Complex Networks in Systems Biology

Modelling – Clustering – Modularity

Costas Bouyioukos

UMR7216, Paris Epigénetique et Destine Cellulaire
Université Paris Diderot

8 novembre 2018

Overview of the course

Introduction
- Contents

Clustering and modularity on complex networks
- Clustering
- Modularity

Modelling
- Introduction
- Metabolism
- Boolean Networks
- Logical Models
- Ordinary Differential Equations
- Petri Networks
- Bayesian Networks

Analysis, modelling and simulation of biological networks

1. Analysis : Modular structure, clustering, community detection.
2. Simulation of network dynamics.
3. Analytical / logical models.
4. Computational models.

Bonous TP :

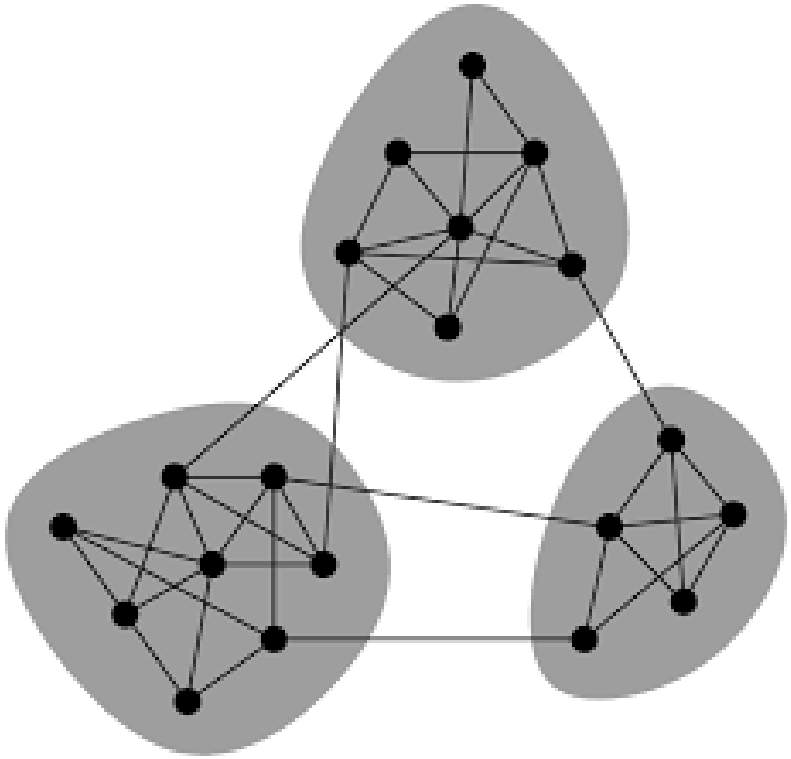
Networks in the 3D space, working and analysing the spatial architecture of genomes with GREAT

absynth.issb.genopole.fr/GREAT/

Graph Clustering – Community Detection

- Graph Clustering and Network Community Detection are two very similar yet distinct concepts.
- Here we will examine two types of algorithms

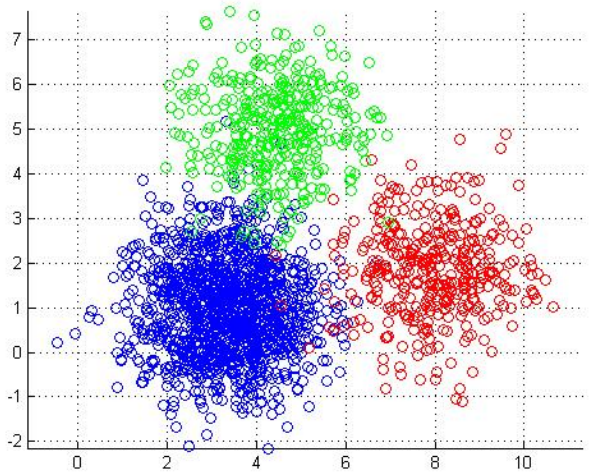
1. Clustering methods originating from statistics, CS and graph theory.
2. Community detection methods from social network science, they appear firstly in social sciences but interest in biology.

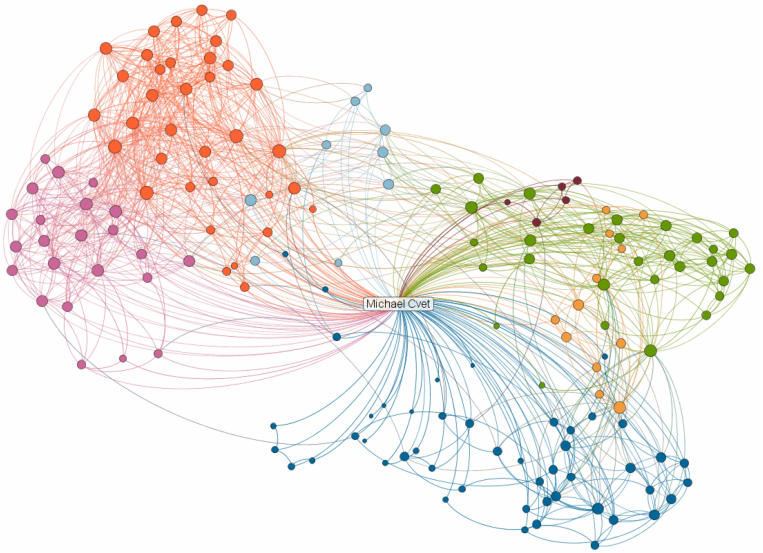


Clustering

General Instroduction

- The collection into a group of entities of the same “type”.
- Two general categories of methods :
  1. Methods based in the calculation of some distance metric.
  2. Methods based in common properties between nodes or edges.

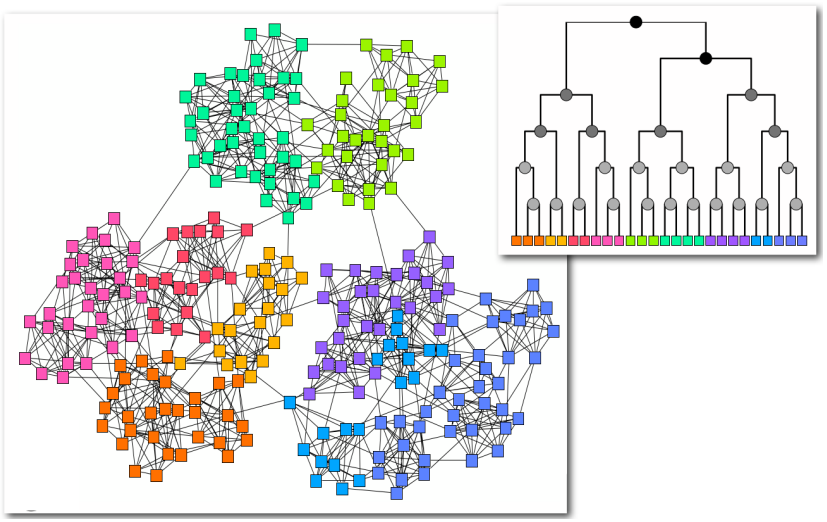




Graph clustering I

Hierarchical Clustering

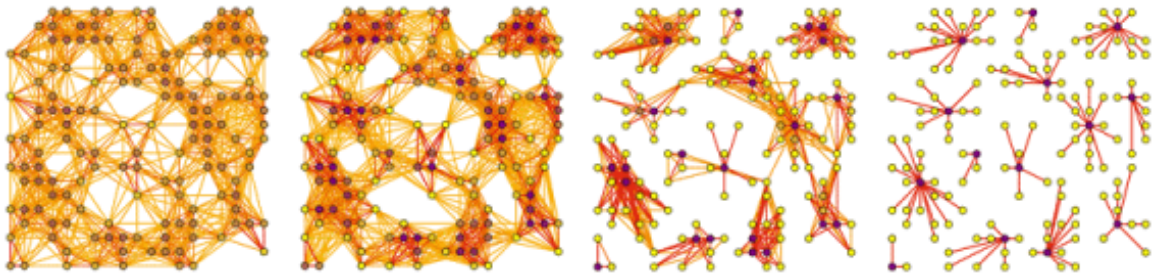
- Calculation of a distance matrix (e.g. BLAST E-values, frequency of peptides in an MS experiment)
  - Agglomerative : Start by pairing the “closest”, recalculate distances and elevate in the hierarchy.
  - Divisive : Start with a huge cluster with all the elements in and split going down in the hierarchy.
- Both approaches require the definition of a threshold in order to identify the clusters (both supervised).



Graph Clustering II

Markov clustering

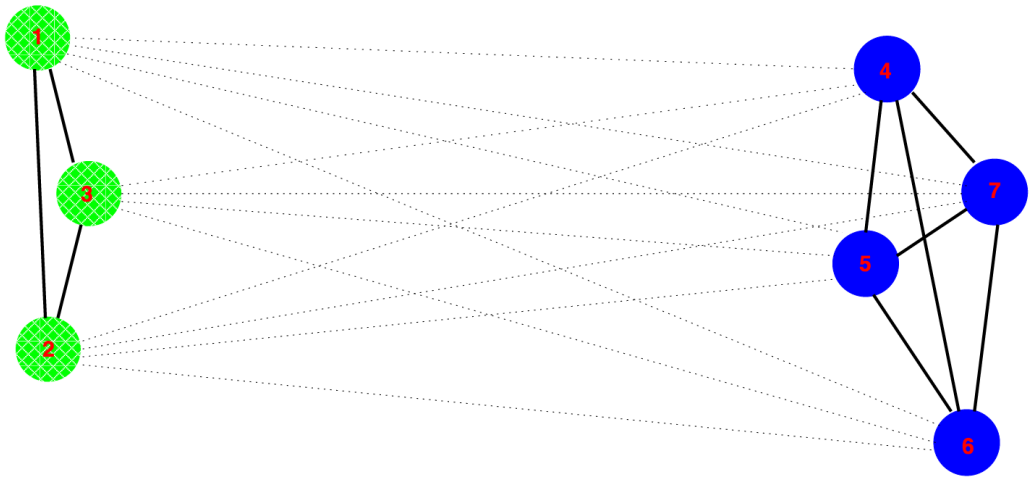
- Markov CLustering (MCL) : The algorithm is based in concept of “random walks”.
  1. The number of paths between two randomly chosen nodes is higher for nodes that belong to the same cluster.
  2. OR : The paths on a graph will less frequently traverse from one cluster to another.
  3. The algorithm uses only a single parameter r the inflation coefficient (for the Markov chain) and it does NOT require prior knowledge of the number of clusters.



Graph clustering II

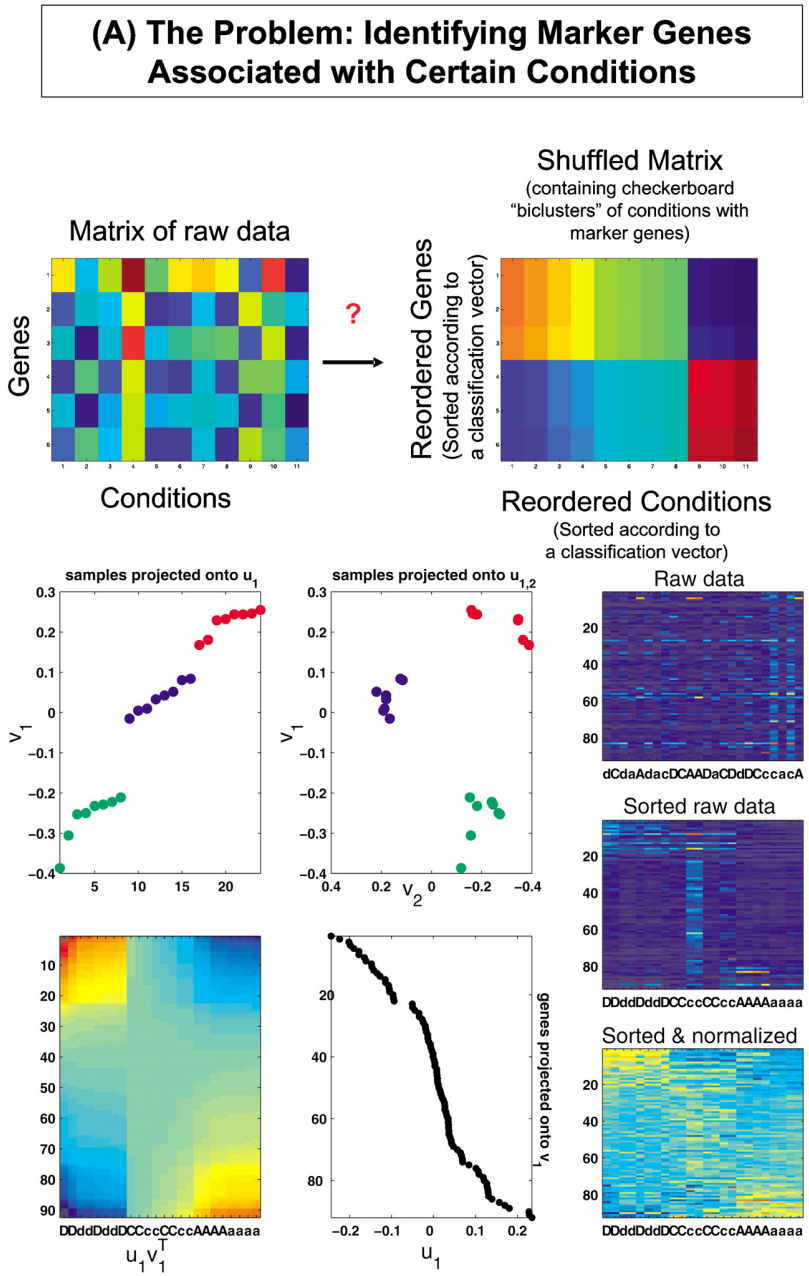
Spectral clustering

- Spectral Clustering : The algorithm is based on the static distribution that a number of particles will have as it diffuses in the network.
  1. In each time-point the network is populated by a specified number of particles (infinite)
  2. Each particle is moving from node to node with a probability *p* which is related to the similarity of the nodes.
  3. It has been solved analytically that a particle will spend more time on nodes which show similarity (and thus are members of clusters) than into others/.



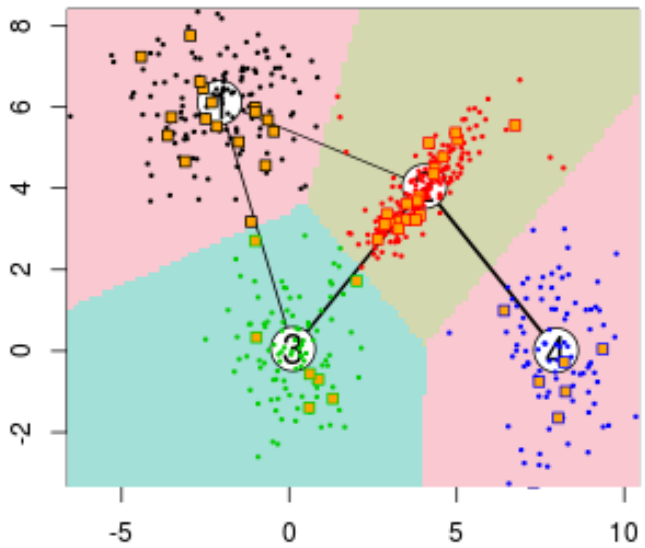
Graph-Data clustering – Biclustering

- **Simultaneous** clustering of genes and conditions.
- Can be solved as an eigenproblem.
- However analytical solutions are never possible.
- We employ heuristic criterion to find a suitable solution.
- The nature of the heuristic criterion defines the type of clusters we discover.



k-means clustering

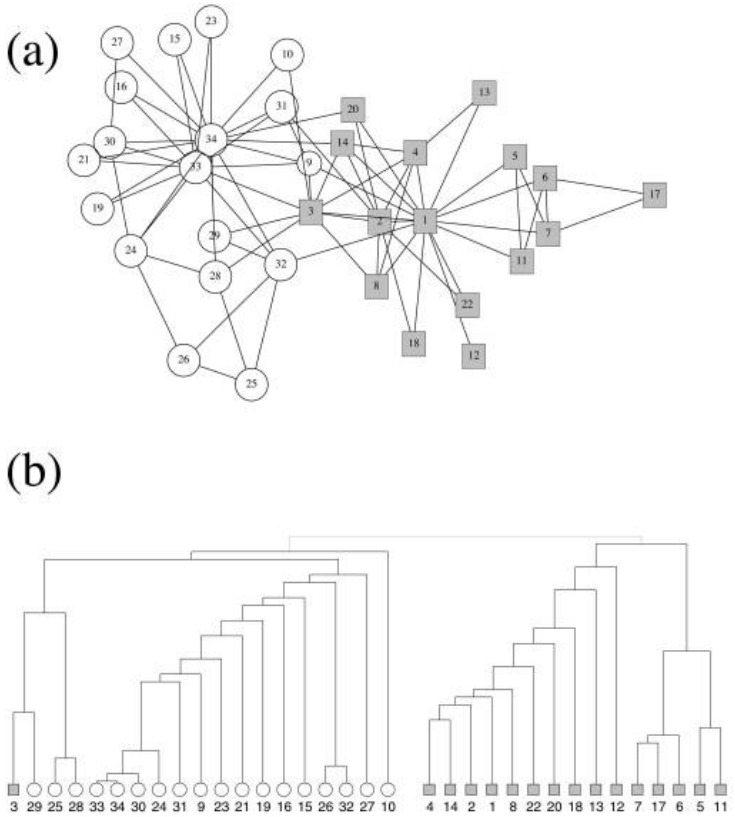
- *Supervised* clustering method (the number of clusters k must be specified).
- Assign randomly k-centroids.
- Calculate the sum of square distances from each centroid.
- Update the centroids in such a way that the SSM will minimise.
- Continue until convergence.
- Drawback : sensitivity to the initial selection of centroids.



Network module (community) detection

- Girvan-Newman : Hierarchical clustering algorithm for the detection of communities in complex systems.

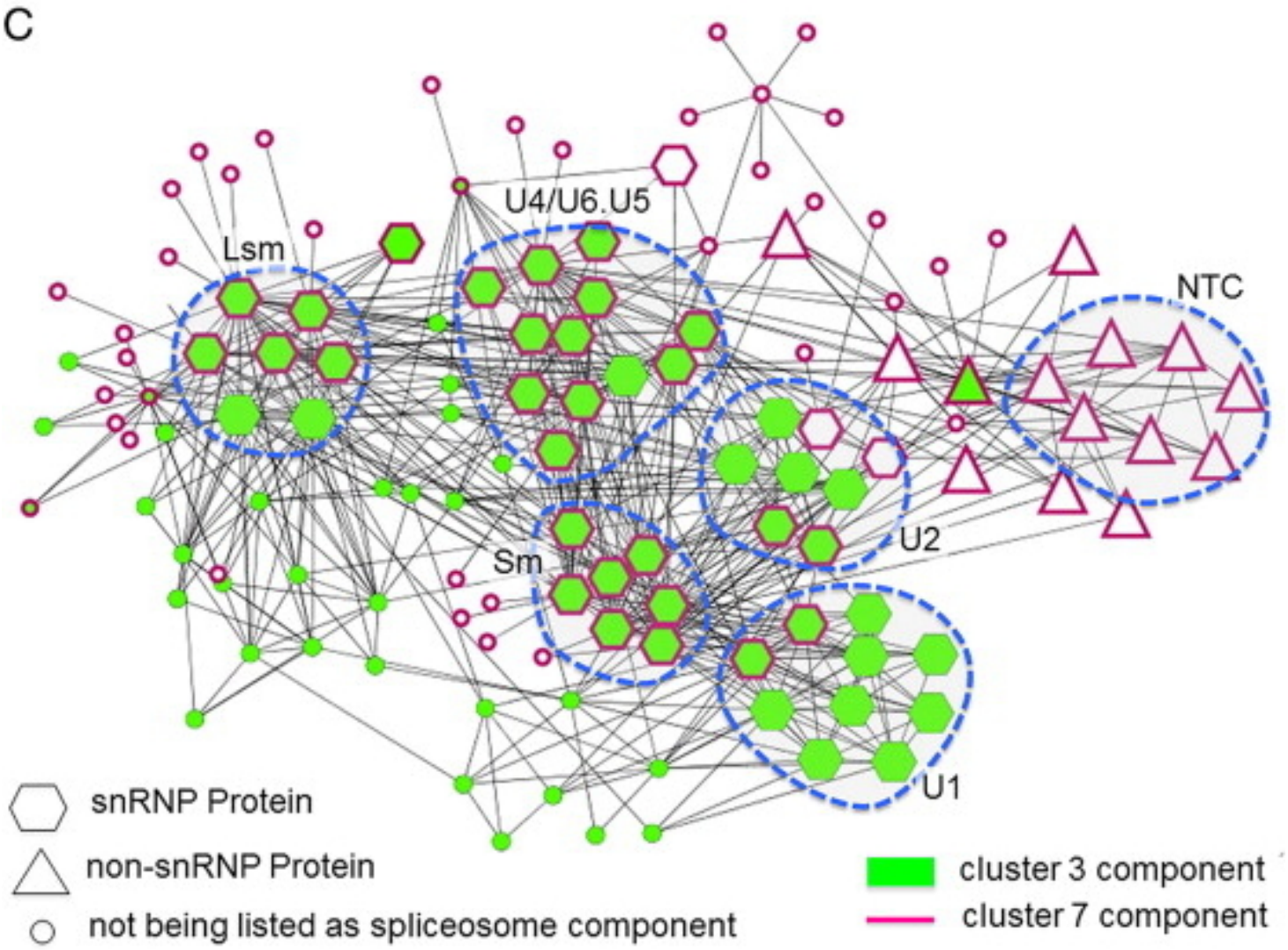
1. Calculating betweenness [1] of each edge.
2. Remove the edge with the highest betweenness
3. Re-calculate edge betweenness
4. Repeat steps 2 to 3 until there are no more edges on the network.



1. Betweenness centrality : Is defined as the number of shortest paths that passing from the edge.

Identify complexes from PPI networks

Integrate Network Clustering



Modelling biological networks

Background

- For which reasons we construct network models.
  - To understand the complexity of a system.
  - To construct a set of hypotheses which can be tested experimentally.

  Models $\rightarrow$ Dynamical Systems
- Basic elements of dynamical systems :
  - State : Describes the properties of the parts of the system in a given time.
  - Transition : Defines the way(s) by which a certain state moves to another.

Definition (Dynamical System)

A system which updates its state with relation to time.

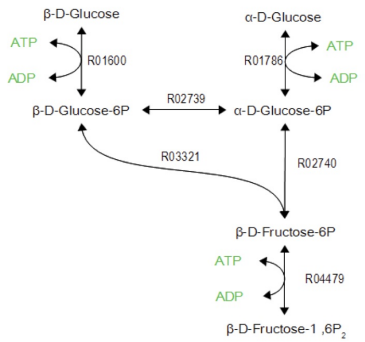
Categories of Network Models

- Metabolic modelling
- Discrete models
  - Discrete state space
    - Boolean networks
    - Logical models
  - Discrete time
    - Difference equations
    - Petri dishes
- Continuous models
  - Continuous time ODEs.
  - Continuous space PDEs.
- Probabilistic models
  - Bayesian Networks

Flux Balance Analysis

Method of analysis and simulation of metabolic network dynamics

- Stoichiometry coefficients of the system
- Matrix of stoichiometric coefficients
- Flux vector of reactants - products.
- System's solution at $A\mathrm{x} = 0$ with a single constrain (growth, sugar uptake etc.)



	R02739	R03321	R02740	R01600	R01786	R04779
C00002	0	0	0	-1	-1	-1
C00008	0	0	0	1	1	1
C00221	0	0	0	-1	0	0
C00267	0	0	0	0	-1	0
C00668	-1	0	-1	0	1	0
C01172	1	-1	0	1	0	0
C05345	0	1	1	0	0	-1
C05378	0	0	0	0	0	1

Modelling gene expression with boolean networks

- The expression state of a gene has only two discrete values ON or OFF (1 or 0).
- The expression state of a gene is updated by a boolean function
- The state of the network is updated in discrete time-steps.

Characteristics of Boolean networks

1. Fixed topology
2. Discrete and synchronised
3. The states and its updates are fully deterministic
4. Boolean functions affects their dynamics

Boolean networks dynamics

- A boolean network consists of :
  1. Our known graph $G = (V, E)$ which determines its topology with the addition of the set of node states $X = \{x_i|1, ..., n\}$.
  2. A set of Boolean function which determine the states' update $B = \{b_i|1, ..., k\}, b_1 : \{0, 1\} \rightarrow \{1, 0\}$
- Each node v_i is associated with a Boolean which takes variables from the connections of the node v_i,
- The state of the node v_i at time-step t is defined as $x_i(t)$, and at time-step $t + 1$ will be :

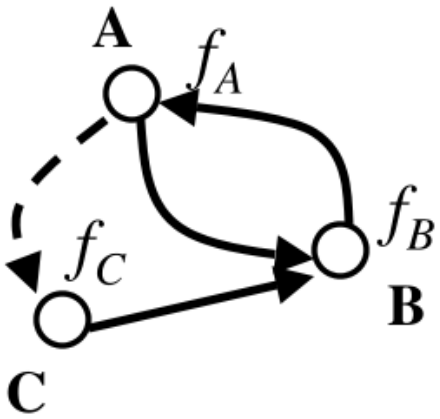
  $x_i(t + 1) = b_i(x_{i1}, x_{i2}, ..., x_{ik})$

  Where x_{ij} are all the states of k connected nodes to v_i.

Boolean networks

Truth tables – Dynamics

- Boolean network



$$f_A(B) = B$$

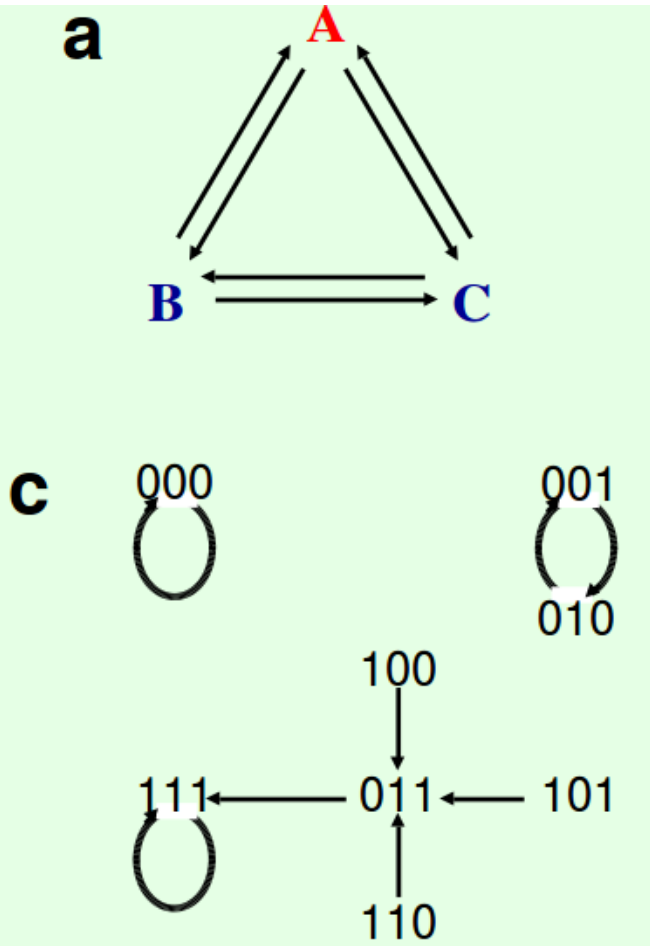
$$f_B(A,C) = A \text{ and } C$$

$$f_C(A) = \text{not } A$$

- Truth table

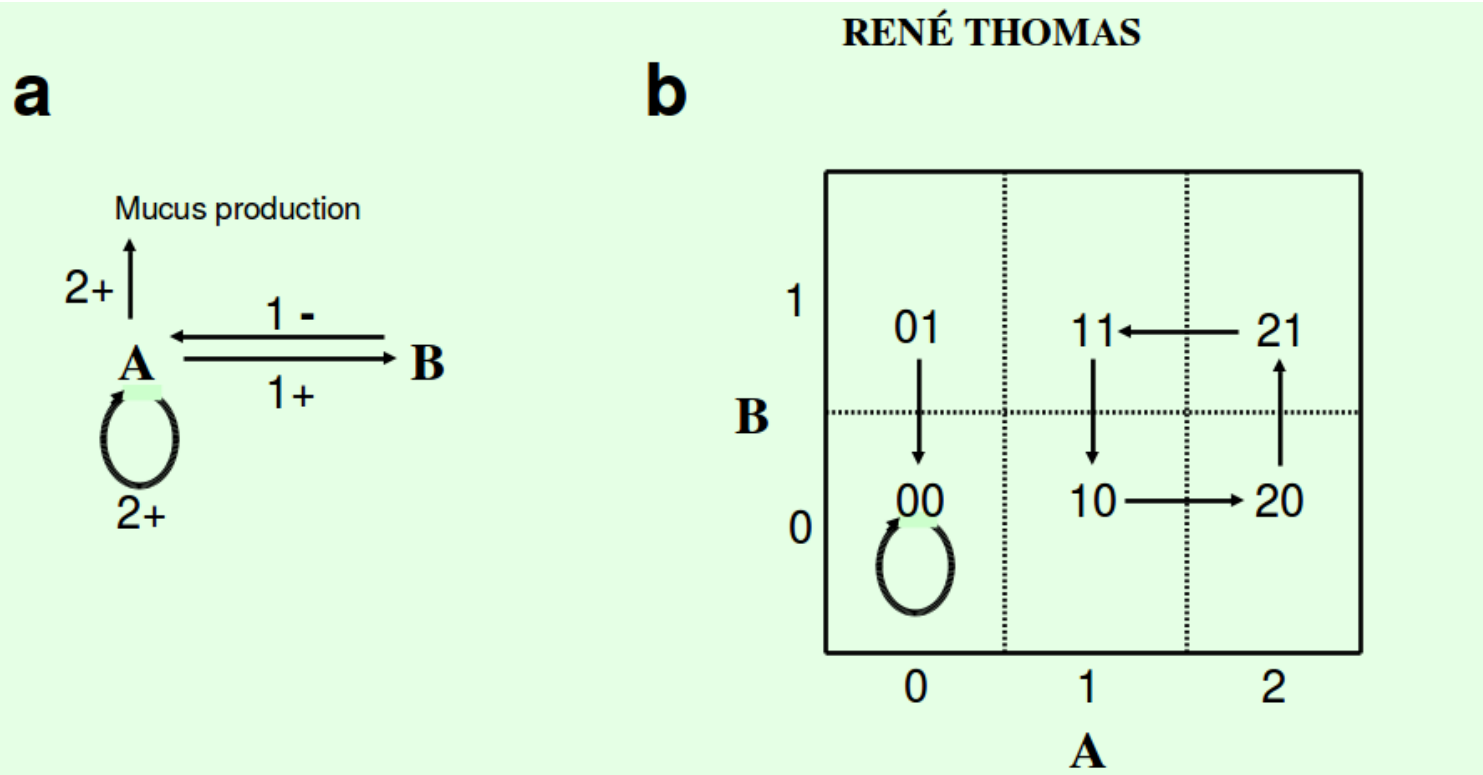
State	INPUT			OUTPUT		
	A	B	C	A'	B'	C'
1	0	0	0	0	0	1
2	0	0	1	0	0	1
3	0	1	0	1	0	1
4	0	1	1	1	0	1
5	1	0	0	0	0	0
6	1	0	1	0	1	0
7	1	1	0	1	0	0
8	1	1	1	1	1	0

Boolean networks : Example



A small boolean network. a) The wiring diagram in a boolean network with three genes (A, B and C), each an input to the other two. **b)** The boolean rules for the diagram shown in a), assuming that gene A represents an *AND* gate, while genes B and C each represent an *OR* gate. **c)** The state transition graph of the boolean network depicted in a) and b). Each triplet of digits correspond to a state for genes A-C, from left to right.

Logical Models René Thomas



A small logical network "à la René Thomas". a) The regulatory interactions for mucus production in the opportunistic pathogen *Pseudomonas aeruginosa*. Two genes, encoding an activator (A) and an inhibitor (B) of mucus production are considered. Each edge in the graph is labeled with the rank number of the threshold, followed by the sign of its regulatory influence (-, inhibition; +, activation). Given parameters (not shown here), a dynamics may be deduced. **b)** The asynchronous state graph. This graph is one among several graphs that would fulfill the constraints based on biological knowledge or hypotheses. A can take any value among {0, 1, 2}, and B among {0, 1}. Thresholds are represented by dashed lines, and transitions by arrows. The graph shows two steady states, one for A = 0, and one cycle: 11 -> 10 -> 20 -> 21 -> 11.

- Logical models appeared as extension of the Boolean framework.

Modelling gene expression

Continuous models ODEs

- The concentrations of moleculs (,RNAs, proteins etc.) are represented by a continuous variable $x_i(t) \in R_{\geq 0}$
- The rate of expression of a gene is analogous to the concentration of its product

$$\frac{d_x}{d_t} = kx$$

- If we solve this differential equation for x we get :

$$x = x_o e^{kt}$$

Modelling gene expression

Continuous models ODEs

- The concentrations of moleculs (,RNAs, proteins etc.) are represented by a continuous variable $x_i(t) \in R_{\geq 0}$
- The rate of expression of a gene is analogous to the concentration of its product

$$\frac{d_x}{d_t} = kx$$

- If we solve this differential equation for x we get :

$$x = x_o e^{kt}$$

Differential Equations

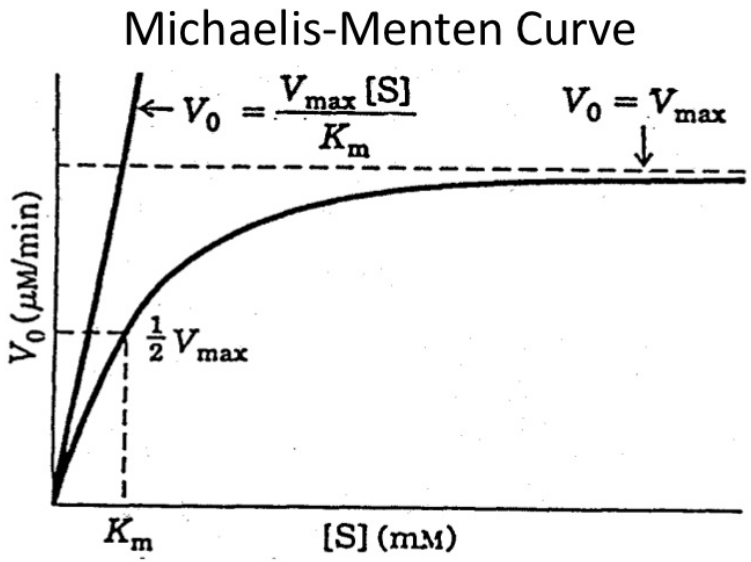
Michaelis-Menten kinetics

- The rate of change $\frac{d_x}{d_t}$ of an mRNA/protein reaches saturation.
- X concentration of mRNA/protein
- γ_x degradation rate of mRNA/protein

$$\frac{dx}{dt} = \frac{V_{max} X}{K_M + X}$$

at equilibrium :

$$X \propto \frac{V_{max}}{\gamma}$$



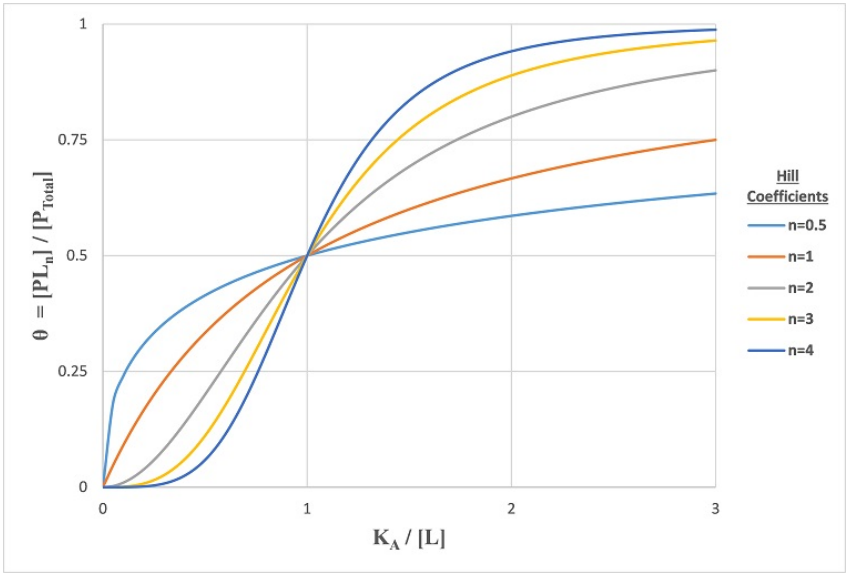
Hills kinetics

Similarities with the discrete models

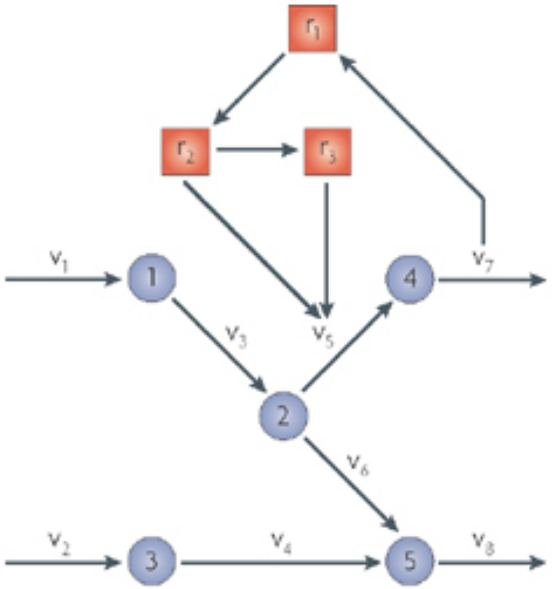
- Hills dynamics more complex as they encompass phenomena of cooperation/competition between regulatory factors.
- Equation similar with the typical M-M, however the addition of Hills exponent n affects the dynamics.

$$\frac{dx}{dt} = \frac{V_{max} X^n}{K_M + X^n}$$

- For $n > 6$ dynamics is converging qualitatively the ones of discrete systems.



Regulated flux balance analysis



Trajectory

r_1	r_2	r_3	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
1	0	1	0.1	0.2	0.1	0.2	0	0.1	0	0.3
0	0	1	0.1	0.2	0.1	0.2	0	0.1	0	0.3
0	1	1	0.2	0.2	0.2	0.2	0.2	0	0.2	0.2
1	1	0	0.2	0.2	0.2	0.2	0.2	0	0.2	0.2
1	0	0	0.2	0.2	0.2	0.2	0.2	0	0.2	0.2
1	0	1	0.1	0.2	0.1	0.2	0	0.1	0	0.3

Regulation functions

$f_{r1}(V_7)$

v_7	r_1
= 0	1
≥ 1	0

$f_{r2}(r_1)$

r_1	r_2
0	1
1	0

$f_{r3}(r_2)$

r_2	r_3
0	1
1	0

$f_{v5}(r_2,r_3)$

r_2	r_3	v_5
0	0	≥ 0
0	1	= 0
1	0	≥ 0
1	1	≥ 0

Stoichiometric matix

1	0	–1	0	0	0	0	0
0	0	1	0	–1	–1	0	0
0	1	0	–1	0	0	0	0
0	0	0	0	1	0	–1	0
0	0	0	1	0	1	0	–1

Constraints

$0 \le v_1, v_2 \le 0.2$

$0 \le v_3, v_4, v_5, v_6 \le 0.4$

$0 \le v_7, v_8 \le 0.3$

Objective function

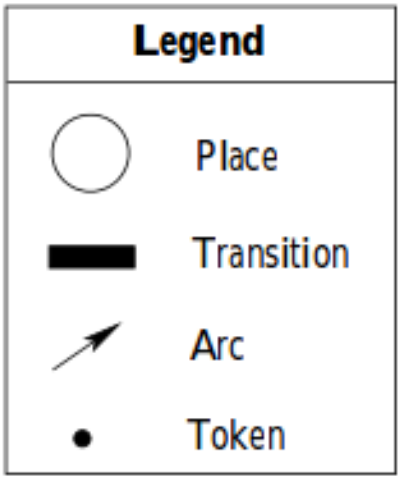
$v_7 + v_8$

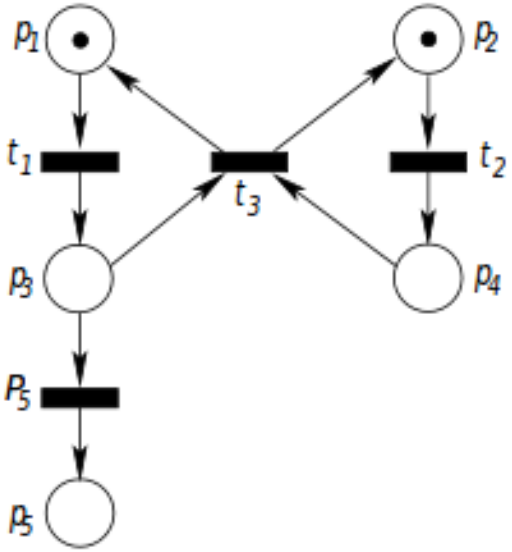
Nature Reviews | Molecular Cell Biology

Provides a complete framework to model the crosstalk between metabolism and gene regulation.

Petri nets models

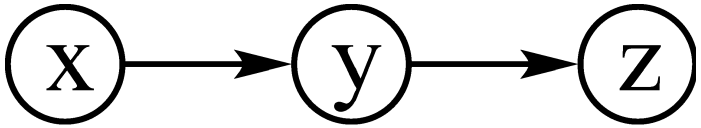
- Petri net :
  1. Modelling method originated from parallel and distributed systems.
  2. Is a bipartite directed graph.
  3. *Parts, transitions* and *states* are the static elements of the graph representing the structure of the system.
  4. *Tokens* represent its dynamical structure :
     The state of the system is determined by the transitions of tokens between parts and their final distribution.



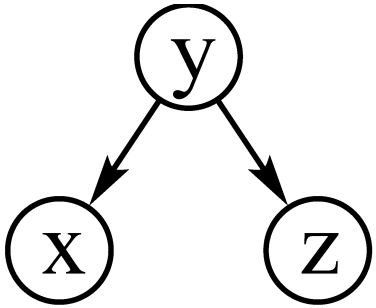


Bayesian network models

- Bayesian network :
  1. A Directed Acyclic Graph (DAG).
  2. Nodes contain random variables which represent concentrations or expression levels.
  3. The conditional probability of every node represents the state of the system (probabilist model)
  4. The joint probability function represents the state of the whole system.



$$p(x,y,z) = p(x)p(y|x)p(z|y)$$
$$p(z|x) = \sum_y p(y|x)p(z|y)$$

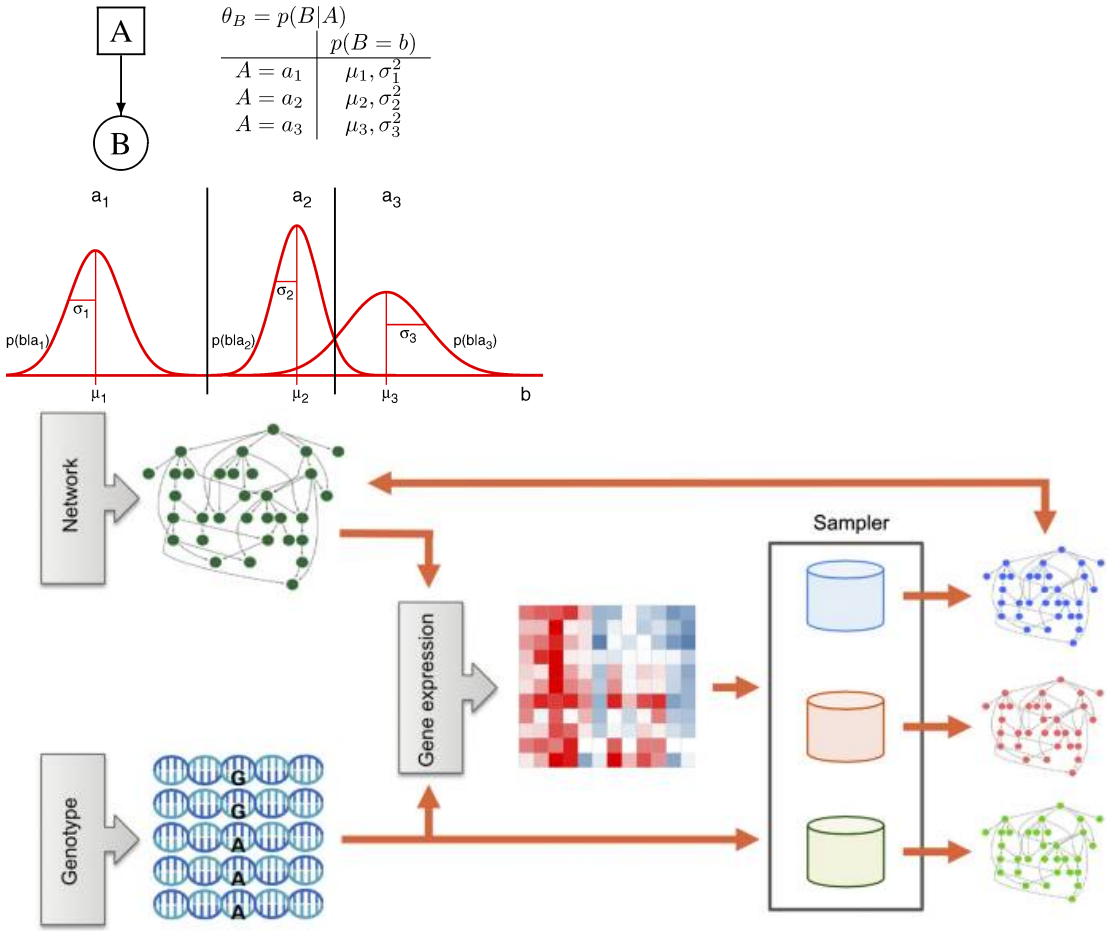


$$p(x,y,z) = p(y)p(x|y)p(z|y)$$
$$p(z|x) = \sum_y p(y)p(x|y)p(z|y)$$

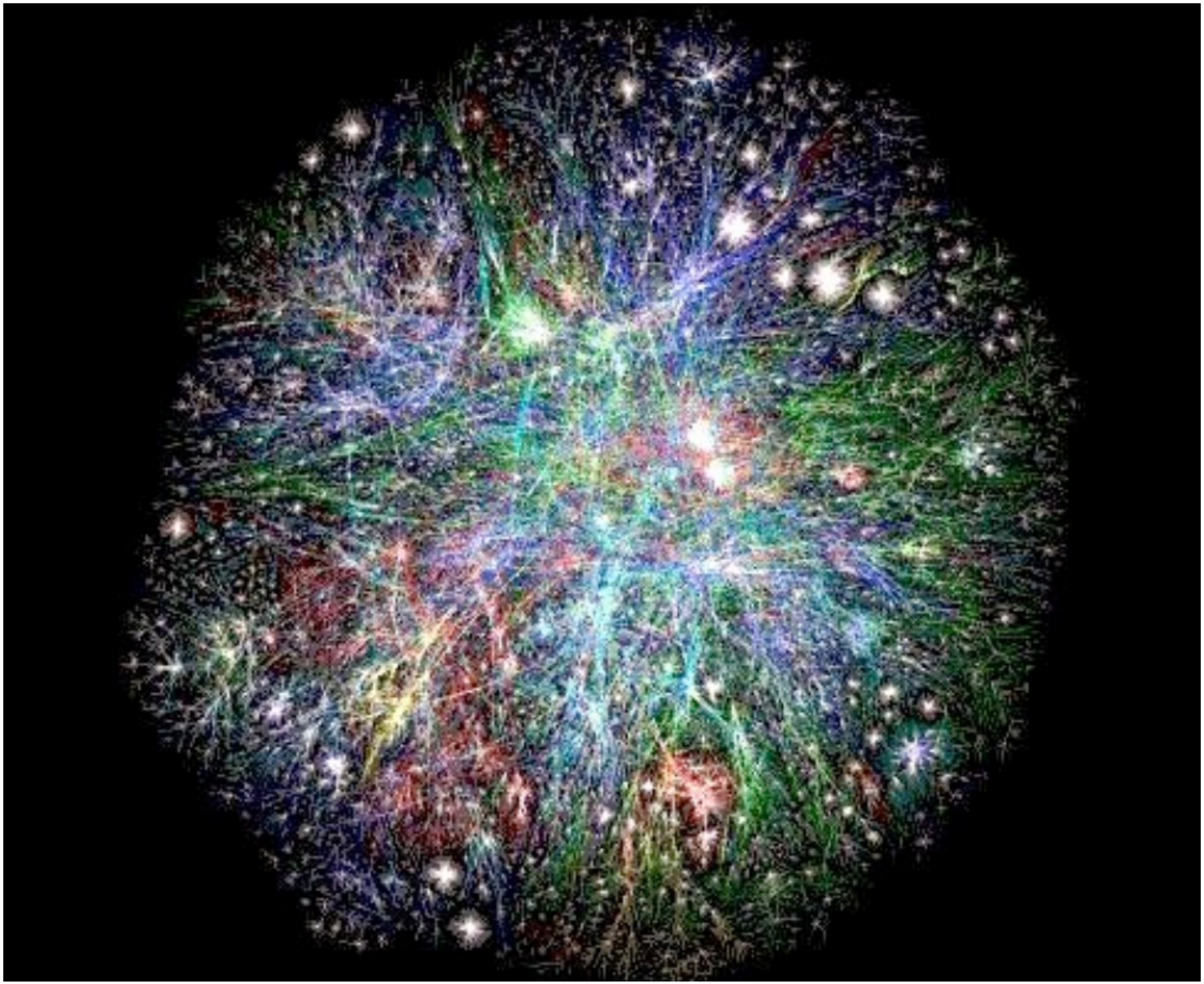
Characteristics of Bayesian networks

- We do not need to know the exact values of the parameters of the system (actually perhaps they do not even exist).
- We rather define a probability density around an interval of accepted values.
- Avoid over-fitting, a very frequent and typical problem of studies when the number of parameters exceeds the number of available samples.
- It is a statistical framework which matches the “statistical”/probabilistic nature of biological systems.
- Constrains : DAG not always representative, difficult to implement dynamics (however there exist dynamic-Bayesian networks)

Bayesian networks – Applications



- The *PDFs* of gene expression are the parameters of the Bayesian model.
- Reverse engineer to infer the structure and parameters of the network.



- Exercise with Cytoscape : From here, for a comprehensive tutorial click here