Complex Networks in Systems Biology

Biological Network Inference

Costas Bouyioukos

UMR7216, Paris Epigénetique et Destine Cellulaire
Université Paris Diderot

13 novembre 2018

Overview of the course

Introduction

- Contents
- Learning biological networks

Inference

- InferenceI – Data driven
- Inference II – Bayesian, Boolean
- Inference II – Differential Equations
- Inference III – Graph models
- Integrative models

Model Evaluation

Learning and inference of biological networks

1. Statistical for *de-novo* inference of networks from data.
2. Bayesian methods.
3. Machine learning methods.

TP :

Hands on experience with a popular method for network inference WGCNA.

Reconstructing network topology

Network representation models are always a prerequisite :

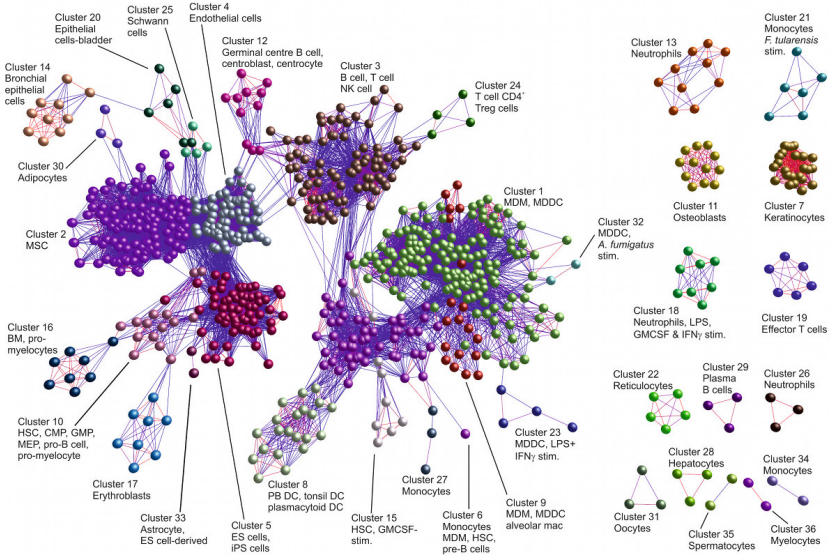
- Boolean networks, boolean functions
- Bayesian networks and dynamics
- Continuous or discrete Ordinary Differential Equations.

but also two more which we have not seen last time :

- Information theory and correlation methods
- Graph theoretical methods

From data to networks

- Relationship between data and networks is two-fold.
- Deluge of data, networks a way to represent the salient features, to compress, to capture the complexity.
- ... but also networks can provide a tool to look at data.
  Like with BioLayout Express3D

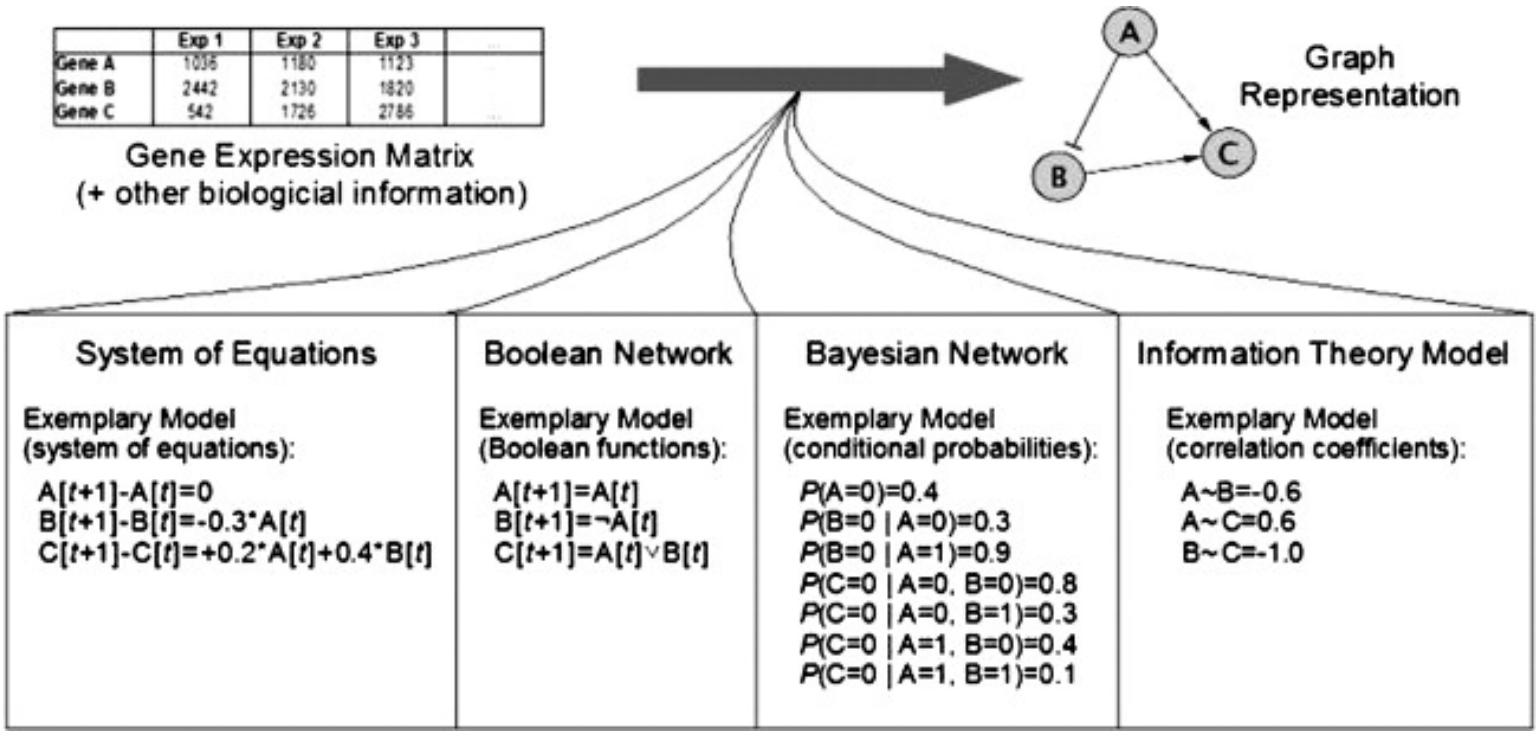


Building network learning models

Goal :

Gene expression data

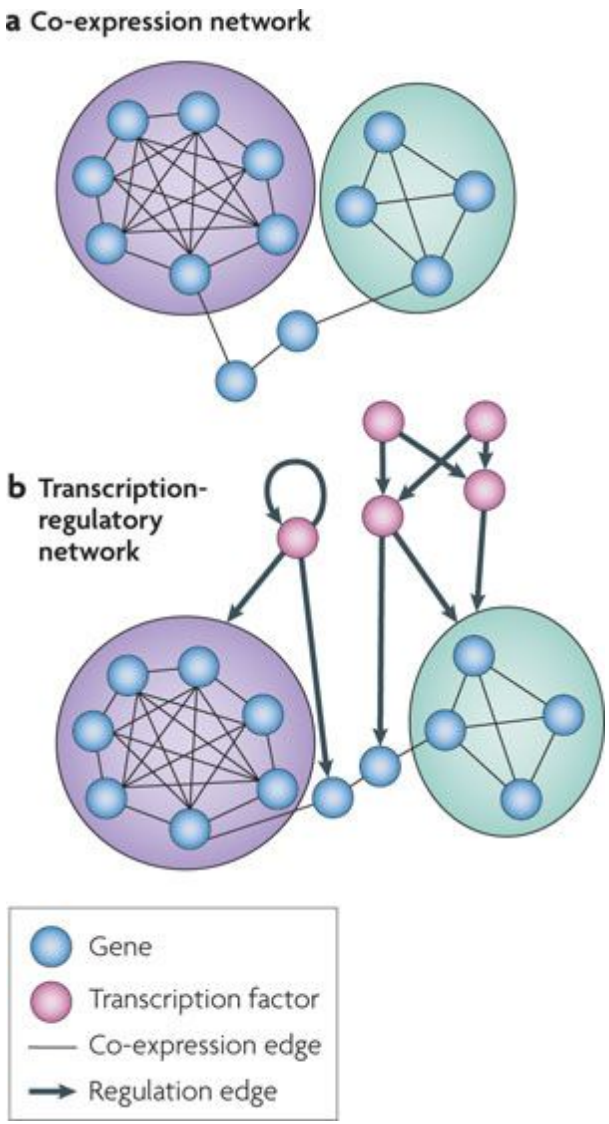
(and other biological information) $\Rightarrow$ Obtain network topology



Module based - Topology based

Two approaches

- Co-regulatory subset of genes - de-novo from a gene expression matrix
- Integrative methods, include information from various types of data.



Network inference - Undetermined problem !

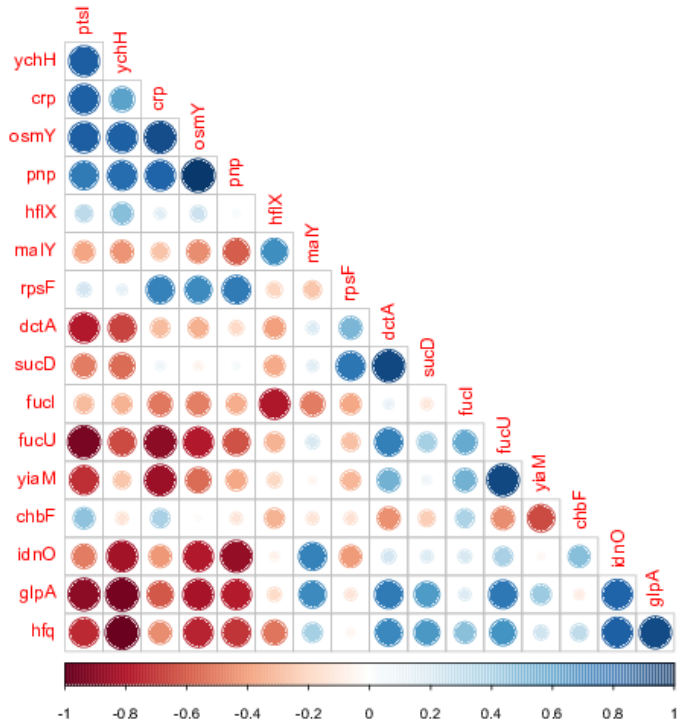
- Network inference is, mathematically, an underdetermined problem.
- large number of theoretically possible interactions between transcription factors (TFs) and their targets far exceeds the number of independent measurements from which the true interactions can be inferred.
- Inference therefore results in many possible solutions that all explain the data equally well, but only a few of these solutions can be biologically true.
- Here we will explore strategies on how to determine WHAT is biologically true.

Information theory/correlation methods

- Computing a matrix of a "characteristic measure".
  - Correlation coefficient
  - Mutual Information
- Pearson :

$$corr(X_i, X_j) = \frac{cov(X_i, X_j)}{\sigma(X_i)\sigma(X_j)}$$

Where $cov(.,.)$ is the covariance between two expression profiles X_i and X_j



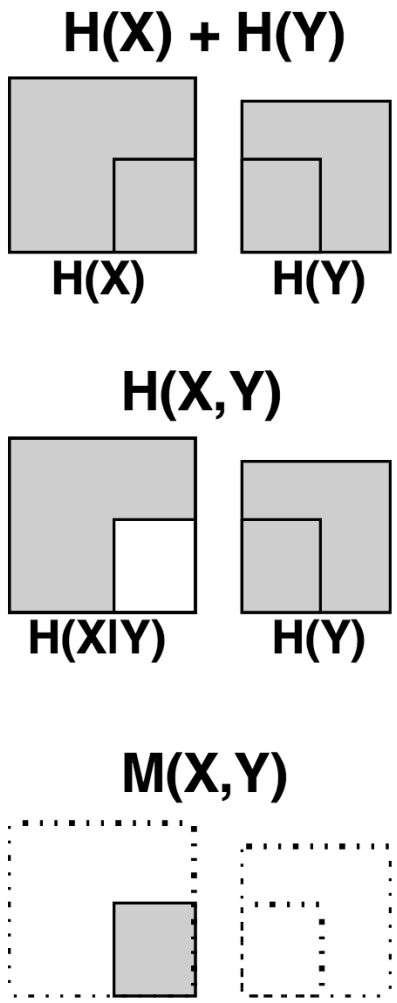
- MI :

$$I(X_i, X_j) = \sum_{x_i \in X_i} \sum_{x_j \in X_j} p(x_i, x_j) log \left(\frac{p(x_i, x_j)}{p(x_i)p(x_j)} \right)$$

Mutual information

Explanation

- Mutual Information (MI) is a information theoretical measure.
- Represents the mutual *dependency* between two random variables.
- It quantifies the amount of information (in bits) that we get for one variable through the other variable.
  i.e. how *much* information they share
- Methods based on MI are generating “Relevance Networks”

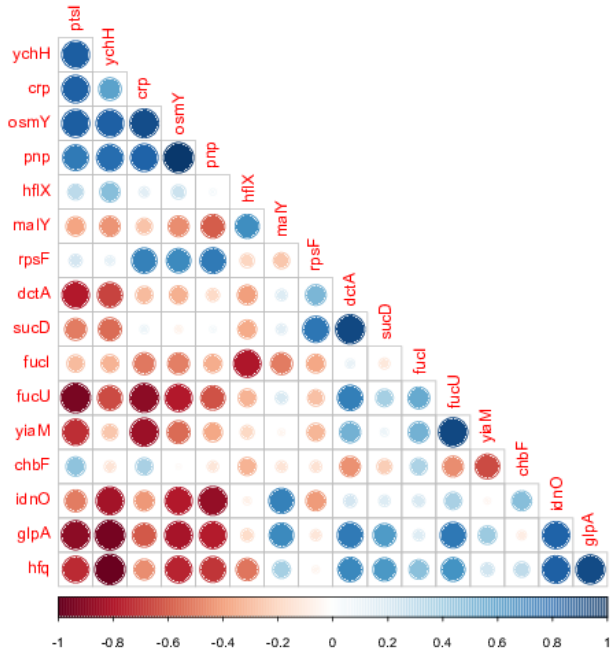


Other measures :

Rank correlations (Spearman, Kendall), Weighted correlation

Information/correlation Inference

- Methods calculate a full matrix of the measure for all ALL against ALL genes.
- Matrix is symmetric, therefore the network we obtain is un-directed.
- Matrix represents a fully connected graph, as due to noise, very few pairs will have zero correlation/MI.
- Need to specify thresholds to "prune" the network :
  - For MI, data processing inequalities.
  - For MI and correlations, compare each pairwise value against a background distribution.



MI : data processing inequality

ARACNE

- Geometric idea to remove indirect interactions based on MI properties.
- If gene *i* interacts with *j* via *k* then :

$$I(X_i, X_j) \leq \min\left(I(X_j, X_k), I(X_k, X_j)\right)$$

- If the above does not hold then there is a direct interaction.
- ARACNE goes in all triplets eliminates indirect edges -> Network pruning -> Inference

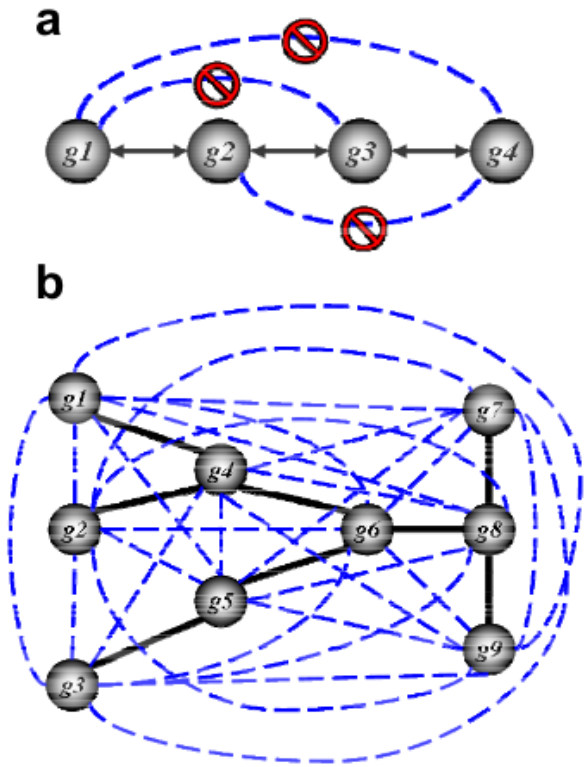
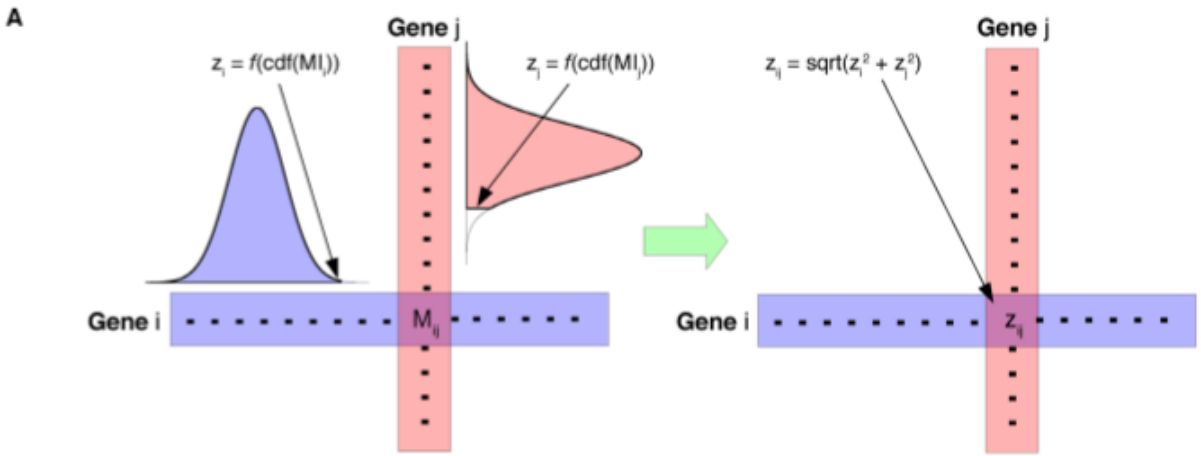


Figure 2
Examples of the data processing inequality. (a) g_1, g_2, g_3, and g_4 are connected in a linear chain relationship. Although all six gene pairs will likely have enriched mutual information, the DPI will infer the most likely path of information flow. For example, $g_1 \leftrightarrow g_3$ will be eliminated because $I(g_1, g_2) > I(g_1, g_3)$ and $I(g_2, g_3) > I(g_1, g_3)$. $g_2 \leftrightarrow g_4$ will be eliminated because $I(g_2, g_3) > I(g_2, g_4)$ and $I(g_3, g_4) > I(g_2, g_4)$. $g_1 \leftrightarrow g_4$ will be eliminated in two ways: first, because $I(g_1, g_2) > I(g_1, g_4)$ and $I(g_2, g_4) > I(g_1, g_4)$, and then because $I(g_1, g_3) > I(g_1, g_4)$ and $I(g_3, g_4) > I(g_1, g_4)$. **(b)** If the underlying interactions form a tree (and MI can be measured without errors), ARACNE will reconstruct the network exactly by removing all false candidate interactions (dashed blue lines) and retaining all true interactions (solid black lines).

Background correction methods

CLR

- Compute a background distribution of the MIs (or the CCs) from the observed values for each gene pair i, j
- The background model will be a set of all the $I(X_i, X_{(1,\ldots,n)})$ and $I(X_{(1,\ldots,n)}, X_j)$
- Then a z-score is calculated for each MI_i, MI_j

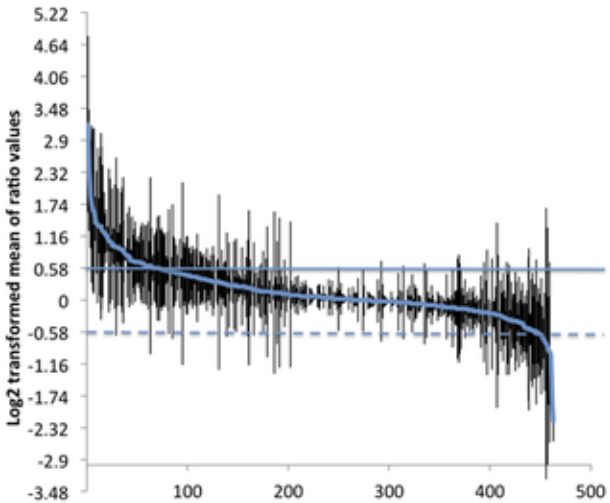


- And the mutual z-score will be $\sqrt{Z_1{}^2 + Z_2{}^2}$
- It takes into account all the gene context for both genes that's why the method is called CLR (Context Likelihood Relatedness)

Median corrected z-score

Gene knock-out Experiments

- Is using the rich information from the expression values of whole genome knock-outs
- If gene *i* interacts with *j* then its expression value is expected to be affected more than the rest of the genes in the knock-out of gene *j*.
- How much.... we can calculate it like this :

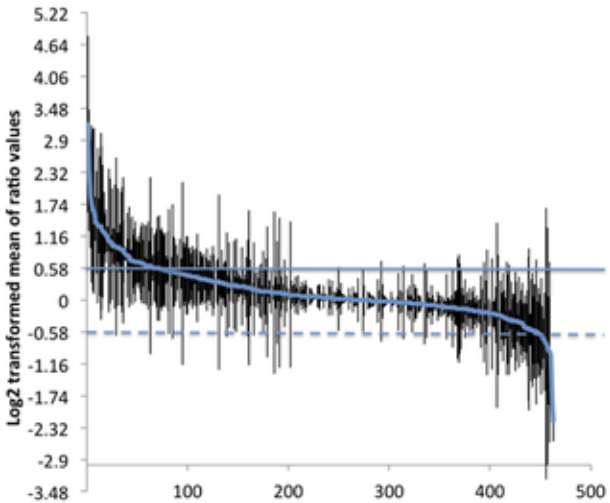


$$z(x_i|x_j^{ko}) = \frac{x_{ij}^{ko} - x_i^{wt}}{\sigma_i}$$

Median corrected z-score

Gene knock-out Experiments

- Is using the rich information from the expression values of whole genome knock-outs
- If gene *i* interacts with *j* then its expression value is expected to be affected more than the rest of the genes in the knock-out of gene *j*.
- How much.... we can calculate it like this :



$$z(x_i|x_j^{ko}) = \frac{x_{ij}^{ko} - x_i^{wt}}{\sigma_i}$$

MI inference application

- Use of the organisational principles of networks to prune many edges and to learn networks which have common properties with the “real world”.
- One of the most widely used methods based on MI is the ARACNE : wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE

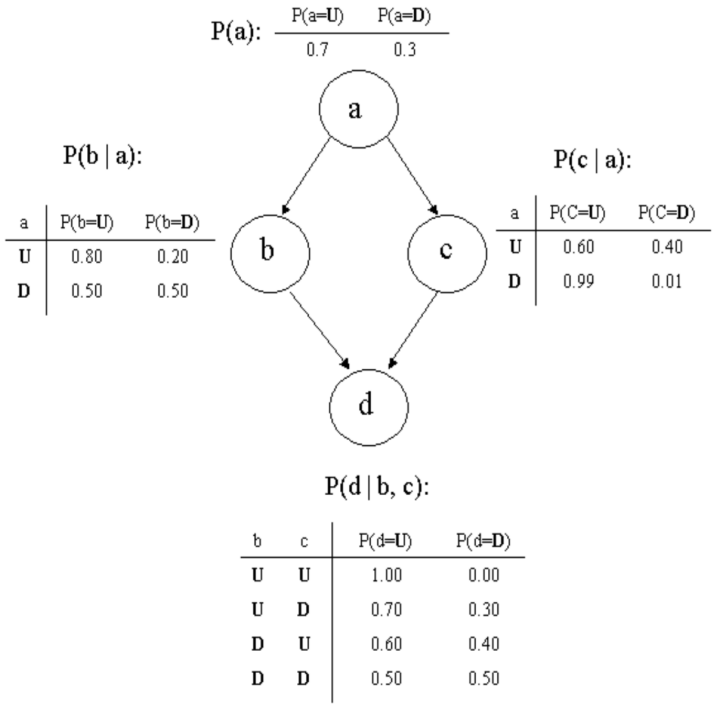
Bayesian and Boolean networks

- Bayesian Networks
  1. Model selection : Specify a DAG (Bayesian Net)
  2. Parametrisation : With the given DAG and the expression table we compute the conditional probabilities.
  3. Model validation : Each DAG gets evaluated according to a score and we select the top scored.
- Boolean nets
  - Target : To find Boolean functions which can “explain” the data from different cell states.
  - Reverse engineering techniques.
  - Current methods incomplete, can only find a set of boolean functions.

Learning Bayesian networks

1. Model Selection : The process of finding the best graph G given the data.
2. Parameter fitting : The process of finding the best set of parameters P that best describes the data.

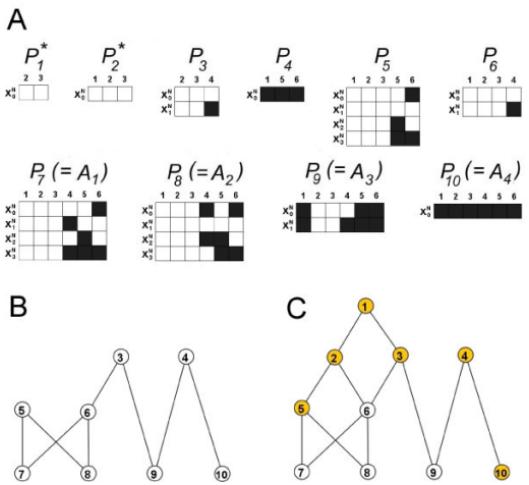
- Parameter fitting : Two very popular (and successful) algorithms :
  1. Bayesian Information Criterion (BIC)
  2. The maximum likelihood ML
  3. The expectation maximisation EM Model selection.
- We can only use heuristics !



Learning Boolean networks

1. Discretisation : All continuous variables are converted to discrete (by introducing thresholds)
2. Each edge is described by a boolean function.
3. Aim : To find ALL boolean functions in such a way that he network describe best the data.

- Reverse engineering : Examining all the possible combinations $\binom{n}{k}$ of boolean functions and employs mutual information criterion to find the co-expressed genes.

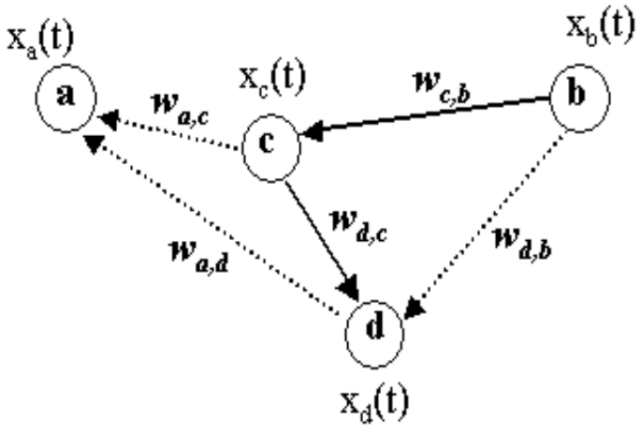


Differential equations

- Ordinary Differential Equations ODEs
  - Build a model of ODEs with linear parameters (e.g. weights of each interaction on the network)
  - Continuous differential equations can be approximated with linear difference equations (discrete in time).
  - Then typical techniques from linear algebra can be employed to solve the linear equations problem (Least square, PLS, SVD, LASSO, etc.)

Differential equations

- Linear Additive Models
- Each interaction is added (or subtracted) from the model and we also have an additional term which represents degradation.
- $\frac{dx_i(t)}{dt} = \mathrm{ext}_i + w_1 x_1(t) + w_2 x_2(t) + \ldots - dx(t)$



$$x_a(t+1) = x_a(t) + w_{a,c} x_c(t) + w_{a,d} x_d(t)$$
$$x_b(t+1) = x_b(t)$$
$$x_c(t+1) = x_c(t) + w_{c,b} x_b(t)$$
$$x_d(t+1) = x_d(t) + w_{d,c} x_c(t) + w_{d,b} x_b(t)$$

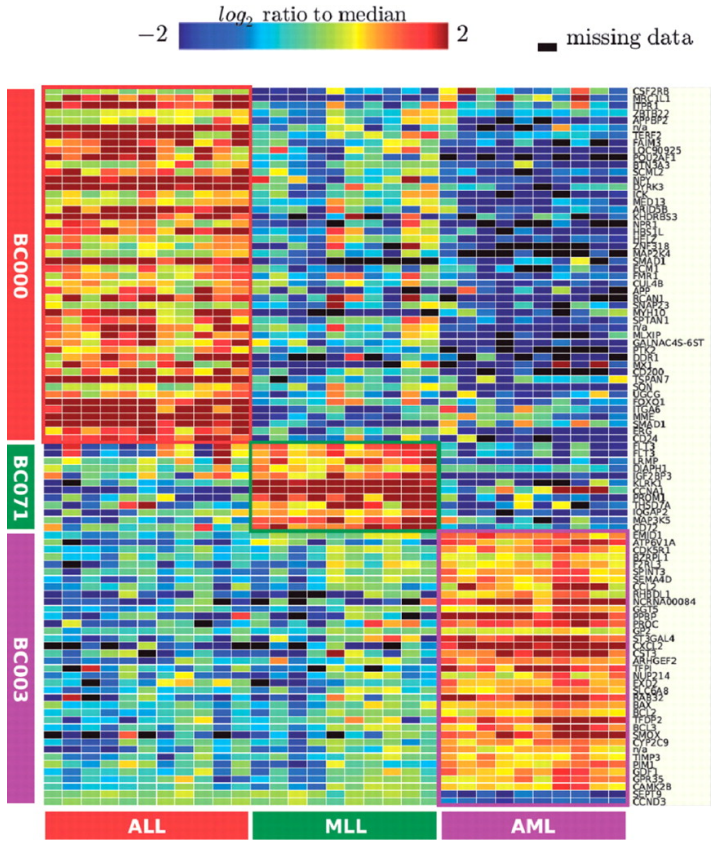
$w_{i,j}$	a	b	c	d
a	0	0	-	-
b	0	0	0	0
c	0	+	0	0
d	0	-	+	0

Graph and Information theory methods

- Graph models
  - Static models of network representation
  - Represent and condense every relation between all kinds of gene regulatory factors.
  - Gaussian Graphical Models.

Graph theory networks

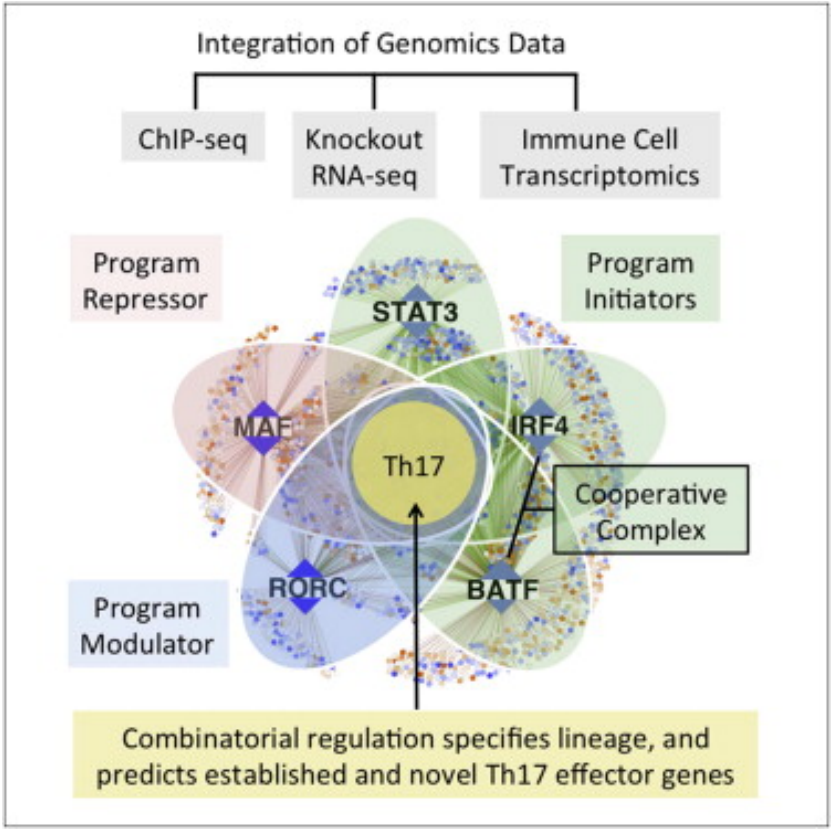
1. Input : gene expression matrix
2. Clustering step and /or biclustering step.
3. We then define two thresholds : One between and one within each group/cluster.
4. From the obtained clusters we define the co-regulated genes.
5. Then we look at the global network properties.



Learning Topology and Dynamics, the Inferelator

Multiple -omics approach

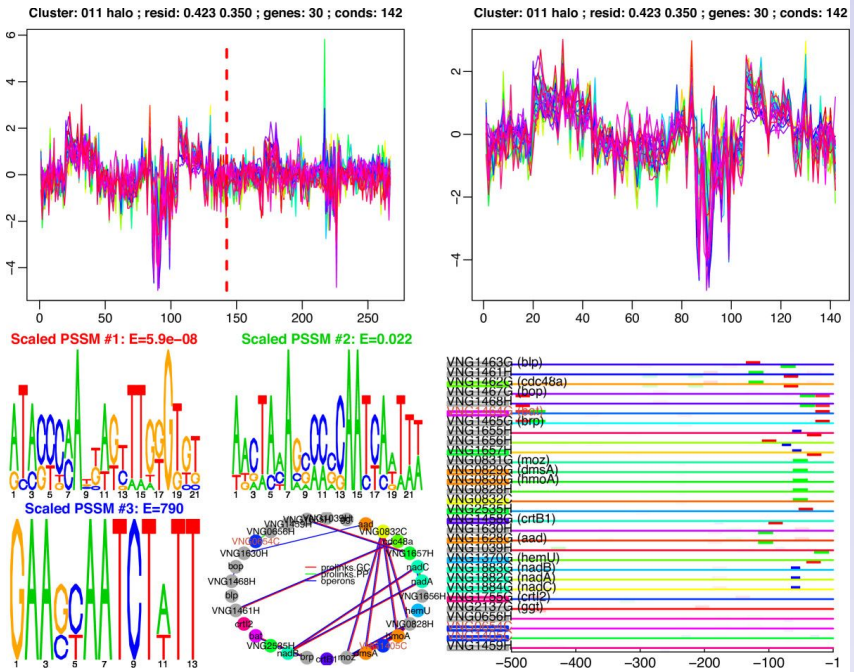
- Combine 4 supplementary datasets. Chip-Seq, RNA-seq, steady state CD4+, public expression data.
- Calculate p values from z-scores, background probability distributions and linear (penalised) regression.
- Calculate scores and integrate networks inside inferelator.



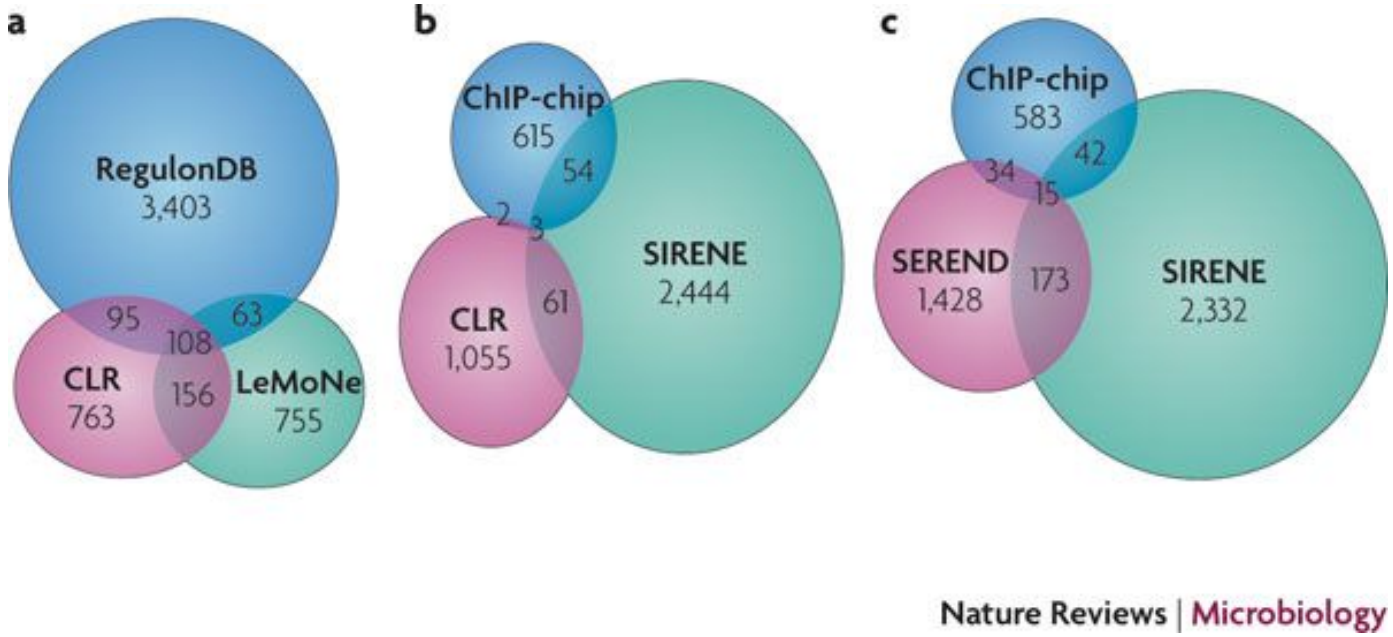
Integrating multiple sources of information

cMonkey

- cMonkey is a bi-clustering based method.
- Finds a set of biclusters in gene-expression profiles.
- Then uses all the possible motifs that can be found upstream each gene in the bi-cluster and builds the network.
- The network is further refined by information from KEGG, COGs, GOs and other functional genomics resources.



Benchmarking is a difficult challenge.

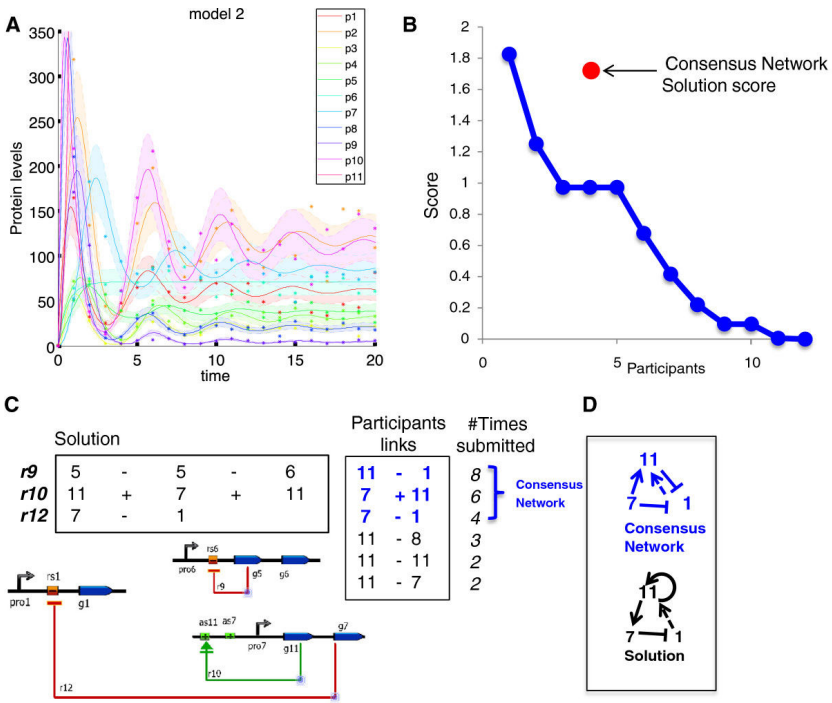


- A standard set of known interaction is composed.
- Standard sets overestimate the false-positive prediction rate, as most genes probably interact with many more TFs than is currently documented.
- To compensate for this, most current studies combine validation based on an external standard with medium-throughput experiments to also validate the new results.

DREAM

Dialogue for Reverse Engineering Assessment of Methods

- A community effort to provide a framework to systematically evaluate network inference methods.
- Parameter inference : define a parameter distance measure in the log scale)
- Topology inference : A score counts the correct source and target genes and the sign. ONLY 3 links are sought.
- Calculate p values from a randomised network distribution.



The End