lymph nodes
*Used as a marker of treatment efficacy*
Breast cancer: an heterogeneous disease

Sorlie et al. 2003.
Repeated observation of breast tumor subtypes in independent gene expression data sets. *PNAS*

Pathologic Complete Response (pCR)
Def.: No residual invasive cancer in the breast or lymph nodes
→ Used as a marker of treatment efficacy
Breast cancer: an heterogeneous disease

Sorlie et al. 2003.
Repeated observation of breast tumor subtypes in independent gene expression data sets. *PNAS*

**pathologic Complete Response (pCR)**

*Def.:* No residual invasive cancer in the breast or lymph nodes

_used as a marker of treatment efficacy*_
Hess et al. 2006.
Pharmacogenomic Predictor of Sensitivity to Preoperative Chemotherapy With Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in Breast Cancer.

Pretreatment gene expression profiling on 133 patients:

<table>
<thead>
<tr>
<th></th>
<th>pCR</th>
<th>not-pCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>Other subtypes</td>
<td>19</td>
<td>85</td>
</tr>
</tbody>
</table>

Table: Clinical Information on the 133 patients includes in the study

Biological issues

Which are the molecular mechanisms underlying the chemo-sensitivity (pCR) or resistance (not-pCR) of basal-like breast tumours?

Pretreatment gene expression profiling on 133 patients:

<table>
<thead>
<tr>
<th></th>
<th>pCR</th>
<th>not-pCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>Other subtypes</td>
<td>19</td>
<td>85</td>
</tr>
</tbody>
</table>

Table: Clinical Information on the 133 patients includes in the study

Biological issues

Which are the molecular mechanisms underlying the chemo-sensitivity (pCR) or resistance (not-pCR) of basal-like breast tumours?
Step 1: Identification of biomarkers

- **100 genes** were identified as differentially expressed ($p < 10^{-3}$)
- **30 first partner genes** have been added to the signature

Step 2: Pathway analysis

- The enrichment test showed **22 KEGG pathways** to be overrepresented ($p < 10^{-2}$)
- **6 core-pathways** have been extracted from the enrichment analysis
Pathway analysis

Tumour cell growth and proliferation mechanism

Adherens Junction

nb: related to TGF-B pathway
(tumour suppressor)
Pathway analysis
Angiogenesis related mechanism

**Calcium signaling pathway**
- Essential role in VEGF

**Axon guidance**
- nb: tumor suppressor via NRP1

**Regulation**

**Potential effect**

**Calcium signaling pathway**
- Calcium induced T Lymphocyte Apoptosis
- Olfactory transduction
- Long term potentiation
- Gliona
- Melanogenesis

**Axon guidance**

**Angiogenesis**
- new blood vessel formation
- induces by hypoxia
- promotes cell motility and distant metastasis

**Renal cell carcinoma**
**MTOR pathway**
**IL8 signaling**
**VEGF signaling pathway**
**Acute myeloid leukemia**
**Insulin signaling pathway**
**Adipokine signaling pathway**
**Olfactory transduction**
**Long term potentiation**
**Gliona**
**Melanogenesis**
**Calcium induced T Lymphocyte Apoptosis**
**Calcium signaling pathway**
**Nitric Oxide Signaling in the Cardiovascular System**

**nb:** Essential role in VEGF

**nb:** Tumour suppressor via NRP1
Regulation of the angiogenesis activity via CALM3

CALM3 gene regulates the activity of AKT1 in breast tumors (Coticchia et al. 2009). Our network suggests that

- this regulation occurs only in the pCR condition
- it may be broken in not-pCR tumors
Results
Mechanism of resistance to chemotherapy

- Hypoxia
- WNT
- AKT
- MTOR
- RAPTOR
- LST8
- HIF
- VEGF
- Angiogenesis
- Tumour growth
- CALM3
- CAMK2G
- pCR
- not-pCR
Conclusion

Improved gene selection

- robustness and reproducibility of the signature

Framework to infer networks on the basis of a biological informative prior over network structures

- less sensitive to the noise inherent to biological data
- reduce the space of possible network
- more relevant network (interpretability)

Breast cancer study:

- key regulations in cancer progression and response to treatment
Conclusion

Improved gene selection
▶ robustness and reproducibility of the signature

Framework to infer networks on the basis of a biological informative prior over network structures
▶ less sensitive to the noise inherent to biological data
▶ reduce the space of possible network
▶ more relevant network (interpretability)

Breast cancer study:
▶ key regulations in cancer progression and response to treatment
Conclusion

Improved gene selection

▶ robustness and reproducibility of the signature

Framework to infer networks on the basis of a biological informative prior over network structures

▶ less sensitive to the noise inherent to biological data
▶ reduce the space of possible network
▶ more relevant network (interpretability)

Breast cancer study:

▶ key regulations in cancer progression and response to treatment
Acknowledgement

Statistic & Genome Laboratory
Christophe Ambroise
Julien Chiquet
Bernard Prum

Pharnext
Mickaël Guedj
The Dream Team: Matthieu, Caroline and Fabrice
Serguei Nabirotchkin
Ilya Chumakov