



**HAL**  
open science

## Sparse factor model for gene co-expression networks

Anne Blum, Magalie Houee, Aldons J. Lulis, Sandrine Lagarrigue, David Causeur

► **To cite this version:**

Anne Blum, Magalie Houee, Aldons J. Lulis, Sandrine Lagarrigue, David Causeur. Sparse factor model for gene co-expression networks. BioNetVisA Workshop, Sep 2014, Strasbourg, France. hal-01210988

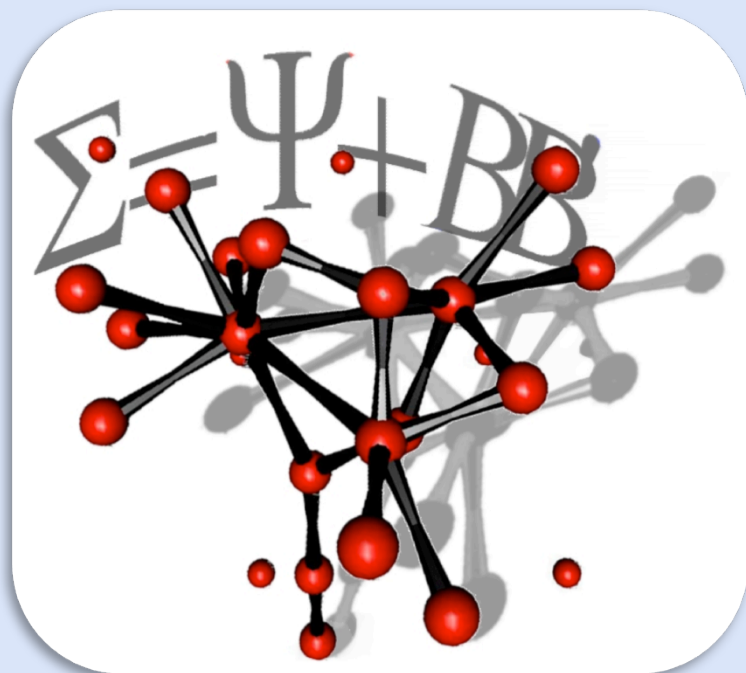
**HAL Id: hal-01210988**

**<https://hal.science/hal-01210988>**

Submitted on 5 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# Sparse factor model for gene co-expression networks

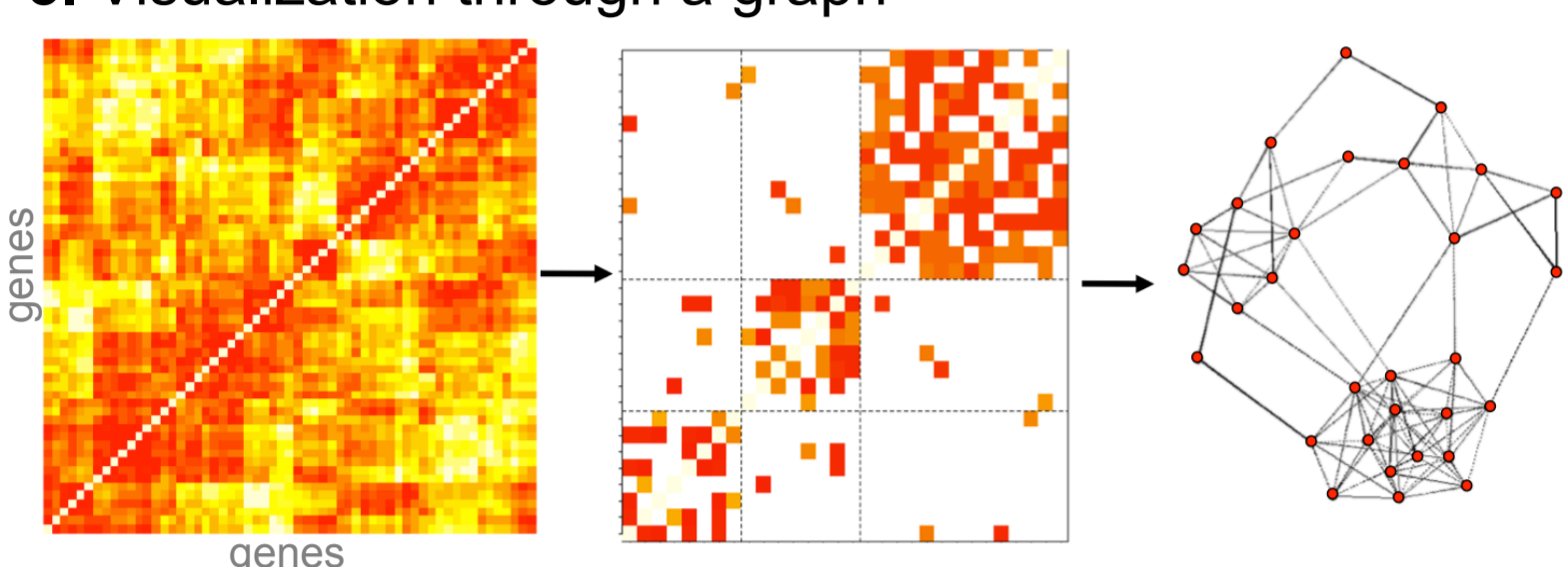
Y. Blum (1), M. Houée (2), AJ. Lusis (1), S. Lagarrigue (3), D. Causeur (2)

(1) Department of Medicine/Division of Cardiology, David Geffen School of Medicine, UCLA, Los Angeles, CA USA, (2) Agrocampus Ouest- Applied Mathematics Department, Rennes, France, (3). Agrocampus Ouest - UMR PEGASE INRA, Agrocampus Ouest, Rennes, France

## Gene co-expression network

**Network construction** (for undirected graphical models)

1. Choose a measure L of the link between 2 genes
2. Decision rule: is L(y<sub>i</sub>, y<sub>j</sub>) different from 0 ?
3. Visualization through a graph



### Measure of the link

**Linear measures**

• **Pearson correlation:**  $corr(y_i, y_j) = \frac{Cov(y_i, y_j)}{\sqrt{Var(y_i)Var(y_j)}}$

simple and intuitive

→ **Co-expression networks**

• **Partial correlation:**  $corr(y_i, y_j | y_{i,j})$

→ Gaussian Graphical Model (GGM)

**Non-linear measures**

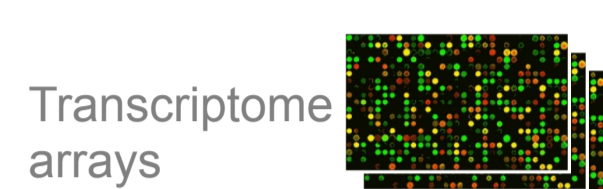
• **Mutual information MI:**

$MI(x, y) = \sum_{i=1}^I \sum_{j=1}^J P(x=i, y=j) \log \frac{P(x=i, y=j)}{P(x=i)P(y=j)}$

→ Information-theory-based methods

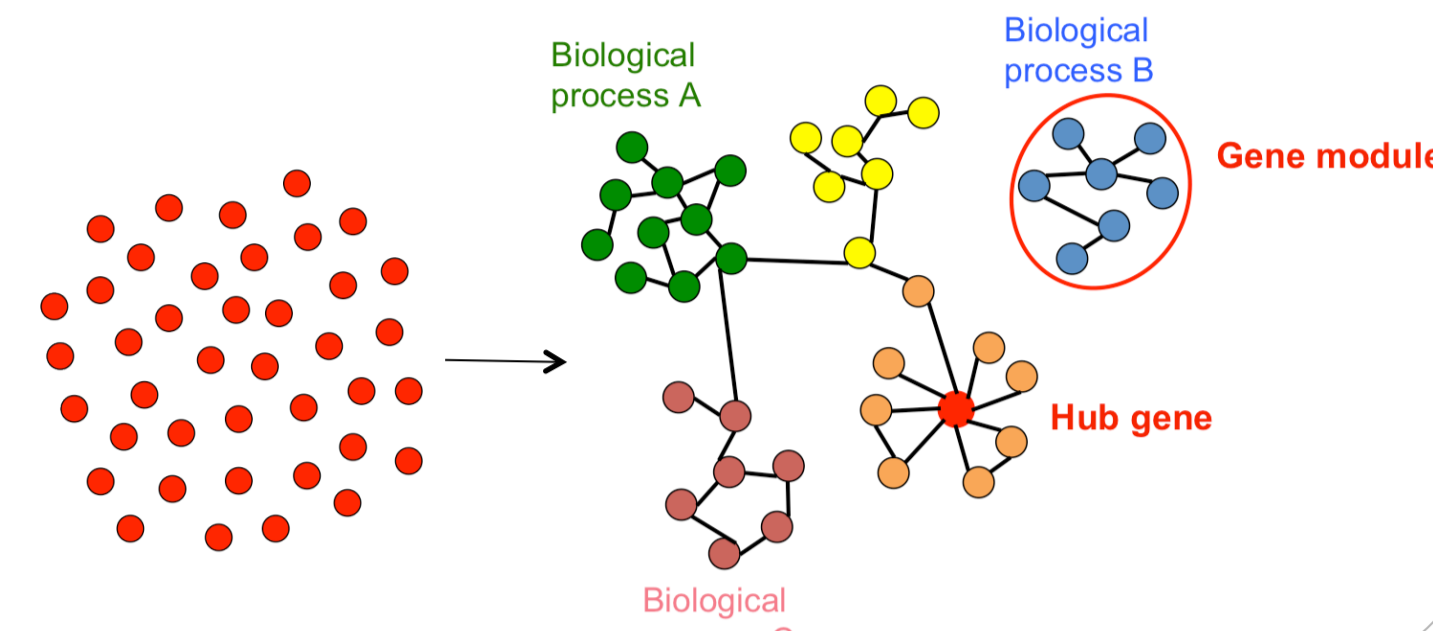
### Challenges

**High dimension:**  $n \ll m$



Gene expression data

**Sparsity assumption:** within a set of genes, only a few are interacting (Tegner et al 2003)



## Sparse factor model

### Factor model

**Key idea:** to structure the dependence for correlation estimation (less parameters to estimate).

Y dataset with n rows (individuals) and m columns (genes).

$Y \sim N_m(\mu, \Sigma)$

Correlations between genes are described by a small number **q** of factors containing a common dependence:

$$\Sigma = \Psi + BB'$$

where  $\Psi$  is a diagonal matrix and B represents the  $m \times q$  matrix of loadings

### Estimation of the parameters

Estimation of  $\Psi$  and B: **EM algorithm** (Rubin & Thayer 1982)

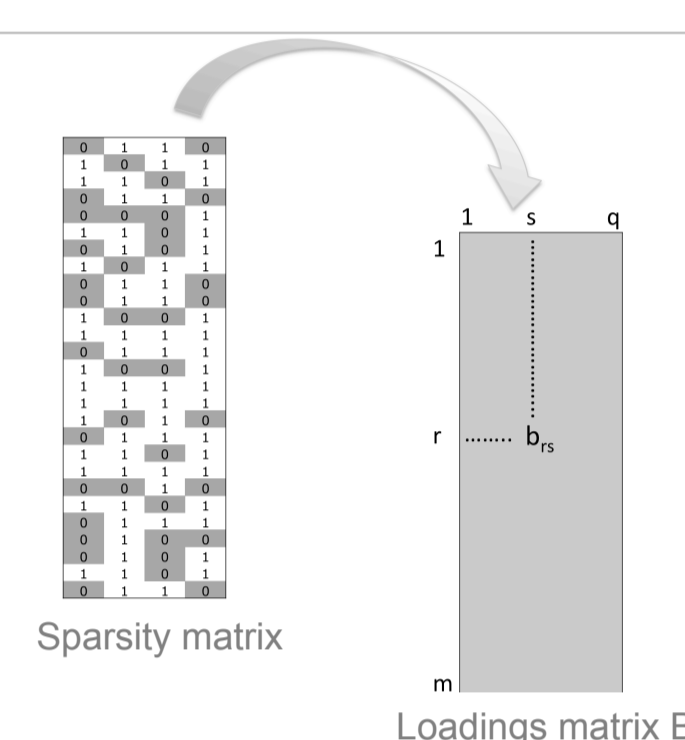
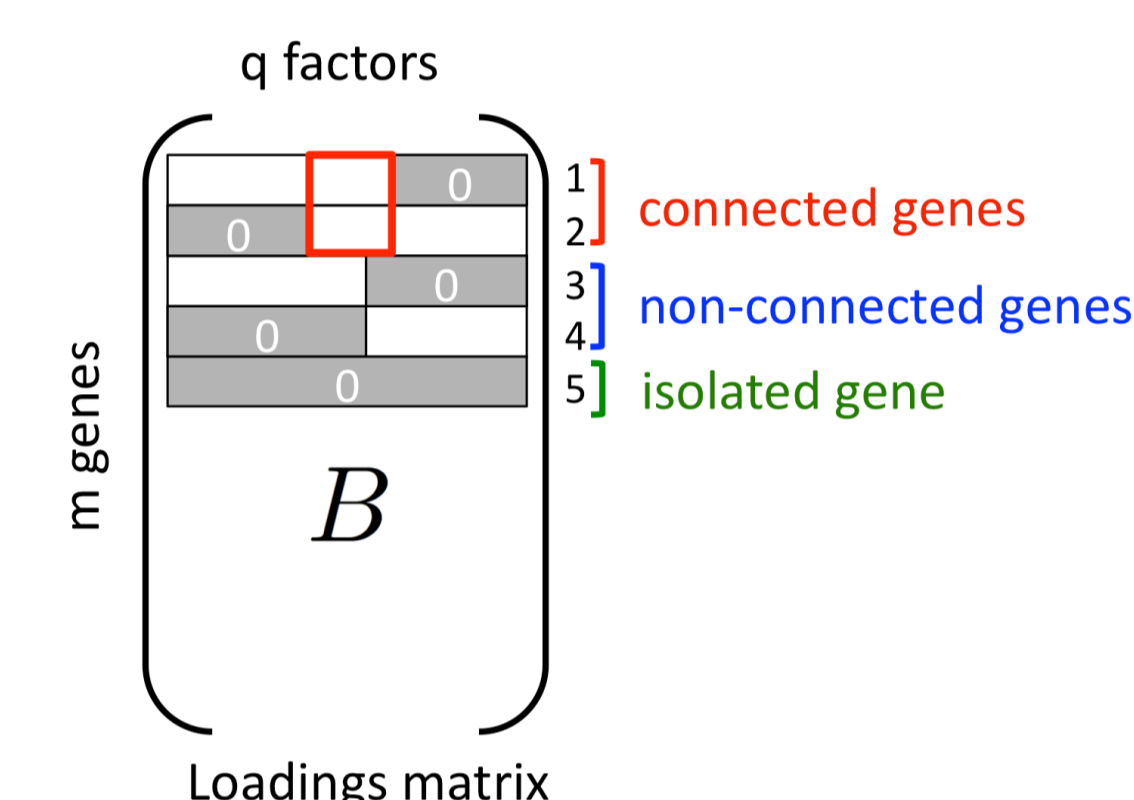
Deviance:

$$D(\Psi, B) \propto \log \det(\Psi + BB') + \text{trace}[S(\Psi + BB')^{-1}]$$

→ Calculate and minimize the expectation of  $D(\Psi, B; Z)$

### Sparse adapted model

The topology of the network can be deduced from the loadings matrix B:



### Adapted EM algorithm

**M-step:** minimizing the expected deviance under sparsity constraints:

$$D(\Psi, B; Z) + \lambda' R' \text{vec}(B) \quad \text{Lagrange multiplier approach}$$

where  $\lambda$  is the vector of Lagrange multipliers and R the constraints matrix deduced from the sparsity matrix.

## Sparsity using LASSO penalization

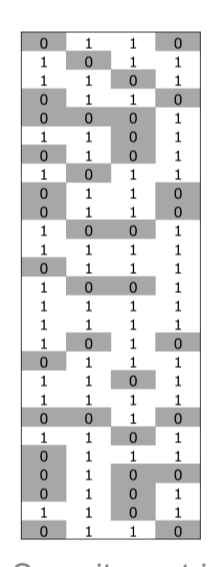
### Inference on the sparsity

**L<sub>1</sub>- regularization**

$$\min_{\Psi, B} D(\Psi, B) + \lambda \sum_{k=1}^m \sum_{i=1}^q |b_{ki}|$$

$\lambda$  chosen by minimization of BIC

$$BIC(\lambda) = D(\hat{\Psi}_\lambda, \hat{B}_\lambda) + 2\#\{(k, i), \hat{b}_{\lambda, ki} \neq 0\}$$



### Comparison study

**Comparable approaches:**

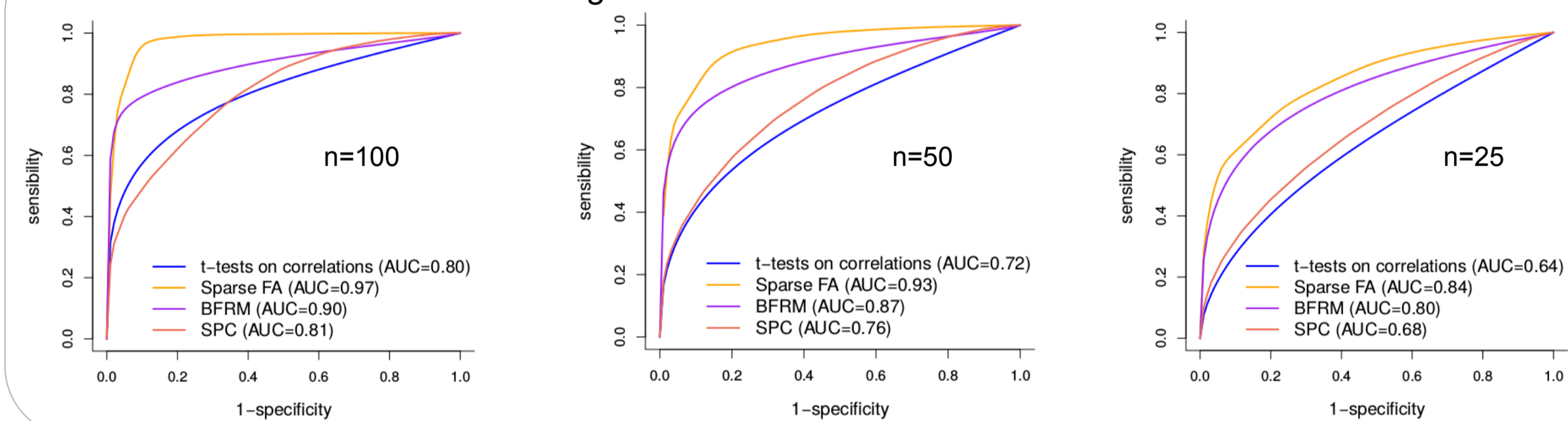
- **SPC:** Sparse Principal Components (Witten et al, *Biostatistics* 2009)
- **BFRM:** Bayesian Factor Regression Model (West, *Bayesian Statistics* 2003)

### Simulations

1000 datasets simulated from this true correlation structure (using  $n=10000$ )  
**5 modules:**  $m_1=200$  genes,  $m_2=150$ ,  $m_3=80$ ,  $m_4=60$ ,  $m_5=40$



Identification of the true edges: Mean ROC curves over 1000 simulations



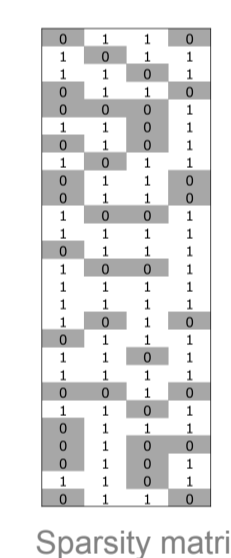
⇒ **Better estimation of the network**

## Sparsity using biological knowledge

### Inference on the sparsity

**Biological information:**

- Gene Ontology
- Transcription factors,
- eQTL regulation
- ...



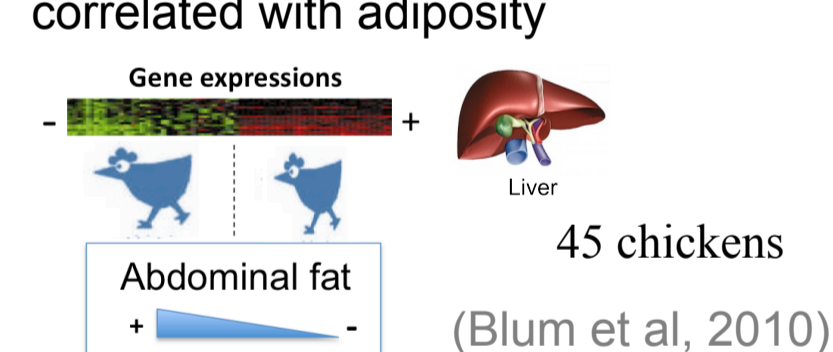
	Term 1	Term 2	Term 3	...	Term X
gene 1	0	1	1	...	0
gene 2	1	0	0	...	0
gene 3	0	0	1	...	1
...	...	...	...	...	...
gene X	1	0	0	...	0

Example of biological constrain matrix

### Example of application

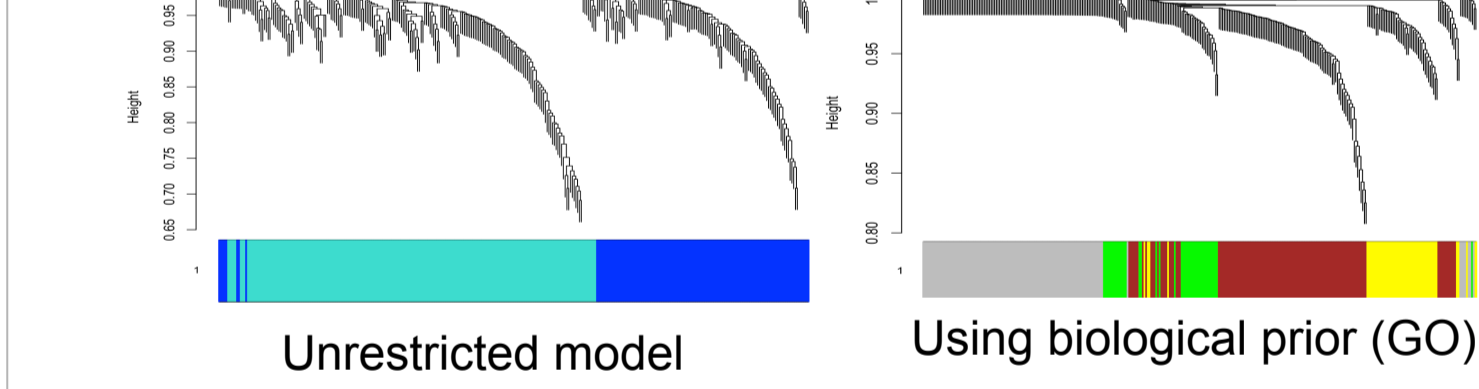
**Gene expression dataset:**

~ 400 hepatic genes expression correlated with adiposity



**Gene modules detection using WGCNA** (Horvath et al 2008)

Gene dendrograms (WGCNA output)



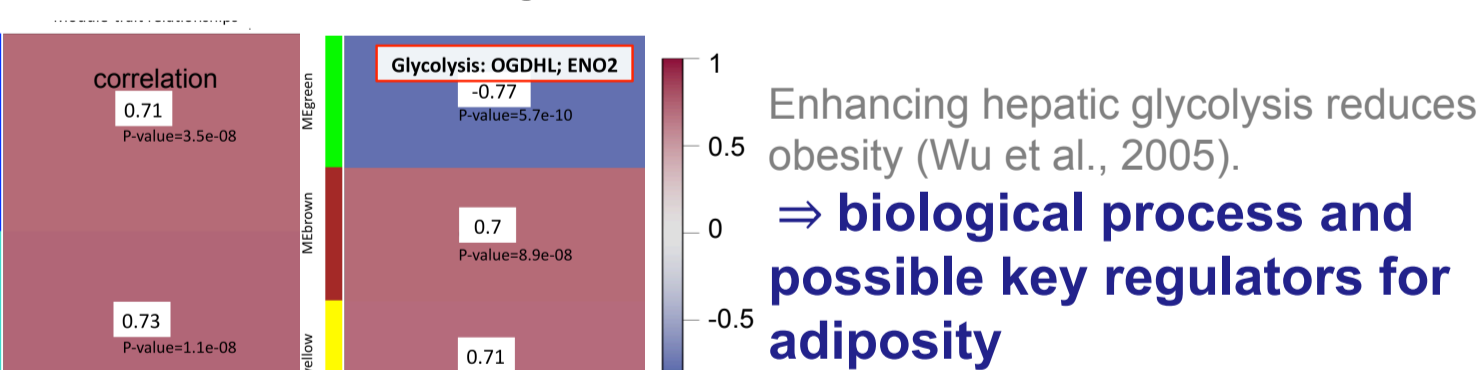
⇒ **a more precise structure is revealed**

### Using Biological information only



⇒ **the model account both for biological knowledge and expression intensity**

### Module-adiposity relationship



Enhancing hepatic glycolysis reduces obesity (Wu et al., 2005).  
 ⇒ **biological process and possible key regulators for adiposity**

⇒ **New biological insight in gene module detection**

## FANET sparse Factor Analysis for Network modeling R package

**emfas.ccdlasso:** this function proposes a LASSO-regularized EM algorithm using a cyclic coordinate descent algorithm to fit a sparse factor model for correlation



**INPUT**  
**data:** gene expression dataset (rows=genes)

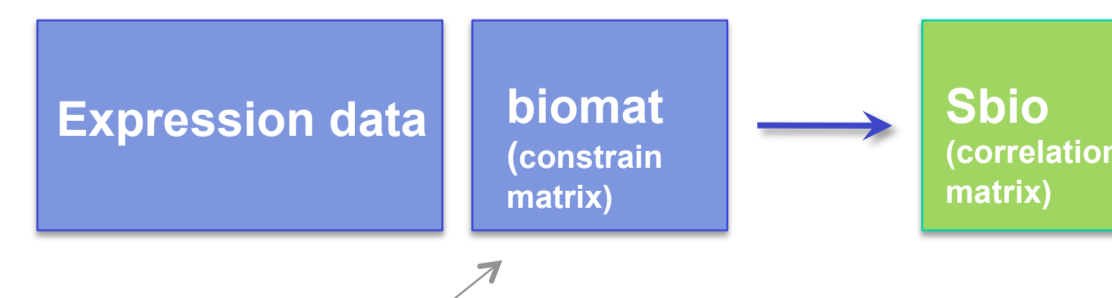
**USAGE**  
**emfas.ccdlasso**(data, maxnbf=10, min.err = 1e-06, max.iter = 100)

**maxnbf:** maximum number of factors in the model  
**min.err:** stopping criterion value for iterations in ccd algorithm  
**max.iter:** stopping criterion value for iterations in ccd algorithm

**OUTPUT**  
**Slasso:** estimated correlation matrix

## FANET sparse Factor Analysis for Network modeling R package

**SparseEmfaBio:** this function uses the biological constrains matrix to introduce sparsity in the factor model for correlation.



**biomat:** constructed by the user or by using the function `Construct.BioMatrix` (uses `biomat` R package)

**INPUT**  
**data:** gene expression dataset (rows=genes)  
**biomat:** biological constrain matrix

**USAGE**  
**SparseEmfaBio**(data,biomat, min.err = 1e-06)

**min.err:** Stopping criterion value for iterations

**OUTPUT**  
**Sbio:** estimated correlation matrix

## References

**More about Factor Analysis for expression data:**

- Blum Y, et al. A Factor Model to Analyze Heterogeneity in Gene Expression. *BMC Bioinformatics*, 2010, 11:368.
- C. Friguet, et al. A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104(488):1406–1415, 2009.

**WGCNA R package:**

P. Langfelder and S. Horvath. *Wgcna: an r package for weighted correlation network analysis*. *BMC bioinformatics*, 9(1):559, 2008.

## R package FANet and tutorial

yuna-blum.com/links/FANet