Mining co-expression networks

Nathalie Villa-Vialaneix

http://www.nathalievilla.org

INRA, Unité MIA-T, INRA, Toulouse (France)

School for advanced sciences of Luchon
Network analysis and applications

Outline

1. Overview on co-expression network analysis

2. Case study 1: gene network analysis in relations with meat quality

3. Case study 2: gene network analysis in LCD experiment

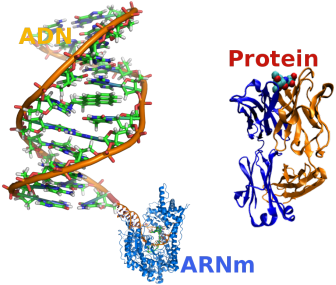
Outline

1 Overview on co-expression network analysis

2 Case study 1: gene network analysis in relations with meat quality

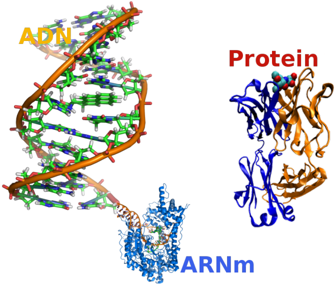
3 Case study 2: gene network analysis in LCD experiment

Transcriptomic data



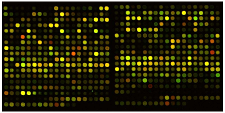
DNA transcripted into mRNA
to produce proteins

Transcriptomic data

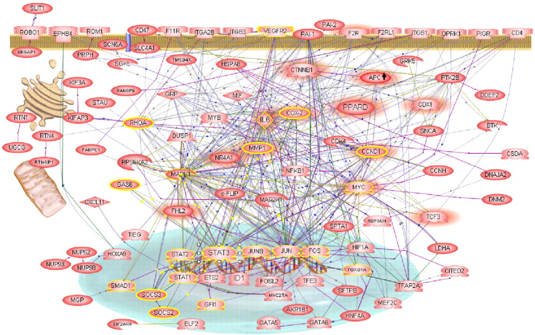


DNA transcripted into mRNA to produce proteins

transcriptomic data: measure of the quantity of mRNA corresponding to a given gene in given cells (blood, muscle...) of a living organism



Systems biology



Some genes' expressions activate or repress other genes' expressions ⇒ understanding the whole cascade helps to comprehend the global functioning of living organisms[1]

[1] Picture taken from: Abdollahi A *et al.*, *PNAS* 2007, **104**:12890-12895. © 2007 by National Academy of Sciences

Standard issues in network analysis

Inference

Giving expression data, how to build a graph whose edges represent the direct links between genes?

Example: co-expression networks built from microarray/RNAseq data (nodes = genes; edges = significant “direct links” between expressions of two genes)

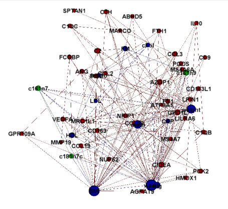
Standard issues in network analysis

Inference

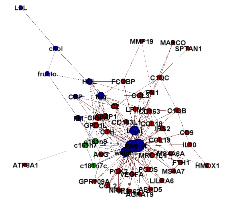
Giving expression data, how to build a graph whose edges represent the direct links between genes?

Graph mining (examples)

1. Network visualization: nodes are not a priori given a position.



Random positions



Positions aiming at representing connected nodes closer

Standard issues in network analysis

Inference

Giving expression data, how to build a graph whose edges represent the direct links between genes?

Graph mining (examples)

1. Network visualization: nodes are not a priori given a position.
2. Important node extraction (high degree, high centrality...)

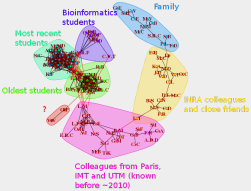
Standard issues in network analysis

Inference

Giving expression data, how to build a graph whose edges represent the direct links between genes?

Graph mining (examples)

1. Network visualization: nodes are not a priori given a position.
2. Important node extraction (high degree, high centrality...)
3. Network clustering: identify "communities"


Family
Bioinformatics students
Most recent students
Oldest students
?
INRA colleagues and close friends
Colleagues from Paris, IMT and UTM (known before ~2010)

Network inference

Data: large scale gene expression data

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

What we want to obtain: a graph/network with

- nodes: genes (a selected sublist of interest[2]; usually, DE genes);
- edges: “strong relations” between gene expressions.

[2] See [Verzelen, 2012] for conditions on respective n/p suited for inference.

Advantages of this network model

1. over raw data: focuses on the strongest direct relationships: irrelevant or indirect relations are removed (more robust) and the data are easier to visualize and understand (track transcription relations).

Advantages of this network model

1. over raw data: focuses on the strongest direct relationships: irrelevant or indirect relations are removed (more robust) and the data are easier to visualize and understand (track transcription relations).
   Expression data are analyzed all together and not by pairs (systems model).

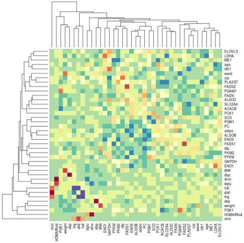
Advantages of this network model

1. over raw data: focuses on the strongest direct relationships: irrelevant or indirect relations are removed (more robust) and the data are easier to visualize and understand (track transcription relations).
   Expression data are analyzed all together and not by pairs (systems model).

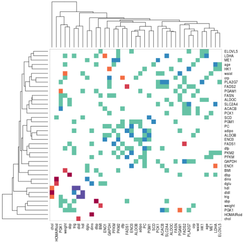
2. over bibliographic network: can handle interactions with yet unknown (not annotated) genes and deal with data collected in a particular condition.

Using *correlations*: relevance of the network [Butte and Kohane, 1999, Butte and Kohane, 2000]

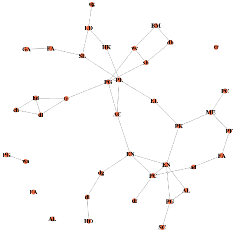
First (naive) approach: calculate correlations between expressions for all pairs of genes, threshold the smallest ones and build the network.



Correlations

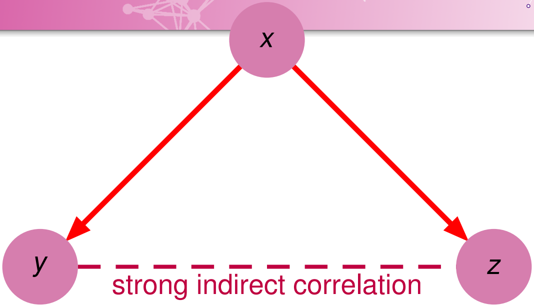


Thresholding

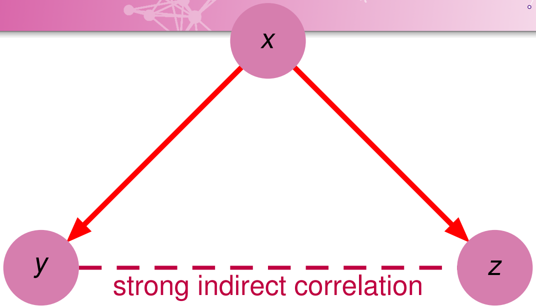


Graph

Using *partial* correlations

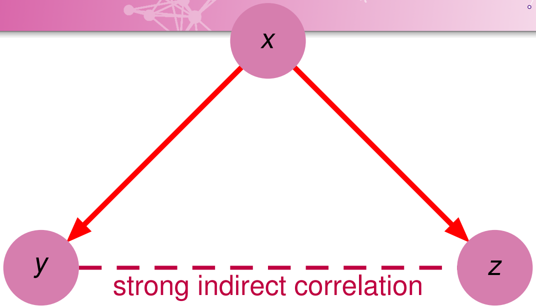


Using *partial* correlations



```
set.seed(2807); x <- rnorm(100)
y <- 2*x+1+rnorm(100,0,0.1); cor(x,y) [1] 0.998826
z <- 2*x+1+rnorm(100,0,0.1); cor(x,z) [1] 0.998751
cor(y,z) [1] 0.9971105
```

Using *partial* correlations



```
set.seed(2807); x <- rnorm(100)
y <- 2*x+1+rnorm(100,0,0.1); cor(x,y) [1] 0.998826
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(x~z)$residuals,lm(y~z)$residuals) [1] 0.7801174
cor(lm(x~y)$residuals,lm(z~y)$residuals) [1] 0.7639094
cor(lm(y~x)$residuals,lm(z~x)$residuals) [1] -0.1933699
```

Partial correlation and GGM

Gaussian Graphical Model framework:

$(X_i)_{i=1,\ldots,n}$ are i.i.d. Gaussian random variables $\mathcal{N}(0,\Sigma)$ (gene expression); then

$$j \longleftrightarrow j' \text{(genes } j \text{ and } j' \text{ are linked)} \Leftrightarrow \mathbb{C}\mathrm{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) \neq 0$$

Partial correlation and GGM

Gaussian Graphical Model framework:

$(X_i)_{i=1,\ldots,n}$ are i.i.d. Gaussian random variables $\mathcal{N}(0,\Sigma)$ (gene expression); then

$$j \longleftrightarrow j' \text{(genes } j \text{ and } j' \text{ are linked)} \Leftrightarrow \mathbb{C}\mathrm{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) \neq 0$$

If (concentration matrix) $S = \Sigma^{-1}$,

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

$\Rightarrow$ Estimate Σ^{-1} to unravel the graph structure

Partial correlation and GGM

Gaussian Graphical Model framework:

$(X_i)_{i=1,\ldots,n}$ are i.i.d. Gaussian random variables $\mathcal{N}(0,\Sigma)$ (gene expression); then

$$j \longleftrightarrow j' \text{(genes } j \text{ and } j' \text{ are linked)} \Leftrightarrow \mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) \neq 0$$

If (concentration matrix) $S = \Sigma^{-1}$,

$$\mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) = -\frac{S_{jj'}}{\sqrt{S_{jj}S_{j'j'}}}$$

$\Rightarrow$ Estimate Σ^{-1} to unravel the graph structure

Problem: Σ: p-dimensional matrix and $n \ll p \Rightarrow (\widehat{\Sigma}^n)^{-1}$ is a poor estimate of S)!

Estimation in GGM

Graphical Gaussian Model estimation

- seminal work:
  [Schäfer and Strimmer, 2005a, Schäfer and Strimmer, 2005b] (with shrinkage and a proposal for a Bayesian test of significance)
  - estimate Σ^{-1} by $(\widehat{\Sigma}^n + \lambda \mathbb{I})^{-1}$
  - use a Bayesian test to test which coefficients are significantly non zero.

Estimation in GGM

Graphical Gaussian Model estimation

- seminal work:
  [Schäfer and Strimmer, 2005a, Schäfer and Strimmer, 2005b] (with shrinkage and a proposal for a Bayesian test of significance)
  - estimate Σ^{-1} by $(\widehat{\Sigma}^n + \lambda \mathbb{I})^{-1}$
  - use a Bayesian test to test which coefficients are significantly non zero.
- sparse approaches:
  [Meinshausen and Bühlmann, 2006, Friedman et al., 2008]: $\forall\, j$, estimate the linear models $X^j = \beta_j^T X^{-j} + \epsilon$ by penalized ML
  $\arg\min_{(\beta_{jj'})_{j'}} \sum_{i=1}^n \left(X_{ij} - \beta_j^T X_i^{-j}\right)^2 + \lambda \|\beta_j\|_{L^1}$, with
  $\|\beta_j\|_{L^1} = \sum_{j'} |\beta_{jj'}|$, L^1 penalty yields to $\beta_{jj'} = 0$ for most j'
  (variable selection)

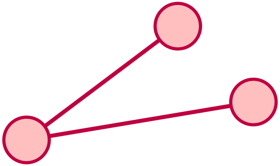
Visualization

Purpose: How to display the nodes in a meaningful and aesthetic way?

Visualization

Purpose: How to display the nodes in a meaningful and aesthetic way?

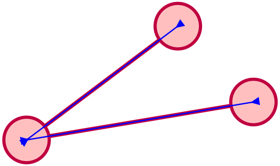
Standard approach: force directed placement algorithms (FDP) (e.g., [Fruchterman and Reingold, 1991])



Visualization

Purpose: How to display the nodes in a meaningful and aesthetic way?

Standard approach: force directed placement algorithms (FDP) (e.g., [Fruchterman and Reingold, 1991])

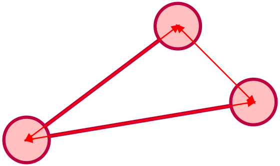


- attractive forces: similar to springs along the edges

Visualization

Purpose: How to display the nodes in a meaningful and aesthetic way?

Standard approach: force directed placement algorithms (FDP) (e.g., [Fruchterman and Reingold, 1991])

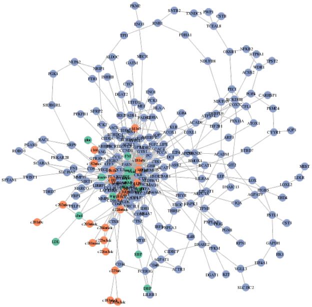


- attractive forces: similar to springs along the edges
- repulsive forces: similar to electric forces between all pairs of vertices

Visualization

Purpose: How to display the nodes in a meaningful and aesthetic way?

Standard approach: force directed placement algorithms (FDP) (e.g., [Fruchterman and Reingold, 1991])



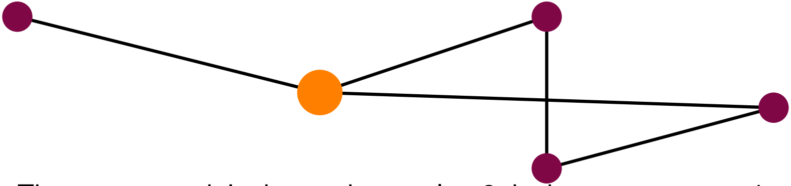
iterative algorithm until stabilization of the vertex positions.

Important node extraction

1. vertex degree: number of edges adjacent to a given vertex. Vertices with a high degree are called hubs: measure of the vertex popularity.

Important node extraction

1. vertex degree: number of edges adjacent to a given vertex. Vertices with a high degree are called hubs: measure of the vertex popularity.
2. vertex betweenness: number of shortest paths between all pairs of vertices that pass through the vertex. Betweenness is a centrality measure (vertices with a large betweenness that are the most likely to disconnect the network if removed).



The orange node's degree is equal to 2, its betweenness to 4.

Vertex clustering

Cluster vertexes into groups that are densely connected and share a few links (comparatively) with the other groups. Clusters are often called communities (social sciences) or modules (biology).

Vertex clustering

Cluster vertexes into groups that are densely connected and share a few links (comparatively) with the other groups. Clusters are often called communities (social sciences) or modules (biology).

Several clustering methods:

- min cut minimization minimizes the number of edges between clusters;
- spectral clustering [von Luxburg, 2007] and kernel clustering uses eigen-decomposition of the Laplacian

$$L_{ij} = \begin{cases} -w_{ij} & \text{if } i \neq j \\ d_i & \text{otherwise} \end{cases}$$

  (matrix strongly related to the graph structure);
- Generative (Bayesian) models [Zanghi et al., 2008];
- Markov clustering simulate a flow on the graph;
- modularity maximization
- ... (clustering jungle... see e.g., [Fortunato and Barthélémy, 2007, Schaeffer, 2007, Brohée and van Helden, 2006])

Modularity optimization

The modularity [Newman and Girvan, 2004] of the partition $(\mathcal{C}_1, \ldots, \mathcal{C}_K)$ is equal to:

$$\mathcal{Q}(\mathcal{C}_1, \ldots, \mathcal{C}_K) = \frac{1}{2m} \sum_{k=1}^{K} \sum_{x_i, x_j \in \mathcal{C}_k} (W_{ij} - P_{ij})$$

with P_{ij}: weight of a “null model” (graph with the same degree distribution but no preferential attachment):

$$P_{ij} = \frac{d_i d_j}{2m}$$

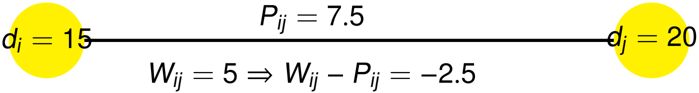
with $d_i = \frac{1}{2} \sum_{j \neq i} W_{ij}$.

Interpretation

A good clustering should maximize the modularity:

- $\mathcal{Q} \nearrow$ when (x_i, x_j) are in the same cluster and $W_{ij} \gg P_{ij}$
- $\mathcal{Q} \searrow$ when (x_i, x_j) are in two different clusters and $W_{ij} \gg P_{ij}$

$(m = 20)$



$d_i = 15$

$P_{ij} = 7.5$

$d_j = 20$

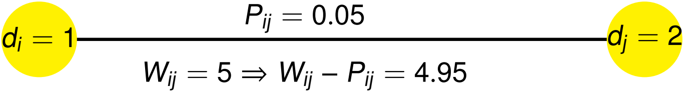
$W_{ij} = 5 \Rightarrow W_{ij} - P_{ij} = -2.5$

i and *j* in the same cluster decreases the modularity

Interpretation

A good clustering should maximize the modularity:

- $Q \nearrow$ when (x_i, x_j) are in the same cluster and $W_{ij} \gg P_{ij}$
- $Q \searrow$ when (x_i, x_j) are in two different clusters and $W_{ij} \gg P_{ij}$ $(m = 20)$



$d_i = 1$

$P_{ij} = 0.05$

$d_j = 2$

$W_{ij} = 5 \Rightarrow W_{ij} - P_{ij} = 4.95$

i and *j* in the same cluster increases the modularity

Interpretation

A good clustering should maximize the modularity:

- $Q \nearrow$ when (x_i, x_j) are in the same cluster and $W_{ij} \gg P_{ij}$
- $Q \searrow$ when (x_i, x_j) are in two different clusters and $W_{ij} \gg P_{ij}$
- Modularity
  - helps separate hubs ($\neq$ spectral clustering or min cut criterion);
  - is not an increasing function of the number of clusters: useful to choose the relevant number of clusters (with a grid search: several values are tested, the clustering with the highest modularity is kept) but modularity has a small resolution default (see **[Fortunato and Barthélémy, 2007]**)

Interpretation

A good clustering should maximize the modularity:

- $\mathcal{Q} \nearrow$ when (x_i, x_j) are in the same cluster and $W_{ij} \gg P_{ij}$
- $\mathcal{Q} \searrow$ when (x_i, x_j) are in two different clusters and $W_{ij} \gg P_{ij}$
- Modularity
    - helps separate hubs ($\neq$ spectral clustering or min cut criterion);
    - is not an increasing function of the number of clusters: useful to choose the relevant number of clusters (with a grid search: several values are tested, the clustering with the highest modularity is kept) but modularity has a small resolution default (see **[Fortunato and Barthélémy, 2007]**)

Main issue: Optimization = NP-complete problem (exhaustive search is not not usable)

Different solutions are provided in **[Newman and Girvan, 2004, Blondel et al., 2008, Noack and Rotta, 2009, Rossi and Villa-Vialaneix, 2011]** (among others) and some of them are implemented in the R package **igraph**.

Outline

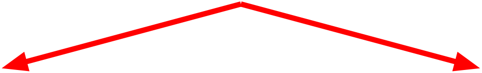
1. Overview on co-expression network analysis

2. Case study 1: gene network analysis in relations with meat quality

3. Case study 2: gene network analysis in LCD experiment

Dataset description

[Villa-Vialaneix et al., 2013]



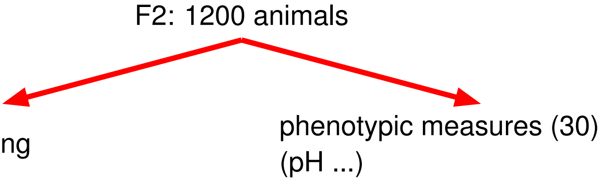
F2: 1200 animals

muscle sampling

phenotypic measures (30) (pH ...)

Dataset description

[Villa-Vialaneix et al., 2013]



F2: 1200 animals

muscle sampling

phenotypic measures (30)
(pH ...)

Used data: 57 F2 pigs (largest variability for PH); transcriptomic data for 272 genes regulated by an eQTL

Problems with these particular data:

- how to understand the relationships between these genes' expression as their co-expression is weaker than between other kind of genes (TF/genes, for instance)?
- how to relate gene expression with a phenotype of interest (muscle pH)?

Inferred network description

Use of [Schäfer and Strimmer, 2005a]

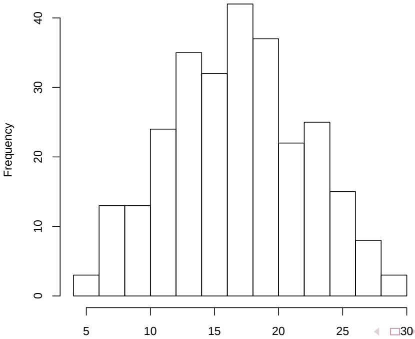
Obtained network: 272 nodes (connected); Density: 6,4%; Transitivity: 25,4%

Inferred network description

Use of [Schäfer and Strimmer, 2005a]

Obtained network: 272 nodes (connected); Density: 6,4%; Transitivity: 25,4%

degree distribution



Inferred network description

Use of [Schäfer and Strimmer, 2005a]

Obtained network: 272 nodes (connected); Density: 6,4%; Transitivity: 25,4%

8 genes both have high degree and high betweenness

BX921641; FTH1; TRIAP1; SLC9A14; GPI; SUZ12; MGP; PRDX4

and several have been identified by the biologist as relevant to meat quality.

Clustering

- clustering with modularity optimization: 7 clusters
- for each cluster, annotated genes submitted to IPA[3] (bibliographic network database): from 71% to 94% of the genes of a single cluster belong to the same IPA network with a biological function associated

[3] https://analysis.ingenuity.com/pa

Relation to muscle pH

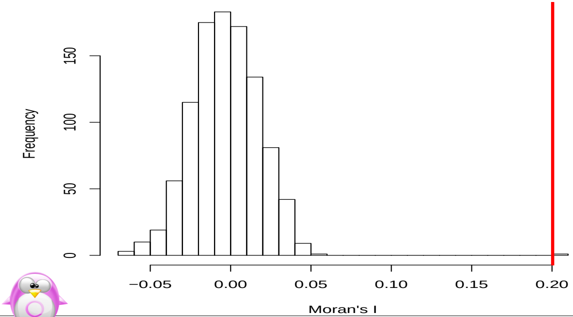
model: label each node of the network with its partial correlation to the muscle pH.

Questions: is there a relation between muscle pH and network structure? is there a relation between clustering and muscle pH?

Relation to muscle pH

model: label each node of the network with its partial correlation to the muscle pH.

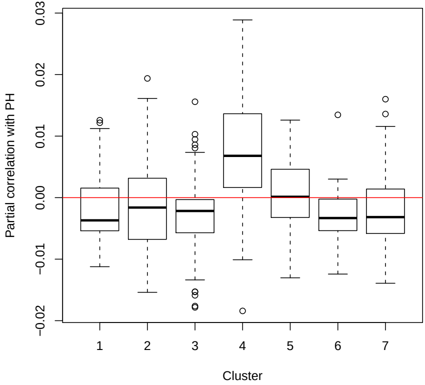
Moran's I (used in spatial statistics): $\mathbf{I} = \frac{\frac{1}{2m}\sum_{i\neq j} w_{ij}\bar{c}_i\bar{c}_j}{\frac{1}{n}\sum_i \bar{c}_i^2}$, where $m = \frac{1}{2}\sum_{i\neq j} W_{ij}$ and c_i is the partial correlation with pH, $\bar{c}_i = c_i - \bar{c}$ with $\bar{c} = \frac{1}{n}\sum_i c_i$. Using a MC simulation (edge permutations):



Moran's *I* is significantly larger than expected: genes tend to be linked to genes which have a similar correlation to muscle pH.

Relation to muscle pH

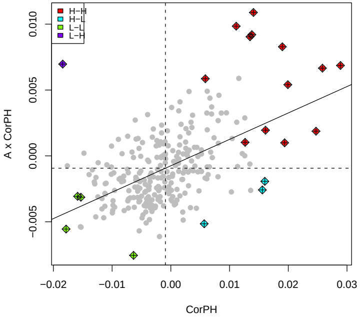
model: label each node of the network with its partial correlation to the muscle pH.



Significant Student test for luster 4: its correlation with pH is larger than for the other clusters

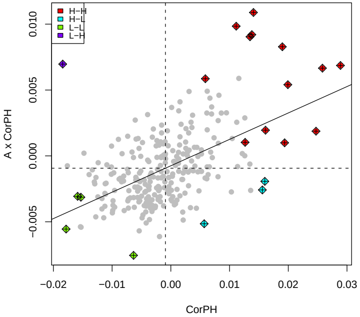
Moran's plot

Moran's plot help to emphasize influential points: *WC* vs *C*



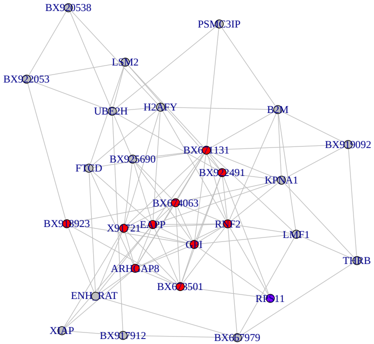
Moran's plot

Moran's plot help to emphasize influential points: *WC* vs *C*



Associated influential measures and tests for finding influential points.

Influential points: example of clusters



Cluster 4

Outline

1 Overview on co-expression network analysis

2 Case study 1: gene network analysis in relations with meat quality

3 Case study 2: gene network analysis in LCD experiment

Data: DIOGENES project

Experimental protocol

135 obese women and 3 times: before LCD, after a 2-month LCD and 6 months later (between the end of LCD and the last measurement, women are randomized into one of 5 recommended diet groups).

At every time step, 221 gene expressions, 28 fatty acids and 15 clinical variables (i.e., weight, HDL, ...)

Data: DIOGENES project

Experimental protocol

135 obese women and 3 times: before LCD, after a 2-month LCD and 6 months later (between the end of LCD and the last measurement, women are randomized into one of 5 recommended diet groups).

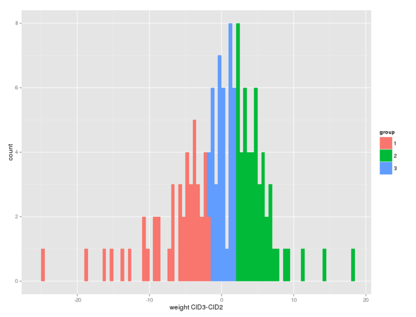
At every time step, 221 gene expressions, 28 fatty acids and 15 clinical variables (i.e., weight, HDL, ...)

Correlations between gene expressions and between a gene expression and a fatty acid levels are not of the same order: inference method must be different inside the groups and between two groups.

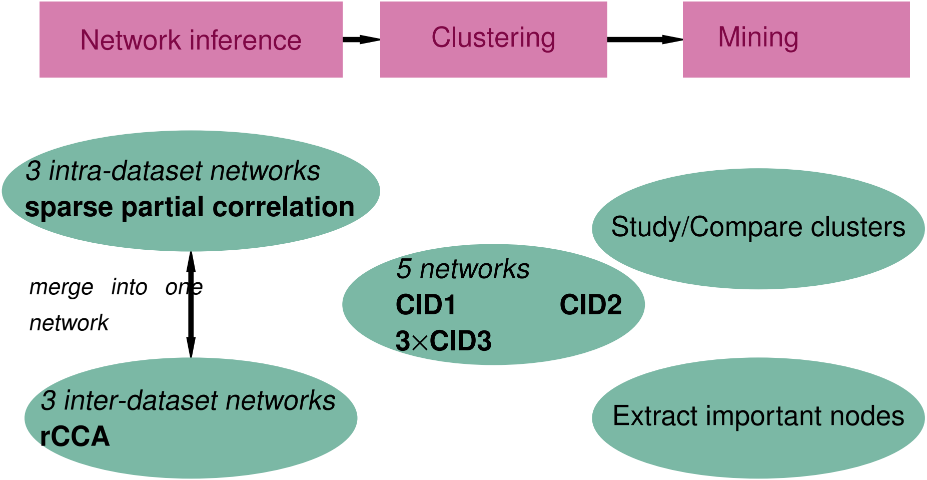
Data: DIOGENES project

Data pre-processing

At CID3, individuals are split into three groups: weight loss, weight regain and stable weight (groups are not correlated to the diet group according to χ^2-test).



Method for CID1, CID2 and 3× CID3



Network inference → Clustering → Mining

3 intra-dataset networks
sparse partial correlation

merge into one network

3 inter-dataset networks
rCCA

5 networks
CID1 **CID2**
3×CID3

Study/Compare clusters

Extract important nodes

Inference

Intra-level networks: use of partial correlations and a sparse approach (graphical Lasso as in the R package **gLasso**) to select edges [Friedman et al., 2008]

Inference

Intra-level networks: use of partial correlations and a sparse approach (graphical Lasso as in the R package **gLasso**) to select edges [Friedman et al., 2008]

Inter-levels networks: use of regularized CCA (as in the R package **mixOmics**) to evaluate strength of the correlations [Lê Cao et al., 2009]

Inference

Intra-level networks: use of partial correlations and a sparse approach (graphical Lasso as in the R package **gLasso**) to select edges [Friedman et al., 2008]

Inter-levels networks: use of regularized CCA (as in the R package **mixOmics**) to evaluate strength of the correlations [Lê Cao et al., 2009]

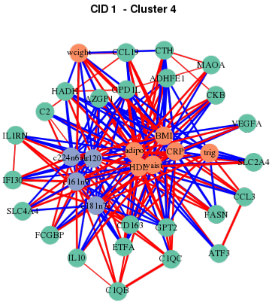
Combination of the 6 informations: tune the number of edges intra or inter-levels so that it is of the order of the number of nodes in the corresponding level(s)

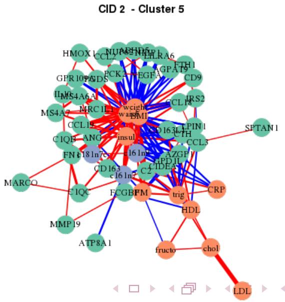
Brief overview on results

5 networks inferred with 264 nodes each:

	CID1	CID2	CID3g1	CID3g2	CID3g3
size LCC	244	251	240	259	258
density	2.3%	2.3%	2.3%	2.3%	2.3%
transitivity	17.2%	11.9%	21.6%	10.6%	10.4%
nb clusters	14 (2-52)	10 (4-52)	11 (2-46)	12 (2-51)	12 (3-54)

clusters were visualized and analyzed for important node extraction


CID 1 - Cluster 4


CID 2 - Cluster 5

Findings

- CID1: clusters were found to be associated to biological functions (fatty acids biosynthesis, adhesion and diapedesis...)
- CID3: for people with weight loss, an unexpected fatty acid was found to be an important node (high betweenness) in a cluster linked to fatty acids biosynthesis

Conclusion

- biological network mining can help the biologist comprehend the complex biological system in its whole
- groups of genes are more robust models, often linked to a biological function, than pairwise relations between genes
- simple tools, such as, numeric characteristics are useful to extract important nodes

Thank you for your attention...

... questions?

Blondel, V., Guillaume, J., Lambiotte, R., and Lefebvre, E. (2008).
Fast unfolding of communites in large networks.
Journal of Statistical Mechanics: Theory and Experiment, P10008:1742–5468.

Brohée, S. and van Helden, J. (2006).
Evaluation of clustering algorithms for protein-protein interaction networks.
BMC Bioinformatics, 7(488).

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.

Fortunato, S. and Barthélémy, M. (2007).
Resolution limit in community detection.
In *Proceedings of the National Academy of Sciences*, volume 104, pages 36–41.
doi:10.1073/pnas.0605965104; URL: http://www.pnas.org/content/104/1/36.abstract.

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

Fruchterman, T. and Reingold, B. (1991).
Graph drawing by force-directed placement.
Software, Practice and Experience, 21:1129–1164.

Lê Cao, K., González, I., and Déjean, S. (2009).
*****Omics: an R package to unravel relationships between two omics data sets.
Bioinformatics, 25(21):2855–2856.

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

Newman, M. and Girvan, M. (2004).
Finding and evaluating community structure in networks.
Physical Review, E, 69:026113.

Noack, A. and Rotta, R. (2009).
Multi-level algorithms for modularity clustering.
In *SEA '09: Proceedings of the 8th International Symposium on Experimental Algorithms*, pages 257–268, Berlin, Heidelberg. Springer-Verlag.

Rossi, F. and Villa-Vialaneix, N. (2011).
Représentation d'un grand réseau à partir d'une classification hiérarchique de ses sommets.
Journal de la Société Française de Statistique, 152(3):34–65.

Schaeffer, S. (2007).
Graph clustering.
Computer Science Review, 1(1):27–64.

Schäfer, J. and Strimmer, K. (2005a).
An empirical bayes approach to inferring large-scale gene association networks.
Bioinformatics, 21(6):754–764.

Schäfer, J. and Strimmer, K. (2005b).
A shrinkage approach to large-scale covariance matrix estimation and implication for functional genomics.
Statistical Applications in Genetics and Molecular Biology, 4:1–32.

Verzelen, N. (2012).
Minimax risks for sparse regressions: ultra-high-dimensional phenomenons.
Electronic Journal of Statistics, 6:38–90.

Villa-Vialaneix, N., Liaubet, L., Laurent, T., Cherel, P., Gamot, A., and San Cristobal, M. (2013).
The structure of a gene co-expression network reveals biological functions underlying eQTLs.
PLoS ONE, 8(4):e60045.

von Luxburg, U. (2007).
A tutorial on spectral clustering.
Statistics and Computing, 17(4):395–416.

Zanghi, H., Ambroise, C., and Miele, V. (2008).
Fast online graph clustering via erdös-rényi mixture.
Pattern Recognition, 41:3592–3599.