Random forest for network inférence (in biology)

Nathalie Vialaneix

nathalie.vialaneix@inrae.fr

http://www.nathalievialaneix.eu

ECAS
European Courses in Advanced Statistics

SFdS

ECAS-SFdS course on random forest
Fréjus (France), October 8-13, 2023

RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

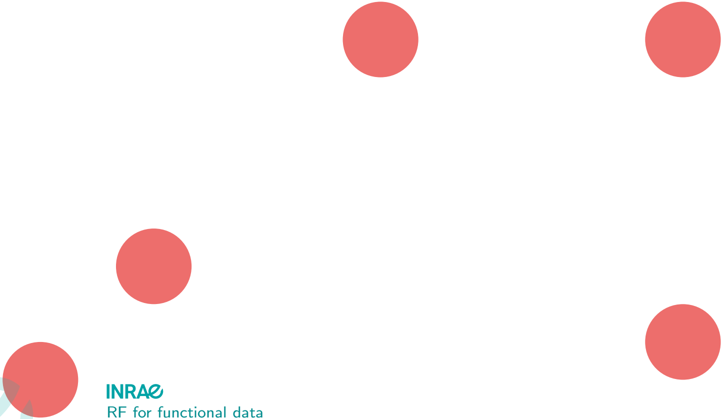
What is a network/graph?

Mathematical object used to model relational data between entities.

What is a network/graph?

Mathematical object used to model relational data between entities.

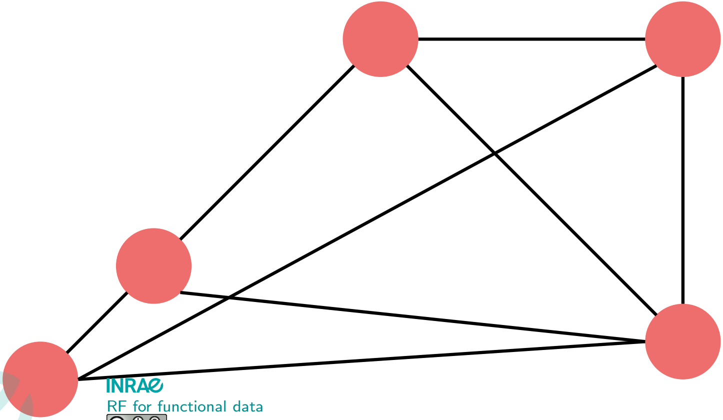
The entities are called the nodes or the vertices



What is a network/graph?

Mathematical object used to model relational data between entities.

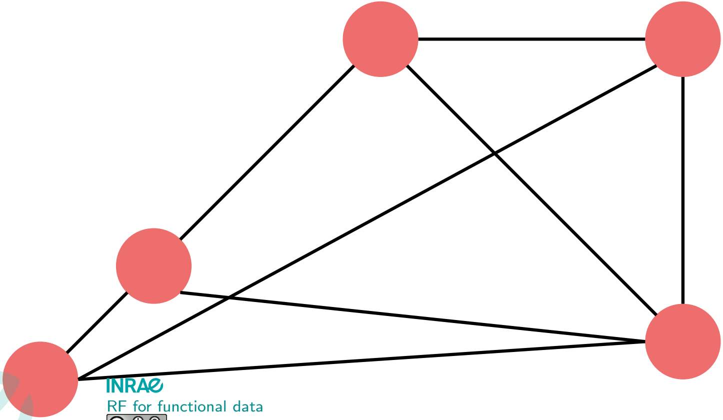
A relation between two entities is modeled by an edge



What is a network/graph?

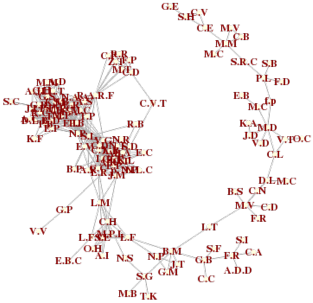
Mathematical object used to model relational data between entities.

A relation between two entities is modeled by an edge + edges can even be oriented



(non biological) Examples

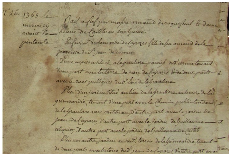
Social network: nodes: persons - edges: 2 persons are connected (“friends”)



(Natty’s facebook[1] network)

(non biological) Examples

Modeling a large corpus of medieval documents



N°26. 1365. Le mercredy avant la pentecote

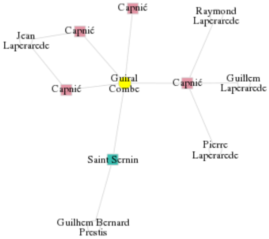
Bail a fief par messire arnaud de roquefeuil et dame
helene de Castelnau son epouse
En faveur de bernarde de Cayrazes fille de feu arnaud de la
paroisse de St jean de cornus
D'une maison située a la grauliere paroisse dud. cornus tenant
d'une part avec la terre de jean de Cayrazes et de deux parts
avec les rues publiques dud. lieu de la grauliere
Plus d'un jardin situé au lieu de la grauliere au terroir de la
grimoardie, tenant d'une part avec le chemin public tendant
de la grauliere vers castelnau, d'autre part avec le jardin de
jean de Cayrazes, d'autre part avec le jardin de Guillaume
aliquier, d'autre part avec le jardin de Guillaume de Castel
Plus un autre jardin audit terroir de la grimoardie, tenant
de deux parts avec la terre dud. jean de Cayrazes, d'autre part avec

Notarial acts (mostly "baux à fief", more precisely, land charters) established in a "seigneurie" named "Castelnau Montratier", written between 1250 and 1500, involving tenants and lords.[a]

[a] http://graphcomp.univ-tlse2.fr

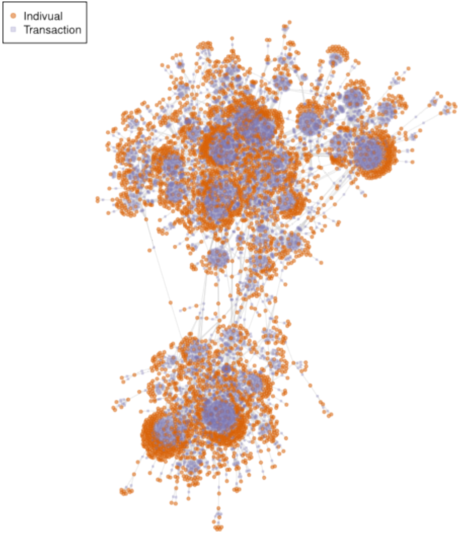
(non biological) Examples

Modeling a large corpus of medieval documents



- nodes: transactions and individuals (3 918 nodes)
- edges: an individual is directly involved in a transaction (6 455 edges)

(non biological) Examples



Standard issues associated with networks

Inference

Given data, how to build a graph whose edges represent the “dependency relationship” between variables?

Standard issues associated with networks

Inference

Given data, how to build a graph whose edges represent the “dependency relationship” between variables?

Graph mining (examples)

1. Network visualization
2. Network clustering

Standard issues associated with networks

Inference

Given data, how to build a graph whose edges represent the “dependency relationship” between variables? Random forest is useful here!

Graph mining (examples)

1. Network visualization
2. Network clustering

Network inference in biology: an overview

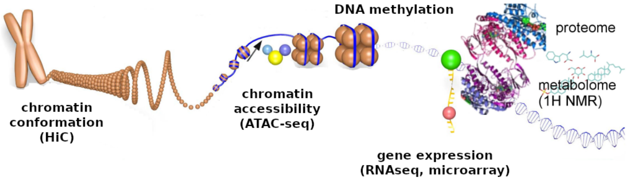
From GGM to random forest

Variants of network inference with random forest

More on tree ensemble methods

Cell molecular mechanisms: gene transcription/translation

Disclaimer: This is way more complicated than what I will tell...!

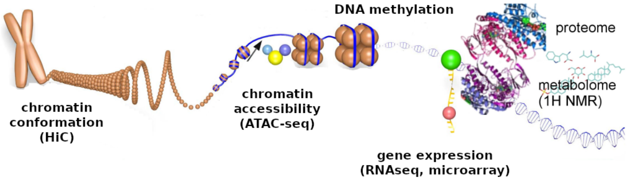


annotation
(genes, exon/intron, ...)

... etc ... etc ...

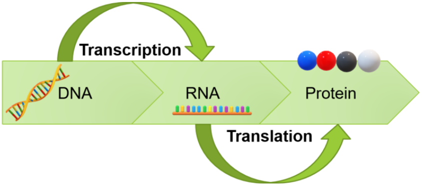
Cell molecular mechanisms: gene transcription/translation

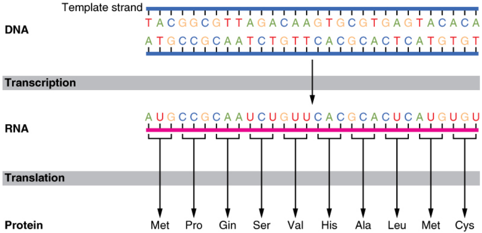
Disclaimer: This is way more complicated than what I will tell...!


chromatin conformation (HiC)
chromatin accessibility (ATAC-seq)
DNA methylation
proteome
metabolome (1H NMR)
gene expression (RNAseq, microarray)

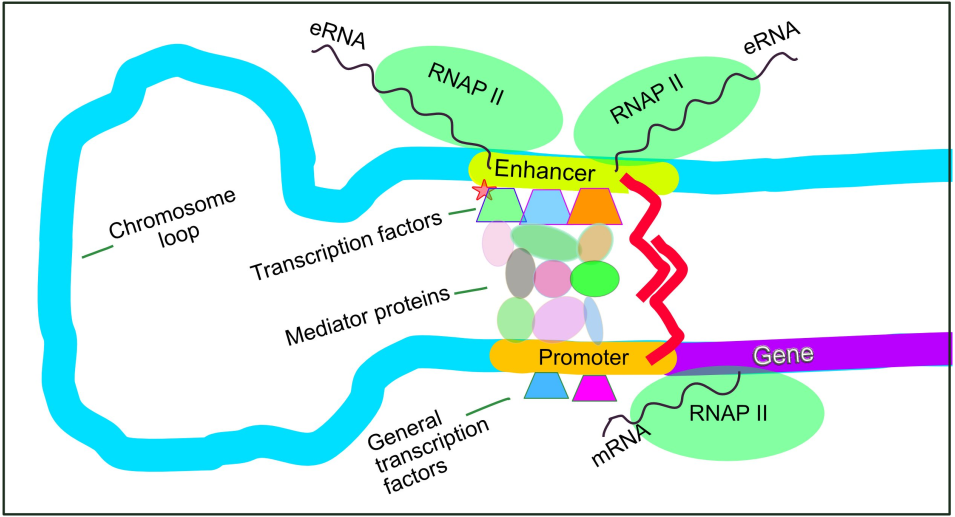
annotation
(genes, exon/intron, ...)

... etc ... etc ...




Template strand
DNA
T A C G G C G T T A G A C A A G T G C G T G A G T A C A C A
A T G C C G C A A T C T G T T C A C G C A C T C A T G T G T
Transcription
RNA
A U G C C G C A A U C U G U U C A C G C A C U C A U G U G U
Translation
Protein
Met
Pro
Gin
Ser
Val
His
Ala
Leu
Met
Cys

Cell molecular mechanisms: (gene-)gene regulation



In short: gene networks

What we would like: use data on gene expression to obtain a network with:

- nodes = genes
- edge = a regulation process of one gene on the other gene

In short: gene networks

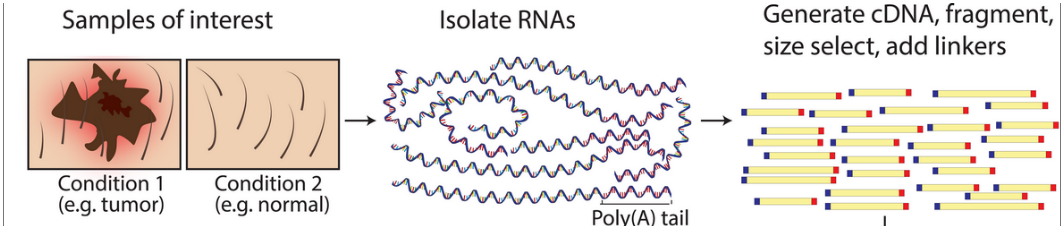
What we would like: use data on gene expression to obtain a network with:

- nodes = genes
- edge = a regulation process of one gene on the other gene

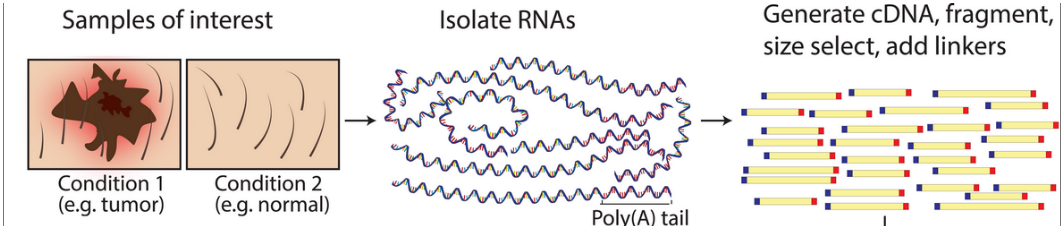
What we approximately actually obtain:

- nodes = genes
- edge = the fact that two genes show similar patterns of expression

Collecting data: gene expression



Collecting data: gene expression



Various techniques:

- continuous data: RT-qPCR, various arrays
- count data: RNA-seq (and its single-cell variant)

Back to a more formal (less biology) description

Data: large scale gene expression data

$$\begin{array}{c}\text{individuals}\\ n \simeq 30/50\end{array} \left\{ \underbrace{X = \begin{pmatrix} . & . & . & . & . & . \\ . & . & X_i^j & . & . & . \\ . & . & . & . & . & . \end{pmatrix}}_{\text{variables (genes expression)},\ p\simeq 10^{3/4}} \right.$$

Back to a more formal (less biology) description

Data: large scale gene expression data

$$\begin{array}{l}\text{individuals}\\ n \simeq 30/50\end{array} \left\{ \underbrace{X = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & X_i^j & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}}_{\text{variables (genes expression)},\; p \simeq 10^{3/4}} \right.$$

What we want to obtain: a network with

- nodes: genes
- edges: some kind of dependency between genes (ideally regulation)

Back to a more formal (less biology) description

Data: large scale gene expression data

$$\begin{array}{c} \text{individuals} \\ n \simeq 30/50 \end{array} \left\{ \underbrace{X = \begin{pmatrix} . & . & . & . & . & . \\ . & . & X_i^j & . & . & . \\ . & . & . & . & . & . \end{pmatrix}}_{\text{variables (genes expression)},\ p \simeq 10^{3/4}} \right.$$

What we want to obtain: a network with

- nodes: genes
- edges: some kind of dependency between genes (ideally regulation)

Note: This is hard to perform genome-widely: Humans $\geq$ 20,000 genes coding for proteins (plus the others), *Bacillus subtilis* $\sim$ 4,000 genes

Main methods used for network inference

- Relevance network: correlation, mutual information
- Partial correlation (Gaussian Graphical Model framework)
- Bayesian network
- Other regression based methods, including:
  - random forest: best in [**Marbach et al., 2012**] / DREAM4 challenge!
  - (of course) deep learning
  - ...

Network inference in biology: an overview

From GGM to random forest

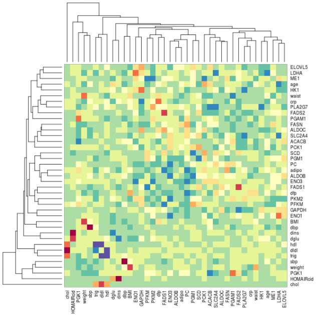
Variants of network inference with random forest

More on tree ensemble methods

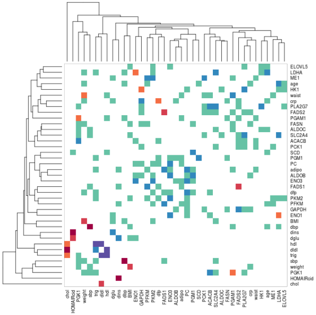
Using *correlations*: relevance network

[Butte and Kohane, 1999, Butte and Kohane, 2000]

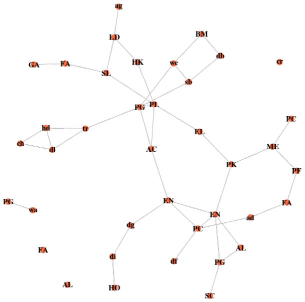
First (naive) approach: correlations + threshold



"Correlations"

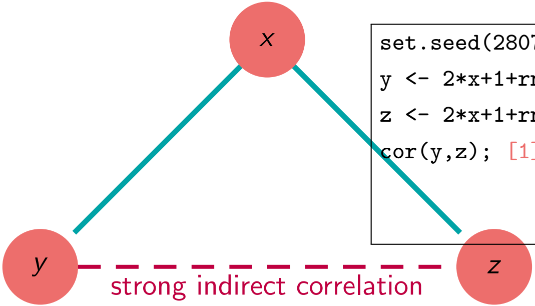


Thresholding



Graph

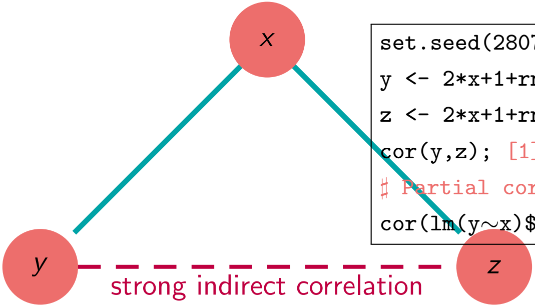
But correlation is not causality...



strong indirect correlation

```
set.seed(2807); x <- runif(100)
y <- 2*x+1+rnorm(100,0,0.1); cor(x,y); [1] 0.9988261
z <- 2*x+1+rnorm(100,0,0.1); cor(x,z); [1] 0.998751
cor(y,z); [1] 0.9971105
```

But correlation is not causality...



```
set.seed(2807); x <- runif(100)
y <- 2*x+1+rnorm(100,0,0.1); cor(x,y); [1] 0.9988261
z <- 2*x+1+rnorm(100,0,0.1); cor(x,z); [1] 0.998751
cor(y,z); [1] 0.9971105
♯ Partial correlation
cor(lm(y~x)$residuals, lm(z~x)$residuals) [1] -0.1933699
```

Partial correlation is also...

For: $(X_i)_{i=1,\ldots,n}$ i.i.d. $\mathcal{N}(0,\Sigma)$ (gene expressions)

- $\mathbb{C}\mathrm{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right)$

Partial correlation is also...

For: $(X_i)_{i=1,\ldots,n}$ i.i.d. $\mathcal{N}(0,\Sigma)$ (gene expressions)

- $\mathbb{C}\mathrm{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right)$
- Related to the entries of Σ^{-1}

Partial correlation is also...

For: $(X_i)_{i=1,\ldots,n}$ i.i.d. $\mathcal{N}(0, \Sigma)$ (gene expressions)

- $\mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right)$
- Related to the entries of Σ^{-1}
- Related to $\beta_{jj'}$ in linear regression models:

$$X^j = \sum_{j' \neq j} \beta_{jj'} X^{j'} + \epsilon_j$$

GGM

For: $(X_i)_{i=1,\ldots,n}$ i.i.d. $\mathcal{N}(0,\Sigma)$ (gene expressions)

- edge between j and j' $\Leftrightarrow \mathbb{C}\mathrm{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) \neq 0$

- Related to the entries of Σ^{-1}

- Related to $\beta_{jj'}$ in linear regression models:

$$X^j = \sum_{j'\neq j} \beta_{jj'} X^{j'} + \epsilon_j$$

GGM

For: $(X_i)_{i=1,\ldots,n}$ i.i.d. $\mathcal{N}(0,\Sigma)$ (gene expressions)

- edge between j and j' $\Leftrightarrow$ $\mathbb{C}\mathrm{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) \neq 0$
- edge between j and j' $\Leftrightarrow$ $\left[\Sigma^{-1}\right]_{jj'} \neq 0$ **[Friedman et al., 2008]**
- Related to $\beta_{jj'}$ in linear regression models:

$$X^j = \sum_{j'\neq j} \beta_{jj'} X^{j'} + \epsilon_j$$

GGM

For: $(X_i)_{i=1,\ldots,n}$ i.i.d. $\mathcal{N}(0,\Sigma)$ (gene expressions)

- edge between j and j' $\Leftrightarrow \mathbb{C}\mathrm{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) \neq 0$

- edge between j and j' $\Leftrightarrow \left[\Sigma^{-1}\right]_{jj'} \neq 0$ **[Friedman et al., 2008]**

- edge between j and j' $\Leftrightarrow \beta_{jj'} \neq 0$ in

$$X^j = \sum_{j'\neq j} \beta_{jj'} X^{j'} + \epsilon_j$$

[Meinshausen and Bühlmann, 2006]

Why restrict yourself at linear regression?

- GGM: Gaussian assumption + fit of p linear regressions

$$\forall\, j = 1, \ldots, p, \qquad X^j = \sum_{j' \neq j} \beta_{jj'} X^{j'} + \epsilon_j$$

Why restrict yourself at linear regression?

- GGM: Gaussian assumption + fit of p linear regressions

$$\forall j = 1, \ldots, p, \qquad X^j = \sum_{j' \neq j} \beta_{jj'} X^{j'} + \epsilon_j$$

Problems: ill-conditionned, only linear dependencies, restricted to Gaussian case.

- Just fit p regressions!

$$\forall j = 1, \ldots, p, \qquad X^j = \sum_{j' \neq j} F_j(X^{j'}) + \epsilon_j$$

F_j: your favorite regression method!

Why restrict yourself at linear regression?

- GGM: Gaussian assumption + fit of p linear regressions

$$\forall\, j = 1, \ldots, p, \qquad X^j = \sum_{j' \neq j} \beta_{jj'} X^{j'} + \epsilon_j$$

Problems: ill-conditionned, only linear dependencies, restricted to Gaussian case.

- Just fit p regressions!

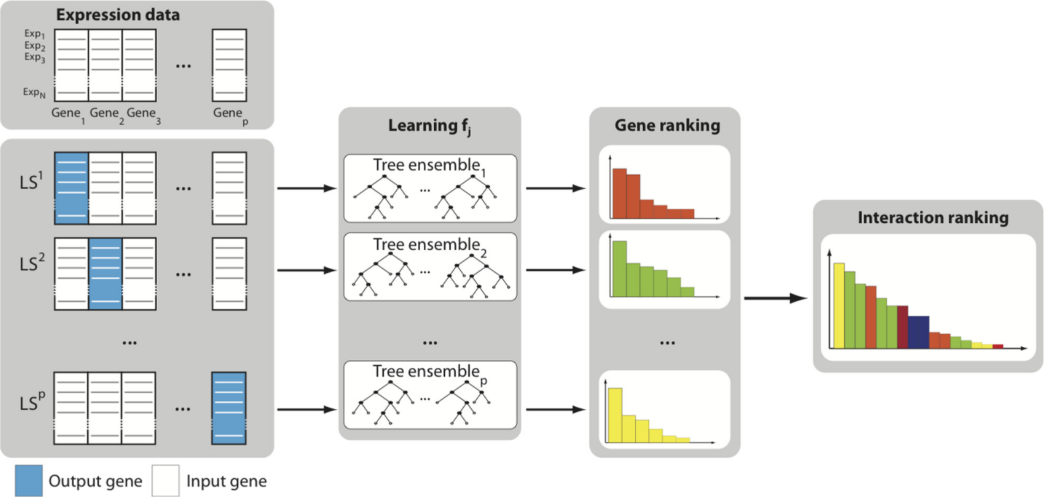
$$\forall\, j = 1, \ldots, p, \qquad X^j = \sum_{j' \neq j} F_j(X^{j'}) + \epsilon_j$$

F_j: your favorite regression method!
But: Direct dependency interpretation is lost.

My favorite regression method? Random forest!

[Huynh-Thu et al., 2010] **GENIE3**

GENIE3: Using feature selection in RF to predict edges



Important notes: orientation

Notation: $w_{jj'}$ weight obtained by $X^{j'}$ to predict X^j

Important notes: orientation

Notation: $w_{jj'}$ weight obtained by $X^{j'}$ to predict X^j

- in general $w_{jj'} \neq w_{j'j}$ which gives a way to obtain oriented edges (not really causality though)

Important notes: weight definition

Notation: $w_{jj'}$ weight obtained by $X^{j'}$ to predict X^j

Important notes: weight definition

Notation: $w_{jj'}$ weight obtained by $X^{j'}$ to predict X^j

- in RF$_j$, $w_{jj'} :=$ Gini index of variable j', *e.g.*, reduction of variance due to splits defined by variable j': Pierre's Tuesday class

$$\sum_{\mathcal{N} \text{ defined by } j'} \left[|\mathcal{N}| \mathrm{Var}\left(X^j_{\mathcal{N}}\right) - |\mathcal{N}_R| \mathrm{Var}\left(X^j_{\mathcal{N}_R}\right) - |\mathcal{N}_L| \mathrm{Var}\left(X^j_{\mathcal{N}_L}\right) \right]$$

$\mathcal{N}$ improperly defines either the node, the split, and the samples assigned to the node.

Important notes: weight definition

Notation: $w_{jj'}$ weight obtained by $X^{j'}$ to predict X^j

- in RF$_j$, $w_{jj'} :=$ Gini index of variable j', *e.g.*, reduction of variance due to splits defined by variable j': Pierre's Tuesday class

$$\sum_{\mathcal{N} \text{ defined by } j'} \left[|\mathcal{N}| \mathrm{Var}\left(X^j_{\mathcal{N}}\right) - |\mathcal{N}_R| \mathrm{Var}\left(X^j_{\mathcal{N}_R}\right) - |\mathcal{N}_L| \mathrm{Var}\left(X^j_{\mathcal{N}_L}\right) \right]$$

$\mathcal{N}$ improperly defines either the node, the split, and the samples assigned to the node.

Advantages:

- fast to compute (compared to MDA obtained by permutation)
- RF can be replaced by Extra-Trees ensemble Pierre's Thursday class

Important notes: weight definition

Notation: $w_{jj'}$ weight obtained by $X^{j'}$ to predict X^j

- in RF$_j$, $w_{jj'} :=$ Gini index of variable j', *e.g.*, reduction of variance due to splits defined by variable j': Pierre's Tuesday class

$$\sum_{\mathcal{N}\text{ defined by }j'} \left[|\mathcal{N}|\mathrm{Var}\left(X^j_{\mathcal{N}}\right) - |\mathcal{N}_R|\mathrm{Var}\left(X^j_{\mathcal{N}_R}\right) - |\mathcal{N}_L|\mathrm{Var}\left(X^j_{\mathcal{N}_L}\right)\right]$$

$\mathcal{N}$ improperly defines either the node, the split, and the samples assigned to the node.

Advantages:

- fast to compute (compared to MDA obtained by permutation)
- RF can be replaced by Extra-Trees ensemble Pierre's Thursday class

Drawback:

- migth be slightly less efficient than MDA

Important notes: ranking

Notation: $w_{jj'}$ weight obtained by $X^{j'}$ to predict X^j

Important notes: ranking

Notation: $w_{jj'}$ weight obtained by $X^{j'}$ to predict X^j

- Can we really rank $(w_{jj'})_{j,j':j\neq j'}$ globally? For a given forest RF$_j$,

$$\sum_{j'} w_{jj'} \qquad \sim \qquad n \times \mathrm{Var}\left(X^j\right)$$

Important notes: ranking

Notation: $w_{jj'}$ weight obtained by $X^{j'}$ to predict X^j

- Can we really rank $(w_{jj'})_{j,j':j\neq j'}$ globally? For a given forest RF_j,

$$\sum_{j'} w_{jj'} \qquad \sim \qquad n \times \mathrm{Var}\left(X^j\right)$$

$\Rightarrow$ sound preprocesssing: reduction of all variables to unit variance

Important notes: ranking

Notation: $w_{jj'}$ weight obtained by $X^{j'}$ to predict X^j

- Can we really rank $(w_{jj'})_{j,j' : j \neq j'}$ globally? For a given forest RF_j,

$$\sum_{j'} w_{jj'} \qquad \sim \qquad n \times \mathrm{Var}\left(X^j\right)$$

$\Rightarrow$ sound preprocesssing: reduction of all variables to unit variance

- output: ranking of the edges based on $(w_{jj'})_{j,j' : j \neq j'}$ $\Rightarrow$ edges require a threshold

Experiments on *Escherichia coli*

- expression data: $n = 907$, $p = 4297$ (microarray)
- “ground truth” network: from RegulonDB (curated but might not be exaustive; 1471 genes only)

Experiments on *Escherichia coli*

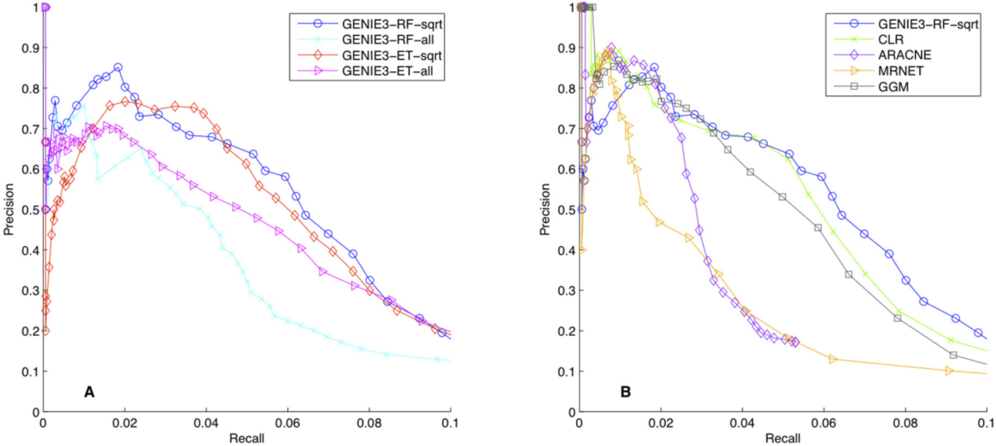
- expression data: $n = 907$, $p = 4297$ (microarray)
- "ground truth" network: from RegulonDB (curated but might not be exaustive; 1471 genes only)

Hyper-parameters:

- # trees: $1,000$
- $m = \sqrt{p-1}$ or $p - 1$ (full)
- RF or ET
- no decision on edges (PR and ROC curves)
- sets of predictors restricted to known regulators

  So: ranking of $(w_{jj'})_{j=1,\ldots,p,\, j':\ \mathrm{reg.}}$ only.

Results


A
Precision
Recall
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.02
0.04
0.06
0.08
0.1
GENIE3−RF−sqrt
GENIE3−RF−all
GENIE3−ET−sqrt
GENIE3−ET−all
B
GENIE3−RF−sqrt
CLR
ARACNE
MRNET
GGM

Network inference in biology: an overview

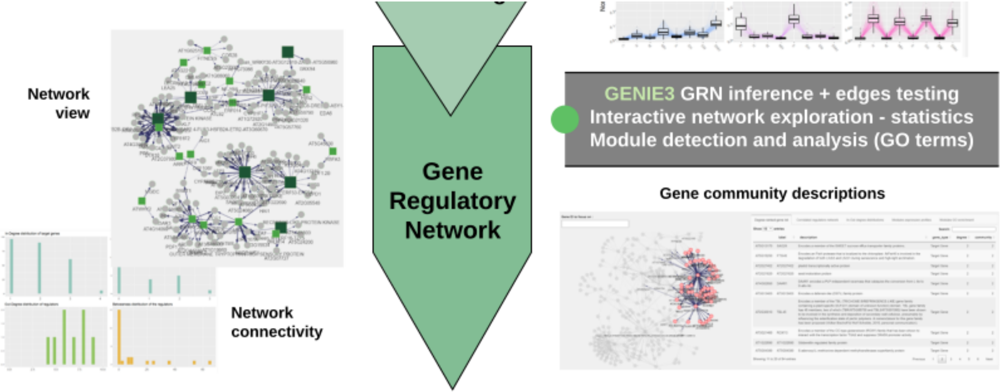
From GGM to random forest

Variants of network inference with random forest

More on tree ensemble methods

DIANE: How to select edges? [Cassan et al., 2021]

https://diane.bpmp.inrae.fr/



Network view

Network connectivity

Gene Regulatory Network

GENIE3 GRN inference + edges testing
Interactive network exploration - statistics
Module detection and analysis (GO terms)

Gene community descriptions

What’s new in **DIANE**?

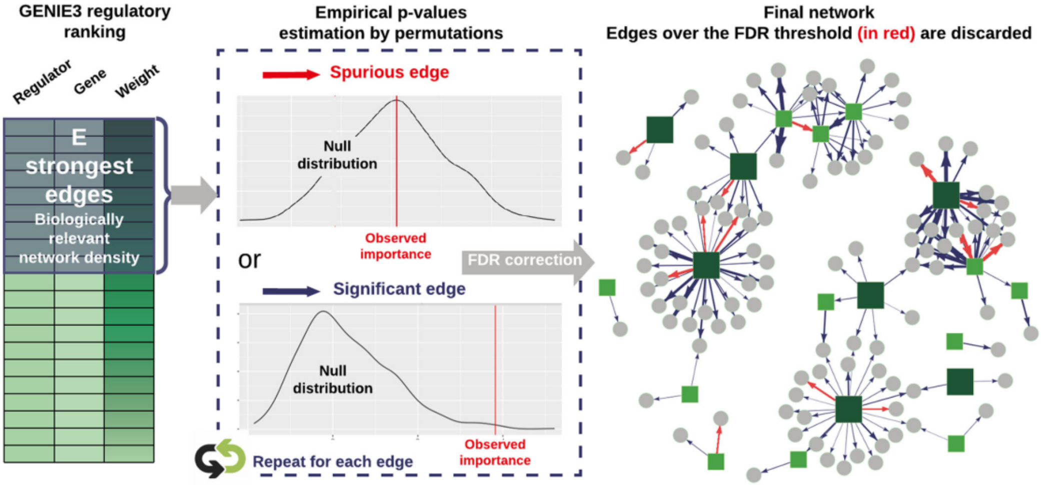
1. predictors are only TF (transcription factors) standard pre-filtering

2. transcription factors (highly correlated) can be grouped into a single gene

3. edges pre-selected using **GENIE3** (threshold based on plausible global density)

4. empirical *p*-value computation based on MDA for final selection with **rfpermute** (using MDA)

What’s new in **DIANE**?



More on edge selection [Aibar et al., 2017]

SCENIC (oriented toward single-cell) / **GENIE3** component selects edges using:

1. weight > 0.001
2. further filters for multiple gene sets (a gene set $=$ a cluster of genes with a TF):
   - top predicted genes for each TF
   - top predictor TF for each gene
   - several weight thresholds
3. further filtering (using biological information on DNA motifs with **RcisTarget**)
   not described here

More on tree ensemble methods [Aibar et al., 2017]

Alternative to **GENIE3** in **SCENIC**: **GRNBoost**

https://github.com/aertslab/GRNBoost

Replace RF method with XGBoost:

- tree ensemble based on boosting   Pierre's Thursday class
- tree depth restricted to 1

More on tree ensemble methods [Aibar et al., 2017]

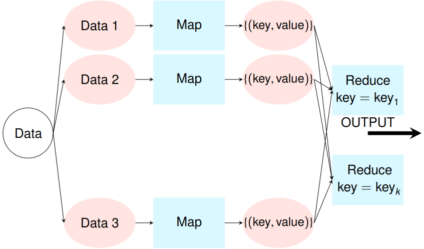
Alternative to **GENIE3** in **SCENIC**: **GRNBoost**

https://github.com/aertslab/GRNBoost

Replace RF method with XGBoost:

- tree ensemble based on boosting   Pierre's Thursday class
- tree depth restricted to 1

Map/Reduce implementation (for spark):


Data
Data 1
Data 2
Data 3
Map
Map
Map
$\{(\text{key}, \text{value})\}$
$\{(\text{key}, \text{value})\}$
$\{(\text{key}, \text{value})\}$
Reduce
$\text{key} = \text{key}_1$
Reduce
$\text{key} = \text{key}_k$
OUTPUT

- map split based on targets
- map output: set of edges (same filters) (not 100% sure)
- reduce: union of output edges

Network inference in biology: an overview

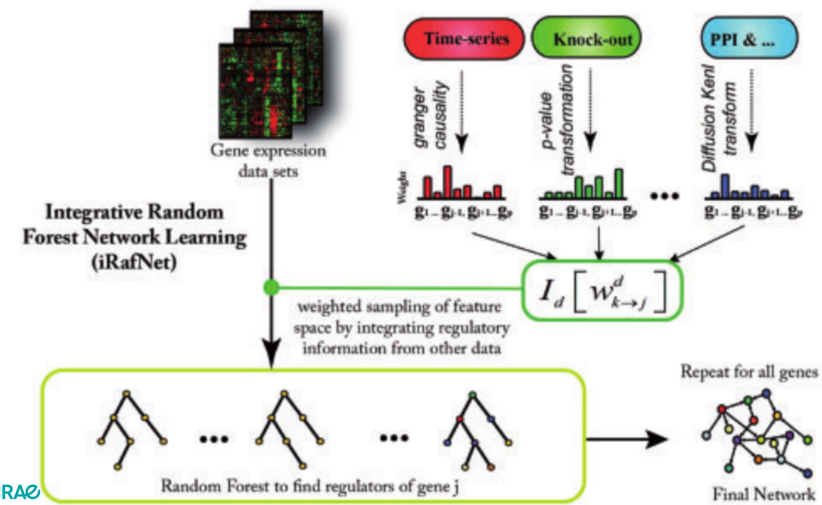
From GGM to random forest

Variants of network inference with random forest

More on tree ensemble methods

Integration of knowledge into Random Forest [Petralia et al., 2015]

IRafNet


Gene expression data sets
Time-series
Knock-out
PPI & ...
granger causality
p-value transformation
Diffusion Kenl transform
Weight
$g_1 .. g_{j-1}, g_{j+1} .. g_p$
...
Integrative Random Forest Network Learning (iRafNet)
$I_d\left[w^d_{k\to j}\right]$
weighted sampling of feature space by integrating regulatory information from other data
Repeat for all genes
Random Forest to find regulators of gene j
Final Network

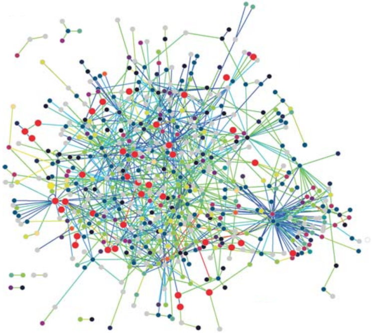
Using prior knowledge as a weight

1. Knowledge (given): modelled by $(w_{jj'}^{\text{prior}})_{jj'}$

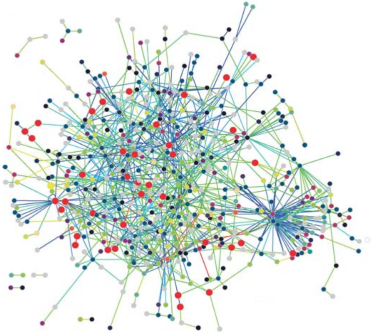
2. in RF_j, change the split rule definition:

   - sample $N \sim \mathcal{U}(\llbracket 1, p \rrbracket)$
   - sample N possible predictors with probability $(w_{jj'}^{\text{prior}})_{j'}$
   - find the best split among them

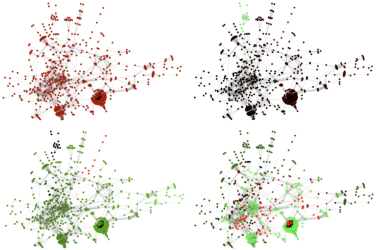
Example of prior weights: Protein-Protein Interactions (PPI)



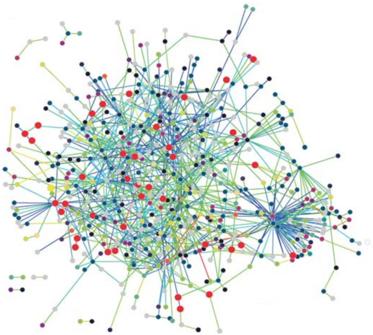
Example of prior weights: Protein-Protein Interactions (PPI)



1. from PPI network, Laplacian $L = D - P^{\text{ppi}}$ with $P^{\text{ppi}}_{jj'} \in \{0, 1\}$
   Why? L eigendecomposition $\sim$ graph structure
   [Rapaport et al., 2007].

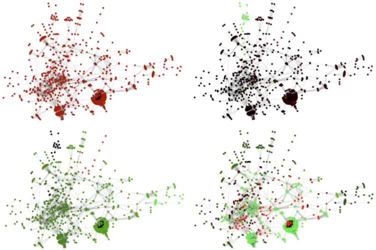


Example of prior weights: Protein-Protein Interactions (PPI)



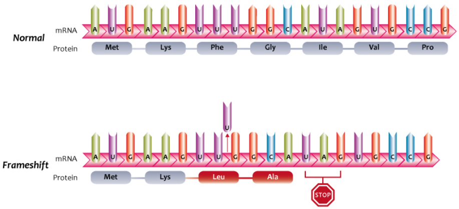
1. from PPI network, Laplacian $L = D - P^{\mathrm{ppi}}$ with $P^{\mathrm{ppi}}_{jj'} \in \{0,1\}$

   Why? L eigendecomposition $\sim$ graph structure [Rapaport et al., 2007].



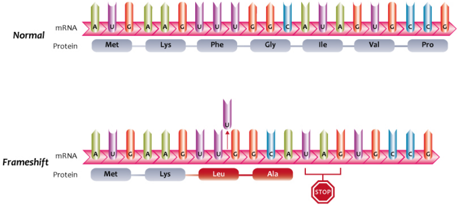
2. $W^{\mathrm{ppi}} = e^{-L}$ (heat kernel [Kondor and Lafferty, 2002])

Example of prior weights: knockouts


Normal
mRNA
A U G A A G U U U G G C A U A G U G C C G
Protein
Met
Lys
Phe
Gly
Ile
Val
Pro
U
Frameshift
mRNA
A U G A A G U U U G G C A U A G U G C C G
Protein
Met
Lys
Leu
Ala
STOP

Adapted from Campbell NA (ed). Biology, 2nd ed, 1990.

Example of prior weights: knockouts

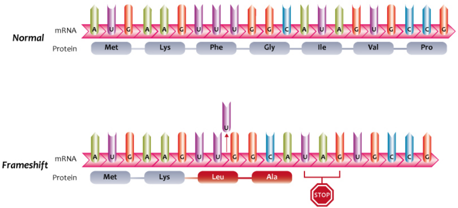

Normal
mRNA
A U G A A G U U U G G C A U A G U G C C G
Protein
Met
Lys
Phe
Gly
Ile
Val
Pro
Frameshift
mRNA
U
A U G A A G U U U G G C A U A G U G C C G
Protein
Met
Lys
Leu
Ala
STOP

Adapted from Campbell NA (ed). Biology, 2nd ed, 1990.

1. $\mathcal{K} \subset [\![1, p]\!]$ knockouts

2. for $j \in \mathcal{K}$ and $j' \in [\![1, p]\!]$, "j affects j'" if expression of j' is significantly different (Student's test) before/after knockout $w_{jj'}^{\mathrm{KO}} := p$-value

Example of prior weights: knockouts



Adapted from Campbell NA (ed). Biology, 2nd ed, 1990.

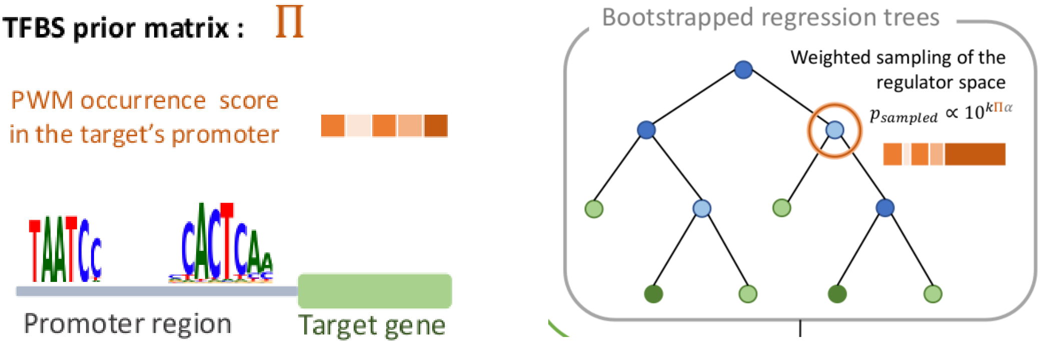
1. $\mathcal{K} \subset [\![1, p]\!]$ knockouts

2. for $j \in \mathcal{K}$ and $j' \in [\![1, p]\!]$, "j affects j'" if expression of j' is significantly different (Student's test) before/after knockout $w_{jj'}^{\mathrm{KO}} := p$-value

3. weights for $j \notin \mathcal{K}$: weigthed average $(w_{\ell j'}^{\mathrm{KO}})_{\ell \in \mathcal{K}}$ using similarity of gene sets that affect j and ℓ

Alternative ways to use priors (and alternative priors)

- using TFBS prior [Cassan et al., 2023]


TFBS prior matrix : Π
PWM occurrence score in the target's promoter
TAATCC
CACTCAA
Promoter region
Target gene
Bootstrapped regression trees
Weighted sampling of the regulator space
$p_{sampled} \propto 10^{k\Pi\alpha}$

Alternative ways to use priors (and alternative priors)

- using TFBS prior [**Cassan et al., 2023**]

- using chromatine accessibility (ATAC-seq) **SCENIC+** [**Bravo González-Blas et al., 2023**] accessible regions + motif enrichment of these regions are used to pre-filter candidate enhancers

Want to know more on network inference?

Some useful benchmarks:

- [Saint-Antoine and Singh, 2023]
- [Kang et al., 2021]
- [Hawe et al., 2019]
- [Marbach et al., 2012]: DREAM5 (simulated and real data)

End of the story!

Questions?

Credits

- Omics image is my own work but using as a base image one of the old illustration of the ENCODE project
- Image on DNA transcription and RNA translation (simplified) is "Transcription and Translation" by Christinelmiller from Wikimedia Commons
- Image on DNA transcription and RNA translation (with sequence) is by OpenStax from Wikimedia Commons
- Image on expression regulation is by Bernstein0275 from Wikimedia Commons
- Image on RNA expression experiment is a taken from **[Griffith et al., 2015]**
- Image of **GENIE3** method is taken from **[Huynh-Thu et al., 2010]**
- Image of **GENIE3** results is taken from **[Huynh-Thu et al., 2010]**
- Image of **DIANE** network inference is taken from **[Cassan et al., 2021]**
- Image of **IRafNet** method is taken from **[Petralia et al., 2015]**
- Image of PPI network is by Häuser et al. from Wikimedia Commons
- Image of Laplacian eigenvector decomposition is taken from **[Rapaport et al., 2007]**
- Images of TFBS priors and weights is taken from **[Cassan, 2022]**

References

(unofficial) Beamer template made with the help of Thomas Schiex, Matthias Zytnicki and Andreea Dreau:
https://forgemia.inra.fr/nathalie.villa-vialaneix/bainrae

Aibar, S., Gonzàlez-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., and Aerts, S. (2017).
SCENIC: single-cell regulatory network inference and clustering.
Nature Methods, 14:1083–1086.

Bravo González-Blas, C., De Winter, S., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V., Poovathingal, S., Wouters, J., Aibar, S., and Aerts, S. (2023).
SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks.
Nature Methods, 20:1355–1367.

Butte, A. and Kohane, I. (1999).
Unsupervised knowledge discovery in medical databases using relevance networks.
In *Proceedings of the AMIA Symposium*, pages 711–715.

Butte, A. and Kohane, I. (2000).
Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.
In *Proceedings of the Pacific Symposium on Biocomputing*, pages 418–429.

Cassan, O. (2022).
Inférence statistique des réseaux de régulation de gènes chez Arabidopsis thaliana en réponse à l 'élévation des teneurs en CO2 athmosphérique.
Phd thesis, Université de Montpelliers.

Cassan, O., Lèbre, S., and Martin, A. (2021).
Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite.
BMC Genomics, 22:387.

Cassan, O., Lecellier, C.-H., Bréhélin, L., Martin, A., and Lèbre, S. (2023).
Integration of transcription factor binding sites to gene expression data improves regression-based gene regulatory network inference in Arabidopsis thaliana.
In preparation.

Friedman, J., Hastie, T., and Tibshirani, R. (2008).
Sparse inverse covariance estimation with the graphical lasso.
Biostatistics, 9(3):432–441.

Griffith, M., Walker, J. R., Spies, N. C., Ainscough, B. J., and Griffith, O. L. (2015).
Informatics for RNA sequencing: a web resource for analysis on the cloud.
PLOS Computational Biology, 11(8):e1004393.

Hawe, J. S., Theis, F. J., and Heinig, M. (2019).
Inferring interaction networks from multi-omics data.
Frontiers in Genetics, 10:535.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010).
Inferring regulatory networks from expression data using tree-based methods.
PLoS ONE, 5(9):e12776.

Kang, Y., Thieffry, D., and Cantini, L. (2021).
Evaluating the reproducibility of single-cell gene regulatory network inference algorithms.
Frontiers in Genetics, 12:362.

Kondor, R. I. and Lafferty, J. (2002).

Diffusion kernels on graphs and other discrete structures.
In Sammut, C. and Hoffmann, A., editors, *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, Sydney, Australia. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Marbach, D., Costello, J. C., Küffner, R., Vega, N., Prill, R. J., Camacho, D. M., Allison, K. R., the DREAM5 Consortium, Kellis, M., and Collins, James J.and Stolovitsky, G. (2012).
Wisdom of crowds for robust gene network inference.
Nature Methods, 9(8):796–804.

Meinshausen, N. and Bühlmann, P. (2006).
High dimensional graphs and variable selection with the Lasso.
Annals of Statistic, 34(3):1436–1462.

Petralia, F., Wang, P., Yang, J., and Zhidong, T. (2015).
Integrative random forest for gene regulatory network inference.
Bioinformatics, 31(12):i197–i205.

Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., and Vert, J.-P. (2007).
Classification of microarray data using gene networks.
BMC Bioinformatics, 8:35.

Saint-Antoine, M. and Singh, A. (2023).
Benchmarking gene regulatory network inference methods on simulated and experimental data.
bioRxiv preprint.