Construction of gene regulatory network from Microarray data

Jérôme Ambroise - SPS – Nov 17 2010

Construction of gene regulatory network from Microarray data

- Biological introduction
  - DNA, RNA, Protein
  - Transcription Factors and Gene Regulatory network
  - Measuring RNA concentration: The Microarray technology

- Gene network reconstruction methods from microarray data
  - Relevance Network (Correlation)
  - Gaussian Graphicals Models (Partial Correlation)
  - SIRENE (Supervised approach: local models)
  - TNIFSED (Supervised approach: global model)

- Evaluation of predictive peformances
  - Area Under the Curve of the Receiver Operating Characteristic (AUC)

- Results and conclusions

Construction of gene regulatory network from Microarray data

- Biological introduction
  - DNA, RNA, Protein
  - Transcription Factors and Gene Regulatory network
  - Measuring RNA concentration: The Microarray technology

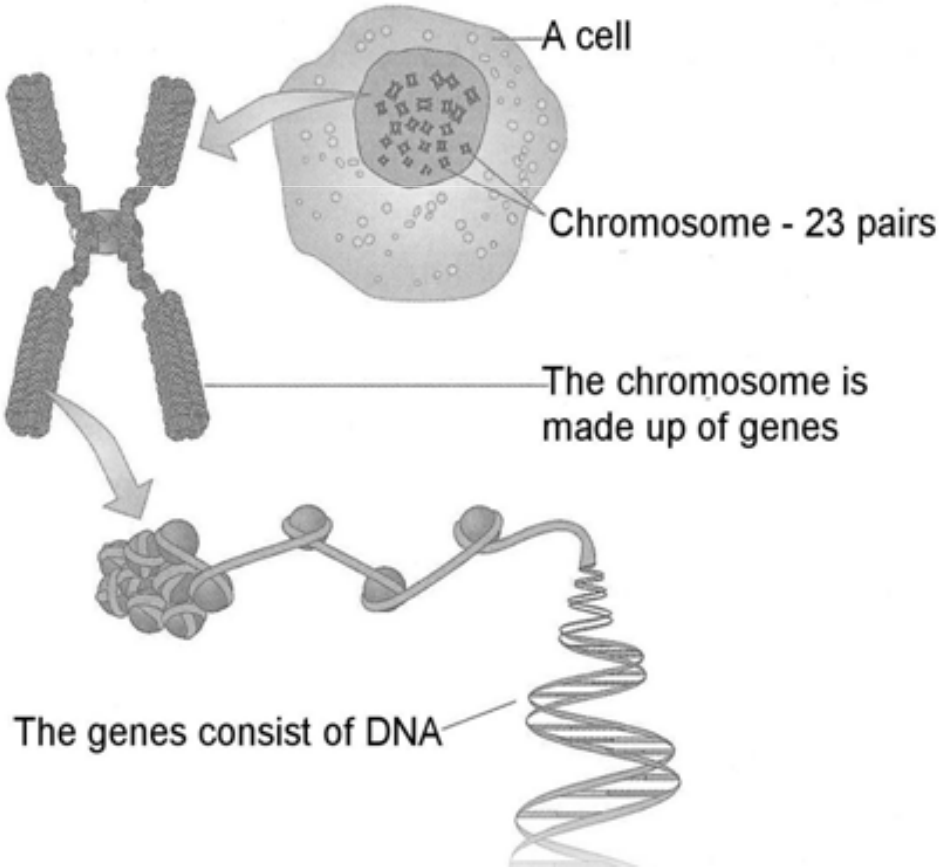
- Gene network reconstruction methods from microarray data
  - Relevance Network (Correlation)
  - Gaussian Graphicals Models (Partial Correlation)
  - SIRENE (Supervised approach: local models)
  - TNIFSED (Supervised approach: global model)

- Evaluation of predictive peformances
  - Area Under the Curve of the Receiver Operating Characteristic (AUC)

- Results and conclusions

DNA, RNA and Proteins

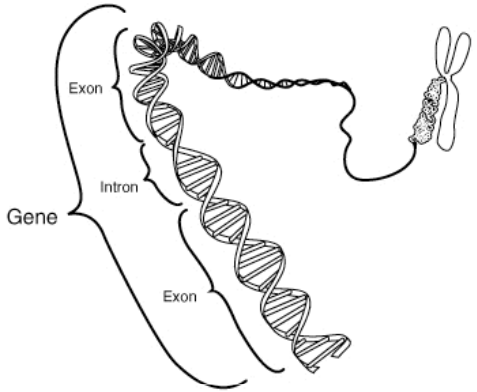
- The human body is composed of a huge number of cells : ~ 10^{14} cells
- Each cell contains 46 Chromosomes constituted of DNA molecules.
- DNA contains the information necessary to the organism development.

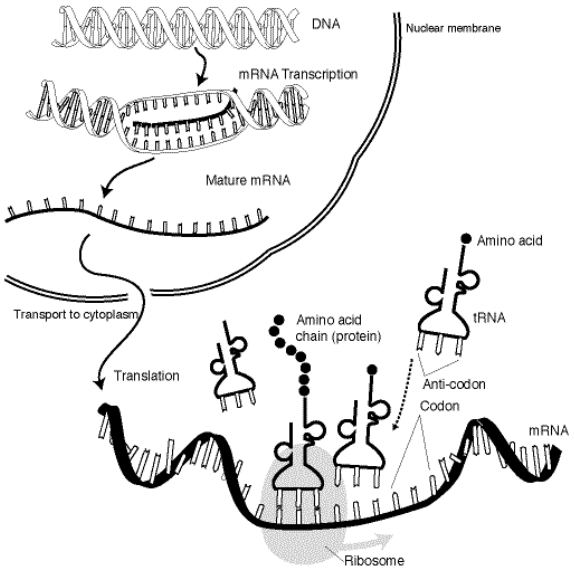


DNA, RNA and Proteins

• DNA is composed of coding (gene) and non-coding regions

• The number of human genes is estimated between 20.000 and 25.000





• gene = sequence of DNA specifying the synthesis of a protein

• Cells can produce proteins in two steps

1: transcription: Production of $_{m}$RNA

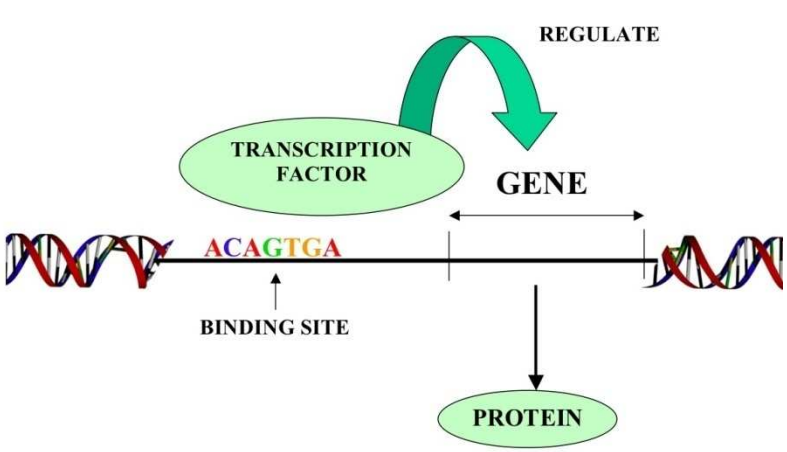
2: Traduction: Traduction of $_{m}$RNA in a protein

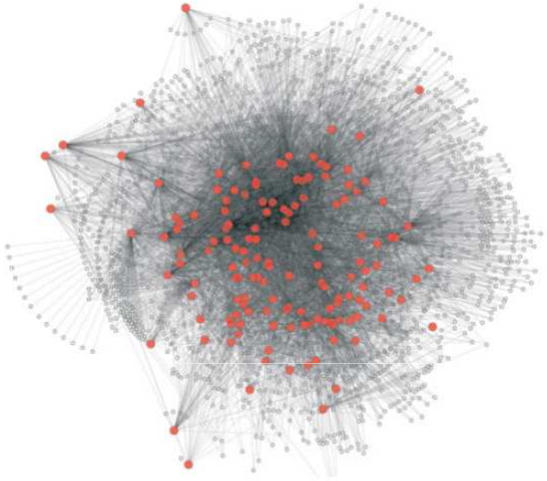
• Transcription is a highly regulated processus.

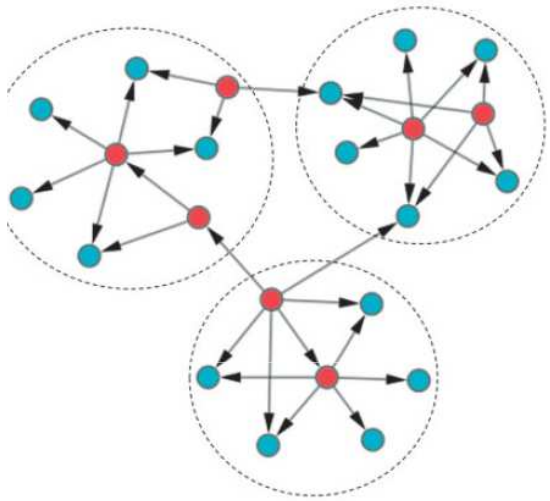
• RNA_m quantity ~ Gene Expression level ~ Gene activity

Transcription Factors – Gene Regulatory Network

Gene activity regulation: Interaction between transcription factors (specific proteins) and their target genes.







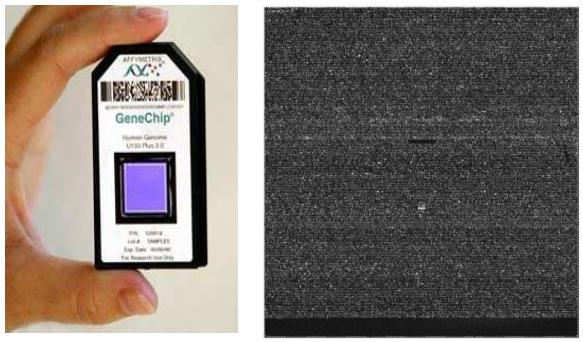
Significant clinical applications:

- More than 150 transcription factors are associated with 300 human diseases.

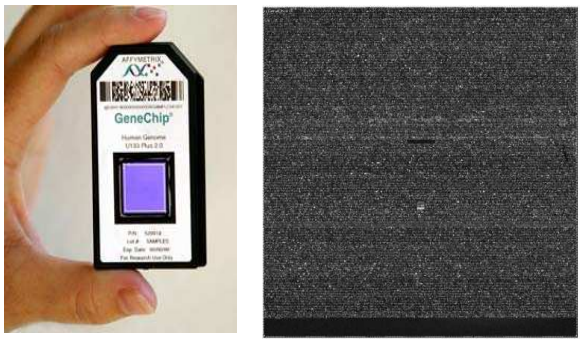
- Disease in human can arise from from mutation of cis-regulatory elements. It can lead to more profound effect than mutation in the coding region.

Gene Expression Microarrays

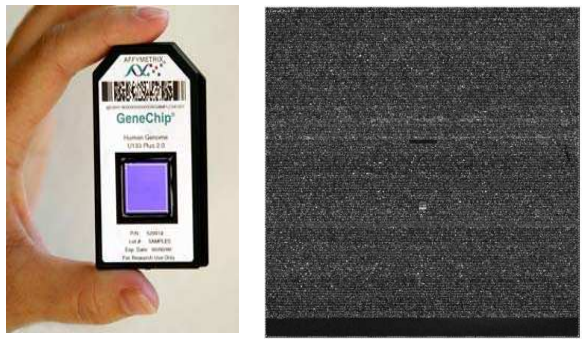
Cellular culture : 10° C



Cellular culture : 20° C



Cellular culture : 30° C



P ~ 20.000 genes

CONDITION	gene 1	gene 2	gene 3	gene 4	gene 5	gene 6	gene 7	gene 8	gene 9	gene 10	gene 11	gene 12	...	gene 22000
10°C	4.795	4.737	4.343	9.348	10.209	10.209	6.032	10.209	10.209	8.693	10.209	10.209	...	10.365
20°C	6.204	5.694	6.032	7.103	6.649	6.649	6.649	6.204	6.649	6.649	6.649	6.649	...	6.204
30°C	8.693	8.488	8.503	10.365	6.086	6.086	6.649	6.086	4.795	6.032	6.204	4.795	...	6.649

Construction of gene regulatory network from Microarray data

- Biological introduction
  - DNA, RNA, Protein
  - Transcription Factors and Gene Regulatory network
  - Measuring RNA concentration: The Microarray technology

- Gene network reconstruction methods from microarray data
  - Relevance Network (Correlation)
  - Gaussian Graphicals Models (Partial Correlation)
  - SIRENE (Supervised approach: local models)
  - TNIFSED (Supervised approach: global model)

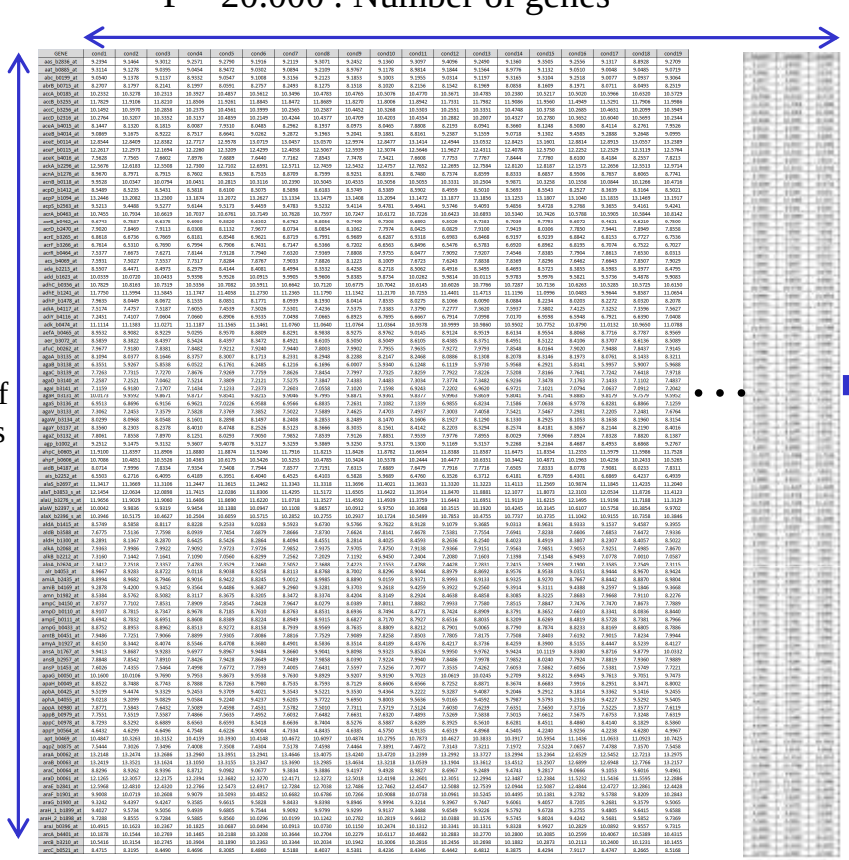
- Evaluation of predictive peformances
  - Area Under the Curve of the Receiver Operating Characteristic (AUC)

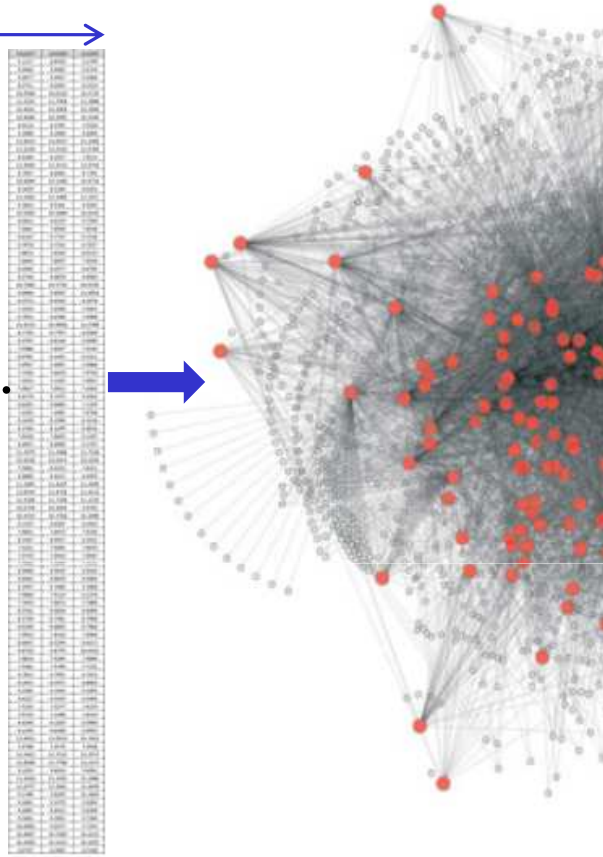
- Results and conclusions

From gene expression data to gene regulation network

P ~ 20.000 : Number of genes

N = 300
Number of
conditions





Gene Regulatory Network Inference:

1: Relevance Network and Gaussian Graphical Models (Unsupervised)

2: Bayesian Network (Unsupervised)

3: Ordinary Differentiall equations (Unsupervised)

4 : SIRENE: Supervised inference of relevance network

5 : TNIFSED: Transcriptional network inference from functional similarity and expression data

Relevance Network : Correlation coefficients

Gene Expression matrix

$$X \propto N_5(\mu, \Sigma)$$

$$\mu = (\mu_1, \ldots . \mu_5)$$

$$\Sigma : \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} \end{pmatrix}$$

CONDITION	gene1	gene2	gene3	gene4	gene5
cond.1	4.795	4.737	4.343	9.348	10.209
cond.2	6.204	5.694	6.032	7.103	6.649
cond.3	8.693	8.488	8.503	10.365	6.086
cond.4	8.653	8.224	8.137	7.819	3.920
cond.5	6.920	6.658	7.102	8.045	11.013
cond.6	7.220	6.984	6.753	11.953	5.350
cond.7	7.996	9.139	9.335	10.703	6.111
cond.8	7.338	7.347	7.657	8.129	6.589
cond.9	8.570	8.518	8.631	6.234	5.324
cond.10	3.265	3.014	3.849	10.455	9.041
cond.11	5.671	4.924	5.745	5.339	13.020
cond.12	8.041	8.496	7.979	9.891	5.161
cond.13	9.008	8.273	8.417	7.436	4.487
cond.14	4.081	4.460	3.735	11.163	8.953
cond.15	10.274	10.304	10.061	10.702	6.180
cond.16	7.747	8.185	7.917	10.709	5.217
cond.17	8.720	8.374	8.735	10.939	8.401
cond.18	2.107	1.882	2.226	10.684	11.670
cond.19	7.464	8.043	8.729	5.482	5.371
cond.20	5.742	5.534	5.561	4.504	9.452

Covariance Matrix

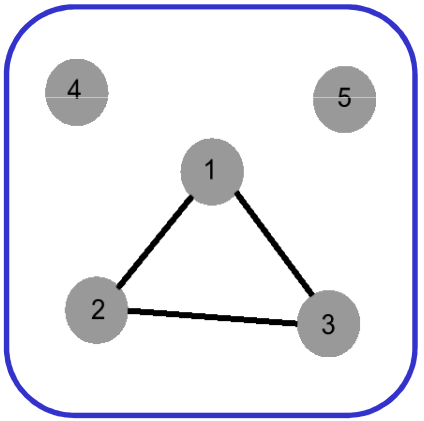
$$\hat{\Sigma} = \hat{\sigma_{ij}} = \frac{1}{N-1}(X - \bar{X})'(X - \bar{X})$$

	gene1	gene2	gene3	gene4	gene5
gene1	4.418	4.507	4.345	-0.235	-3.882
gene2	4.507	4.813	4.587	0.081	-4.231
gene3	4.345	4.587	4.544	-0.332	-3.851
gene4	-0.235	0.081	-0.332	5.048	-0.607
gene5	-3.882	-4.231	-3.851	-0.607	6.852

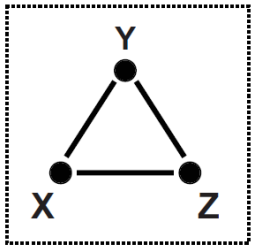
Correlation Matrix

$$\hat{\rho_{ij}} = \frac{\hat{\sigma_{ij}}}{\sqrt{\hat{\sigma_{ii}}\hat{\sigma_{jj}}}}$$

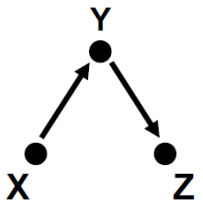
	gene1	gene2	gene3	gene4	gene5
gene1	1.000	0.978	0.970	-0.050	-0.705
gene2	0.978	1.000	0.981	0.017	-0.737
gene3	0.970	0.981	1.000	-0.069	-0.690
gene4	-0.050	0.017	-0.069	1.000	-0.103
gene5	-0.705	-0.737	-0.690	-0.103	1.000

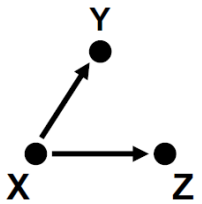


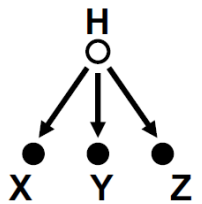
Coexpression



Regulatory network







Gaussian Graphical Models: Partial Correlation coefficients

Gene Expression matrix

$$X \propto N_5(\mu, \Sigma)$$

$$\mu = (\mu_1, ...\mu_5)$$

$$\Sigma : \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} \end{pmatrix}$$

CONDITION	gene1	gene2	gene3	gene4	gene5
cond.1	4.795	4.737	4.343	9.348	10.209
cond.2	6.204	5.694	6.032	7.103	6.649
cond.3	8.693	8.488	8.503	10.365	6.086
cond.4	8.653	8.224	8.137	7.819	3.920
cond.5	6.920	6.658	7.102	8.045	11.013
cond.6	7.220	6.984	6.753	11.953	5.350
cond.7	7.996	9.139	9.335	10.703	6.111
cond.8	7.338	7.347	7.657	8.129	6.589
cond.9	8.570	8.518	8.631	6.234	5.324
cond.10	3.265	3.014	3.849	10.455	9.041
cond.11	5.671	4.924	5.745	5.339	13.020
cond.12	8.041	8.496	7.979	9.891	5.161
cond.13	9.008	8.273	8.417	7.436	4.487
cond.14	4.081	4.460	3.735	11.163	8.953
cond.15	10.274	10.304	10.061	10.702	6.180
cond.16	7.747	8.185	7.917	10.709	5.217
cond.17	8.720	8.374	8.735	10.939	8.401
cond.18	2.107	1.882	2.226	10.684	11.670
cond.19	7.464	8.043	8.729	5.482	5.371
cond.20	5.742	5.534	5.561	4.504	9.452

Covariance Matrix

$$\hat{\Sigma} = \hat{\sigma_{ij}} = \frac{1}{N-1}(X - \bar{X})'(X - \bar{X})$$

	gene1	gene2	gene3	gene4	gene5
gene1	4.418	4.507	4.345	-0.235	-3.882
gene2	4.507	4.813	4.587	0.081	-4.231
gene3	4.345	4.587	4.544	-0.332	-3.851
gene4	-0.235	0.081	-0.332	5.048	-0.607
gene5	-3.882	-4.231	-3.851	-0.607	6.852

Concentration Matrix

$$\hat{\Omega} = (\hat{\omega_{ij}}) = \hat{\Sigma}^{-1}$$

	gene1	gene2	gene3	gene4	gene5
gene1	5.781	-4.582	-0.923	0.277	-0.049
gene2	-4.582	11.526	-6.749	-0.764	0.661
gene3	-0.923	-6.749	7.674	0.531	-0.331
gene4	0.277	-0.764	0.531	0.259	0.006
gene5	-0.049	0.661	-0.331	0.006	0.341

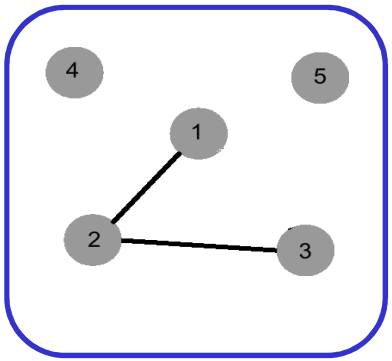
Correlation Matrix

	gene1	gene2	gene3	gene4	gene5
gene1	1.000	0.978	0.970	-0.050	-0.705
gene2	0.978	1.000	0.981	0.017	-0.737
gene3	0.970	0.981	1.000	-0.069	-0.690
gene4	-0.050	0.017	-0.069	1.000	-0.103
gene5	-0.705	-0.737	-0.690	-0.103	1.000

Partial Correlation Matrix

$$\hat{\pi_{ij}} = -\frac{\hat{\omega_{ij}}}{\sqrt{\hat{\omega_{ii}}\hat{\omega_{jj}}}}$$

	gene1	gene2	gene3	gene4	gene5
gene1	1.000	0.561	0.139	-0.226	0.035
gene2	0.561	1.000	0.718	0.442	-0.333
gene3	0.139	0.718	1.000	-0.377	0.204
gene4	-0.226	0.442	-0.377	1.000	-0.021
gene5	0.035	-0.333	0.204	-0.021	1.000



Gaussian Graphical Models: Partial Correlation coefficients

Gene Expression matrix

$$X \propto N_5(\mu, \Sigma)$$

$$\mu = (\mu_1, ...\mu_5)$$

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} \end{pmatrix}$$

CONDITION	gene1	gene2	gene3	gene4	gene5
cond.1	4.795	4.737	4.343	9.348	10.209
cond.2	6.204	5.694	6.032	7.103	6.649
cond.3	8.693	8.488	8.503	10.365	6.086
cond.4	8.653	8.224	8.137	7.819	3.920
cond.5	6.920	6.658	7.102	8.045	11.013
cond.6	7.220	6.984	6.753	11.953	5.350
cond.7	7.996	9.139	9.335	10.703	6.111
cond.8	7.338	7.347	7.657	8.129	6.589
cond.9	8.570	8.518	8.631	6.234	5.324
cond.10	3.265	3.014	3.849	10.455	9.041
cond.11	5.671	4.924	5.745	5.339	13.020
cond.12	8.041	8.496	7.979	9.891	5.161
cond.13	9.008	8.273	8.417	7.436	4.487
cond.14	4.081	4.460	3.735	11.163	8.953
cond.15	10.274	10.304	10.061	10.702	6.180
cond.16	7.747	8.185	7.917	10.709	5.217
cond.17	8.720	8.374	8.735	10.939	8.401
cond.18	2.107	1.882	2.226	10.684	11.670
cond.19	7.464	8.043	8.729	5.482	5.371
cond.20	5.742	5.534	5.561	4.504	9.452

Multiple linear regression models

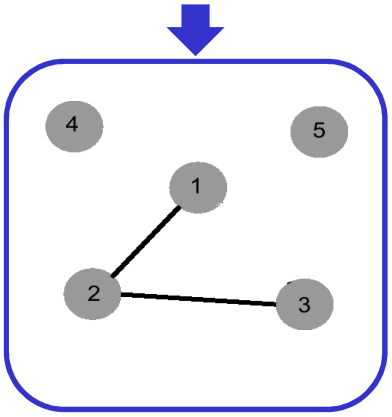
$$\text{minimize} \sum_{i=1}^{n} (y_i - \beta^\top \mathbf{z}_i)^2$$

$$Gene1_n = \beta_2^1\ Gene2_n + \beta_3^1\ Gene3_n + \beta_4^1\ Gene4_n + \beta_5^1\ Gene5_n + \epsilon_n$$
$$Gene2_n = \beta_1^2\ Gene1_n + \beta_3^2\ Gene3_n + \beta_4^2\ Gene4_n + \beta_5^2\ Gene5_n + \epsilon_n$$
$$Gene3_n = \beta_1^3\ Gene1_n + \beta_2^3\ Gene2_n + \beta_4^3\ Gene4_n + \beta_5^3\ Gene5_n + \epsilon_n$$
$$Gene4_n = \beta_1^4\ Gene1_n + \beta_2^4\ Gene2_n + \beta_3^4\ Gene3_n + \beta_5^4\ Gene5_n + \epsilon_n$$
$$Gene5_n = \beta_1^5\ Gene1_n + \beta_2^5\ Gene2_n + \beta_3^5\ Gene3_n + \beta_4^5\ Gene4_n + \epsilon_n$$

Partial Correlation Matrix

$$\hat{\pi_{ij}} = sign(\hat{\beta_j^i}) \sqrt{\hat{\beta_j^i}\ \hat{\beta_i^j}}$$

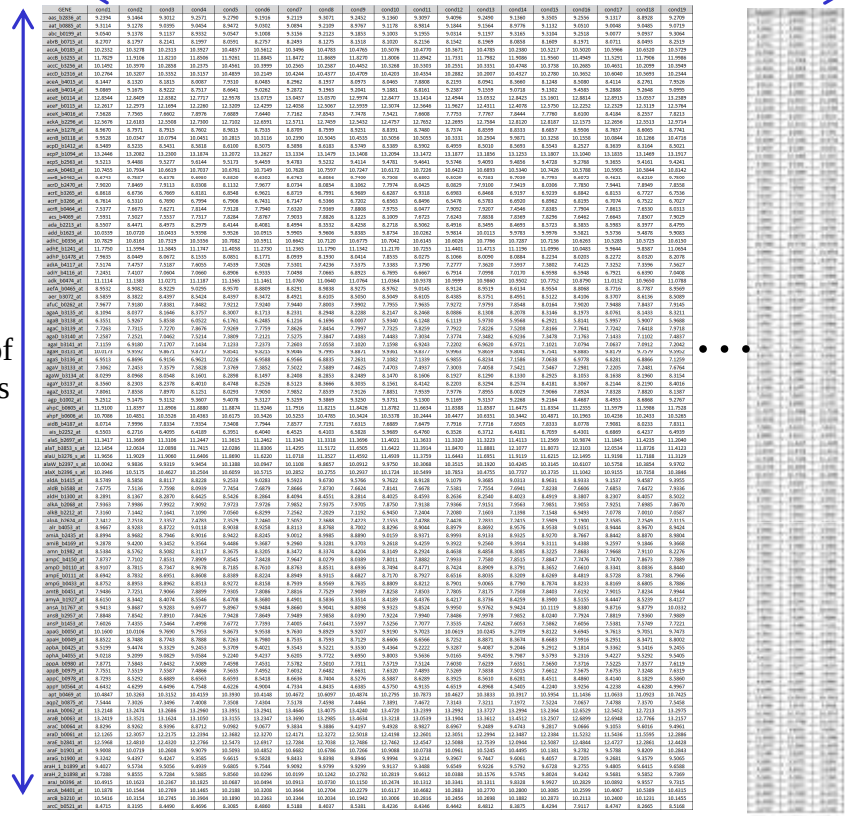
	gene1	gene2	gene3	gene4	gene5
gene1	1.000	0.561	0.139	-0.226	0.035
gene2	0.561	1.000	0.718	0.442	-0.333
gene3	0.139	0.718	1.000	-0.377	0.204
gene4	-0.226	0.442	-0.377	1.000	-0.021
gene5	0.035	-0.333	0.204	-0.021	1.000



Gaussian Graphical Models … when P >> N

P ~ 20.000 : Number of genes

N = 300
Number of conditions



1: Inversion of covariance matrix:

- P >> N : Covariance matrix is not positive-definite and cannot be inverted

$$\hat{\Sigma} = \hat{\sigma_{ij}} = \frac{1}{N-1}(X - \bar{X})'(X - \bar{X})$$

$$\hat{\Omega} = (\hat{\omega_{ij}}) = \hat{\Sigma}^{-1}$$

- Solution (Schafer, Strimmer)

$$\hat{\Sigma}^* = \lambda\, T + (1-\lambda)\, \hat{\Sigma}$$

2: Multiple Linear Regression models:

- P >> N : Multiple linear model with more variables (P) than observation (N)

$$\text{minimize} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}^\top \mathbf{z}_i)^2$$

- Solutions: Penalized estimation

Ridge: $\text{minimize}\ (y - \mathbf{Z}\beta)^\top (y - \mathbf{Z}\beta)\ \text{s.t.}\ \sum_{j=1}^{p} \beta_j^2 \le t$

Lasso: $\text{minimize}\ (\mathbf{y} - \mathbf{Z}\beta)^\top (\mathbf{y} - \mathbf{Z}\beta)\ \text{s.t.}\ \sum_{j=1} |\beta_j| \le t$

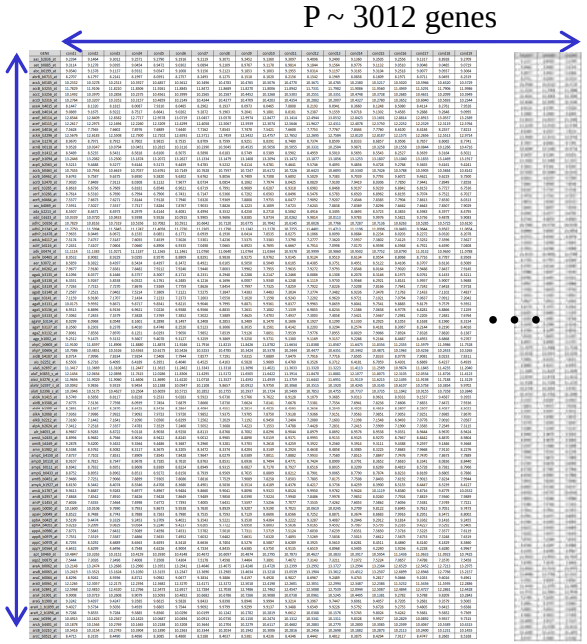
Introduction of bias and reduction of variability

Schafer, J. & Strimmer, K. (2005*b*) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, **4** (1), 1175.

Kramer, N., Schafer, J. & Boulesteix, A. (2009) Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC bioinformatics*, **10** (1), 384.

SIRENE: Local Supervised Model

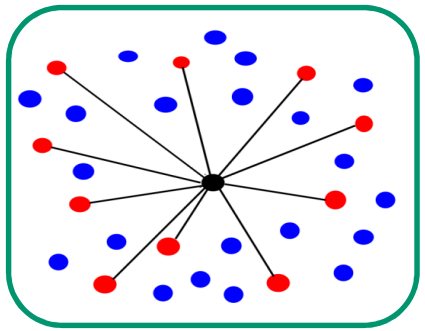
RN , GGM, Bay. Network: UNSUPERVISED => poor predictive performance
SIRENE : SUPERVISED Inference of Regulatory Network
Application of SIRENE on the E.coli bacterium.



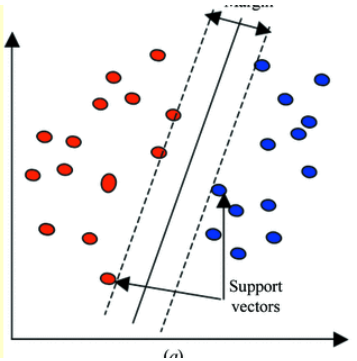
1 model for each transcription factor (132 models)

CONDITION	gene1	gene2	gene3	gene4	gene5	gene6	gene7	gene8	gene9	gene10	gene11	gene12	gene13	gene14	gene15	gene16	gene17	gene18	gene19	gene20
cond.1	4.795	4.737	4.343	9.348	10.209	10.209	10.209	10.209	10.209	10.209	10.209	4.343	9.348	10.209	10.209	10.209	10.209	10.209	10.209	10.209
cond.2	6.204	5.694	6.032	7.103	6.649	6.649	6.649	6.649	6.649	6.649	6.649	6.032	7.103	6.649	6.649	6.649	6.649	6.649	6.649	6.649
cond.3	8.693	8.488	8.503	10.365	6.086	6.086	6.086	6.086	6.086	6.086	6.086	8.503	10.365	6.086	6.086	6.086	6.086	6.086	6.086	6.086
cond.4	8.653	8.224	8.137	7.819	3.920	3.920	3.920	3.920	3.920	3.920	3.920	8.137	7.819	3.920	3.920	3.920	3.920	3.920	3.920	3.920
cond.5	6.920	6.658	7.102	8.045	11.013	11.013	11.013	11.013	11.013	11.013	11.013	7.102	8.045	11.013	11.013	11.013	11.013	11.013	11.013	11.013
cond.6	7.220	6.984	6.753	11.953	5.350	5.350	5.350	5.350	5.350	5.350	5.350	6.753	11.953	5.350	5.350	5.350	5.350	5.350	5.350	5.350
cond.7	7.996	9.139	9.335	10.703	6.111	6.111	6.111	6.111	6.111	6.111	6.111	9.335	10.703	6.111	6.111	6.111	6.111	6.111	6.111	6.111
cond.8	7.338	7.347	7.657	8.129	6.589	6.589	6.589	6.589	6.589	6.589	6.589	7.657	8.129	6.589	6.589	6.589	6.589	6.589	6.589	6.589
cond.9	8.570	8.518	8.631	6.234	5.324	5.324	5.324	5.324	5.324	5.324	5.324	8.631	6.234	5.324	5.324	5.324	5.324	5.324	5.324	5.324
cond.10	3.265	3.014	3.849	10.455	9.041	9.041	9.041	9.041	9.041	9.041	9.041	3.849	10.455	9.041	9.041	9.041	9.041	9.041	9.041	9.041
cond.11	5.671	4.924	5.745	5.339	13.020	13.020	13.020	13.020	13.020	13.020	13.020	5.745	5.339	13.020	13.020	13.020	13.020	13.020	13.020	13.020
cond.12	8.041	8.496	7.979	9.891	5.161	5.161	5.161	5.161	5.161	5.161	5.161	7.979	9.891	5.161	5.161	5.161	5.161	5.161	5.161	5.161
cond.13	9.008	8.273	8.417	7.436	4.487	4.487	4.487	4.487	4.487	4.487	4.487	8.417	7.436	4.487	4.487	4.487	4.487	4.487	4.487	4.487
cond.14	4.081	4.460	3.735	11.163	8.953	8.953	8.953	8.953	8.953	8.953	8.953	3.735	11.163	8.953	8.953	8.953	8.953	8.953	8.953	8.953
cond.15	10.274	10.304	10.061	10.702	6.180	6.180	6.180	6.180	6.180	6.180	6.180	10.061	10.702	6.180	6.180	6.180	6.180	6.180	6.180	6.180
cond.16	7.747	8.185	7.917	10.709	5.217	5.217	5.217	5.217	5.217	5.217	5.217	7.917	10.709	5.217	5.217	5.217	5.217	5.217	5.217	5.217
cond.17	8.720	8.374	8.735	10.939	8.401	8.401	8.401	8.401	8.401	8.401	8.401	8.735	10.939	8.401	8.401	8.401	8.401	8.401	8.401	8.401
cond.18	2.107	1.882	2.226	10.684	11.670	11.670	11.670	11.670	11.670	11.670	11.670	2.226	10.684	11.670	11.670	11.670	11.670	11.670	11.670	11.670
cond.19	7.464	8.043	8.729	5.482	5.371	5.371	5.371	5.371	5.371	5.371	5.371	8.729	5.482	5.371	5.371	5.371	5.371	5.371	5.371	5.371
cond.20	5.742	5.534	5.561	4.504	9.452	9.452	9.452	9.452	9.452	9.452	9.452	5.561	4.504	9.452	9.452	9.452	9.452	9.452	9.452	9.452
cond.21	10.274	10.304	10.061	10.702	6.180	6.180	6.180	6.180	6.180	6.180	6.180	10.061	10.702	6.180	6.180	6.180	6.180	6.180	6.180	6.180
cond.22	5.742	5.534	5.561	4.504	9.452	9.452	9.452	9.452	9.452	9.452	9.452	5.561	4.504	9.452	9.452	9.452	9.452	9.452	9.452	9.452
cond.23	6.204	5.694	6.032	7.103	6.649	6.649	6.649	6.649	6.649	6.649	6.649	6.032	7.103	6.649	6.649	6.649	6.649	6.649	6.649	6.649
cond.24	5.742	5.534	5.561	4.504	9.452	9.452	9.452	9.452	9.452	9.452	9.452	5.561	4.504	9.452	9.452	9.452	9.452	9.452	9.452	9.452
cond.25	5.742	5.534	5.561	4.504	9.452	9.452	9.452	9.452	9.452	9.452	9.452	5.561	4.504	9.452	9.452	9.452	9.452	9.452	9.452	9.452
cond.26	7.338	7.347	7.657	8.129	6.589	6.589	6.589	6.589	6.589	6.589	6.589	7.657	8.129	6.589	6.589	6.589	6.589	6.589	6.589	6.589
cond.27	5.742	5.534	5.561	4.504	9.452	9.452	9.452	9.452	9.452	9.452	9.452	5.561	4.504	9.452	9.452	9.452	9.452	9.452	9.452	9.452
cond.28	4.795	4.737	4.343	9.348	10.209	10.209	10.209	10.209	10.209	10.209	10.209	4.343	9.348	10.209	10.209	10.209	10.209	10.209	10.209	10.209

RegulonDB



Support Vector Machine
target gene / other genes

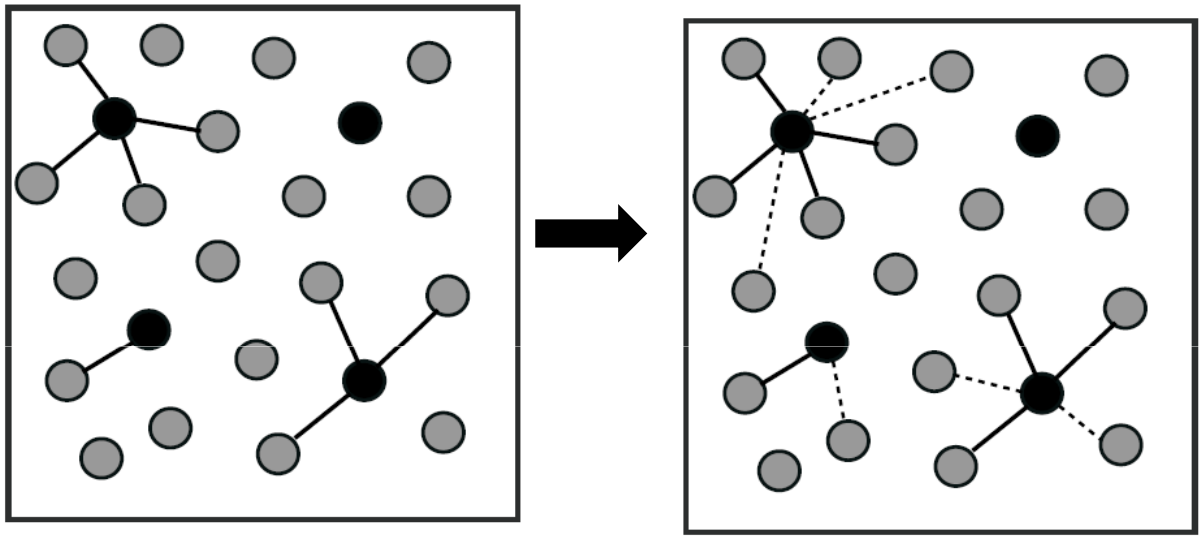


Mordelet, F. & Vert, J. (2008) SIRENE: supervised inference of regulatory networks. *Bioinformatics*, **24** (16), i76.

TNIFSED : Global Supervised Model

SIRENE : Local Supervised Classifier

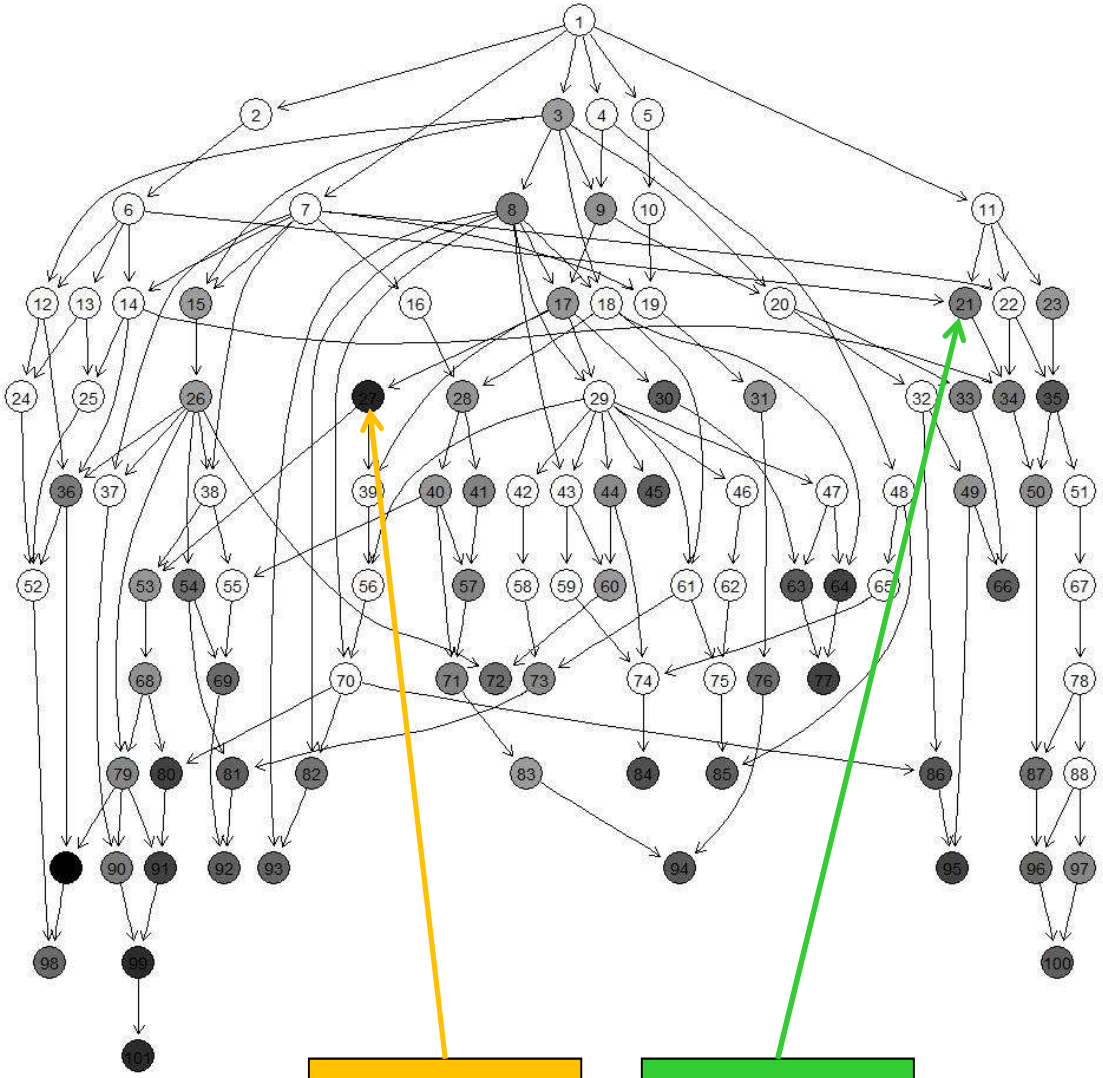
- Good predictive performances
- Unable to predict target genes of ‘orphan’ TF



TNIFSED : Global Supervised Classifier integrating :

- Correlation Coefficients (Expression Matrix)
- Partial Correlation Coefficients (Expression Matrix)
- Molecular function Similarity (Gene Ontology)
- Cellular Component Similarity (Gene Ontology)
- Biological Process Similarity (Gene Ontology)

Functional Similarity (cc, mf, bp)



1	biological_process
2	biological regulation
3	developmental process
4	multicellular organismal process
5	localization
6	regulation of biological process
7	cellular process
8	anatomical structure development
9	multicellular organismal development
10	localization of cell
11	metabolic process
12	regulation of developmental process
13	negative regulation of biological process
14	regulation of cellular process
15	cellular developmental process
16	cellular component organization and biogenesis
17	system development
18	anatomical structure morphogenesis
19	cell motility
20	pattern specification process
21	regulation of metabolic process
22	cellular metabolic process
23	primary metabolic process
24	negative regulation of developmental process
25	negative regulation of cellular process
26	cell differentiation
27	nervous system development
28	cellular structure morphogenesis
29	organ development
30	exocrine system development

Wang, J., Du, Z., Payattakool, R., Yu, P. & Chen, C. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23** (10), 1274.

TNIFSED : Global Supervised Model

Application of TNIFSED to the bacterium E coli :

- RegulonDB : 132 Transcription factor
- Expression Matrix : 3012 genes (3011 potential target genes)
- 132 * 3011 = 397.452 training examples (2407 positives; 395.045 negative)
- 5 – fold Cross validation

Training dataset

Transcription factor	Gene	Correlation	Partial Correlation	Similarity cc	Similarity bp	Similarity mf	Interaction
acrR	aas_	0.377	0.019	0.487	0.000	0.000	0
acrR	aat_	0.088	0.003	1.000	0.348	0.000	0
acrR	abrB	0.263	0.004	0.487	0.000	0.000	0
acrR	accA	0.515	0.005	1.000	0.000	0.000	0
acrR	accB	0.388	0.005	0.000	0.441	0.000	0
acrR	accC	0.346	0.007	1.000	0.000	0.000	0
acrR	accD	0.186	0.004	1.000	0.000	0.000	0
acrR	aceA	0.167	0.003	1.000	0.312	0.000	0
acrR	aceB	0.153	0.001	1.000	0.312	0.000	0
acrR	aceE	0.290	0.003	0.000	0.260	0.000	0
acrR	aceF	0.309	0.003	0.000	0.260	0.000	0
acrR	aceK	0.133	0.007	0.000	0.312	0.000	0
acrR	ackA	0.208	0.002	1.000	0.240	0.000	0
acrR	acnA	0.366	0.003	0.000	0.232	0.000	0
acrR	acnB	0.115	0.013	0.000	0.241	0.000	0
acrR	acpP	0.051	0.008	0.000	0.441	0.000	0
acrR	acpS	0.129	0.008	1.000	0.000	0.000	0
acrR	acrA	0.576	0.016	0.487	1.000	0.000	1
acrR	acrB	0.494	0.027	0.487	1.000	0.000	1
acrR	acrD	0.175	0.004	0.487	1.000	0.000	0

Training of the logistic regression model

$$\log\left[\frac{P(x_1, x_2, \cdots, x_p)}{1 - P(x_1, x_2, \cdots, x_p)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

Prediction of interaction

New dataset

Transcription factor	Gene	Correlation	Partial Correlation	Similarity cc	Similarity bp	Similarity mf
acrR	malT	0.384	0.005	1.000	0.626	0.000
acrR	malX	0.140	0.003	0.000	0.280	0.000
acrR	malY	0.150	0.004	1.000	0.280	0.000
acrR	malZ	0.119	0.004	1.000	0.268	0.000
acrR	manA	0.154	0.004	1.000	0.513	0.000
acrR	manX	0.108	0.013	1.000	0.280	0.000
acrR	manY	0.105	0.011	0.487	0.280	0.000
acrR	manZ	0.052	0.022	0.487	0.280	0.000
acrR	maoC	0.340	0.014	0.000	0.280	0.000
acrR	map_	0.389	0.004	0.000	0.491	0.000
acrR	marA	0.092	0.003	1.000	0.880	0.000

Interaction
0.0012
0.0003
0.0003
0.0071
0.0060
0.0003
0.0051
0.0003
0.0017
0.0043
0.0003

Construction of gene regulatory network from Microarray data

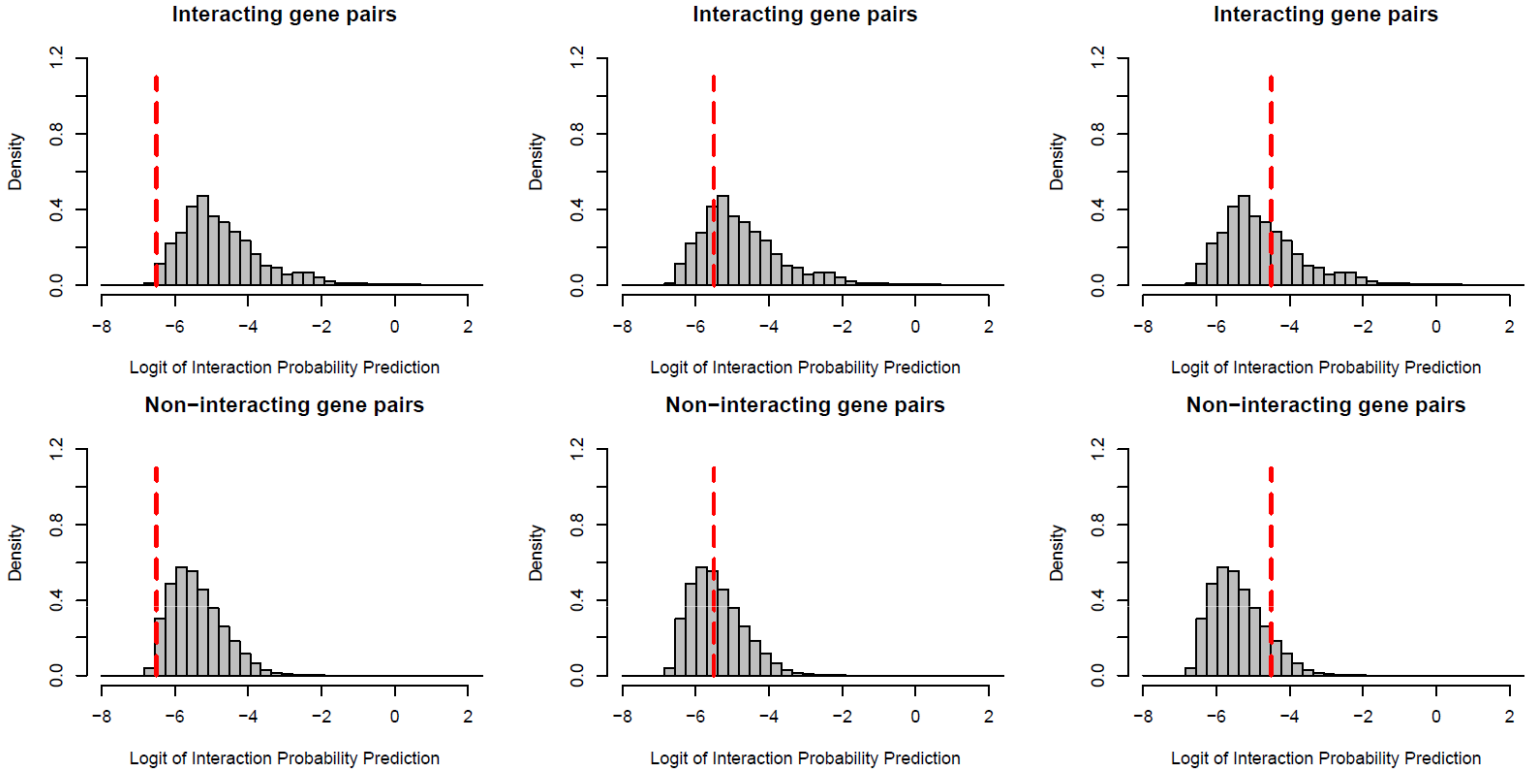
- Biological introduction
  - DNA, RNA, Protein
  - Transcription Factors and Gene Regulatory network
  - Measuring RNA concentration: The Microarray technology

- Gene network reconstruction methods from microarray data
  - Relevance Network (Correlation)
  - Gaussian Graphicals Models (Partial Correlation)
  - SIRENE (Supervised approach: local models)
  - TNIFSED (Supervised approach: global model)

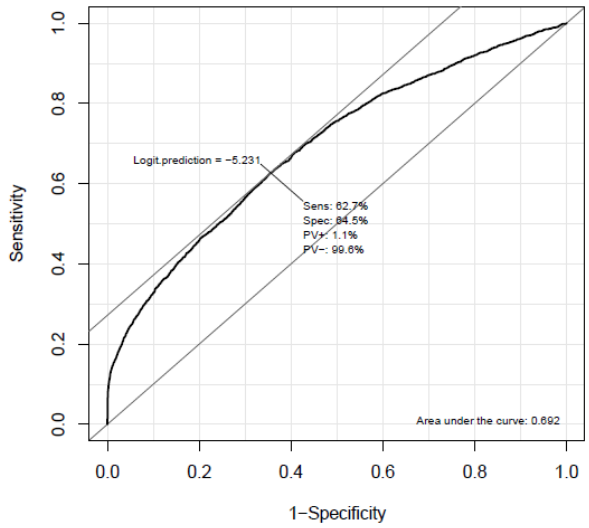
- Evaluation of predictive peformances
  - Area Under the Curve of the Receiver Operating Characteristic (AUC)

- Results and conclusions

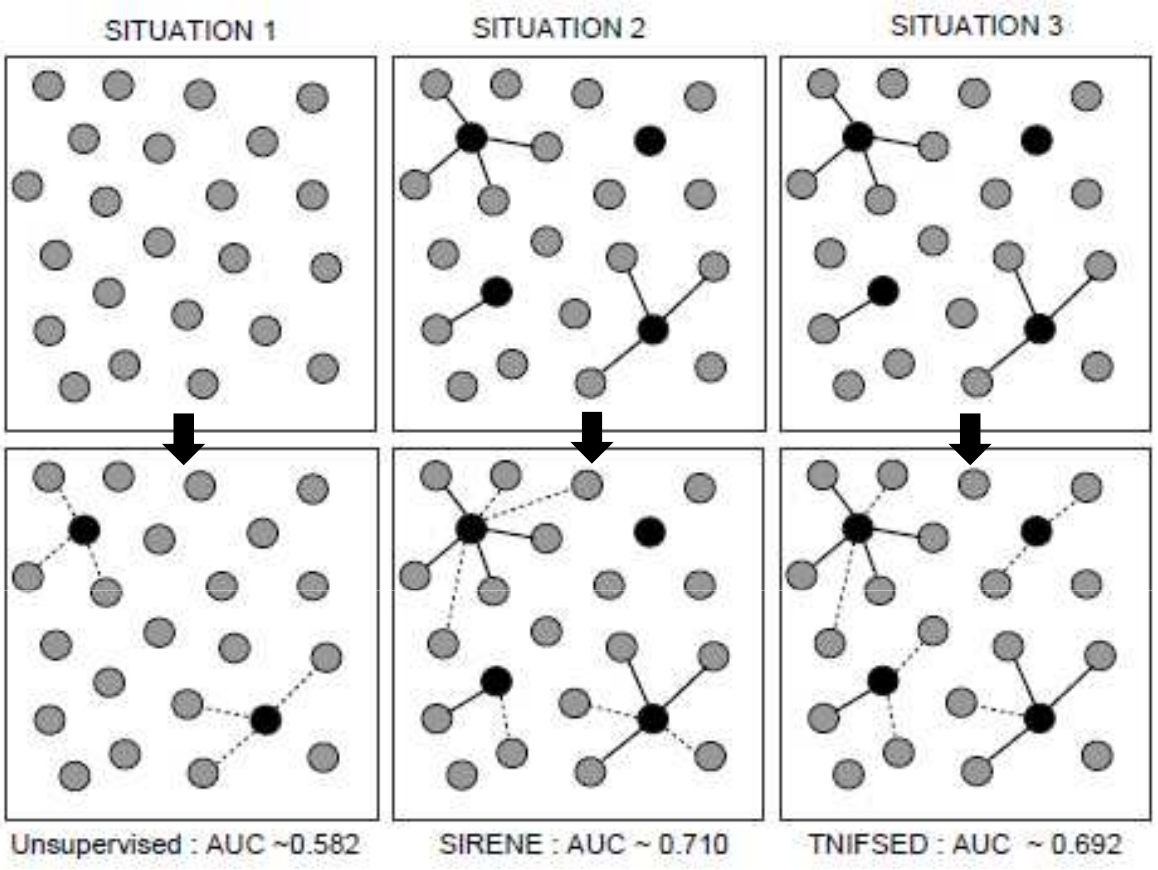
Evaluation of predictive performance: AUC of the ROC



		Interaction (as determined by RegulonDB)	
		Positive	Negative
Test outcome	Positive	TRUE POSITIVE	FALSE POSITIVE
	Negative	FALSE NEGATIVE	TRUE NEGATIVE
		↓	↓
		Sensitivity	Specificity



Results and conclusions



Situation 1: Organism with no or very litle known interactions => unsupervised method (GGM)

Situation 2: Organism with some known interactions => SIRENE (unable for ‘orphans’ TF)

Situation 3: Organism with some known interactions => TNIFSED (OK for ‘orphans TF)

Questions ?