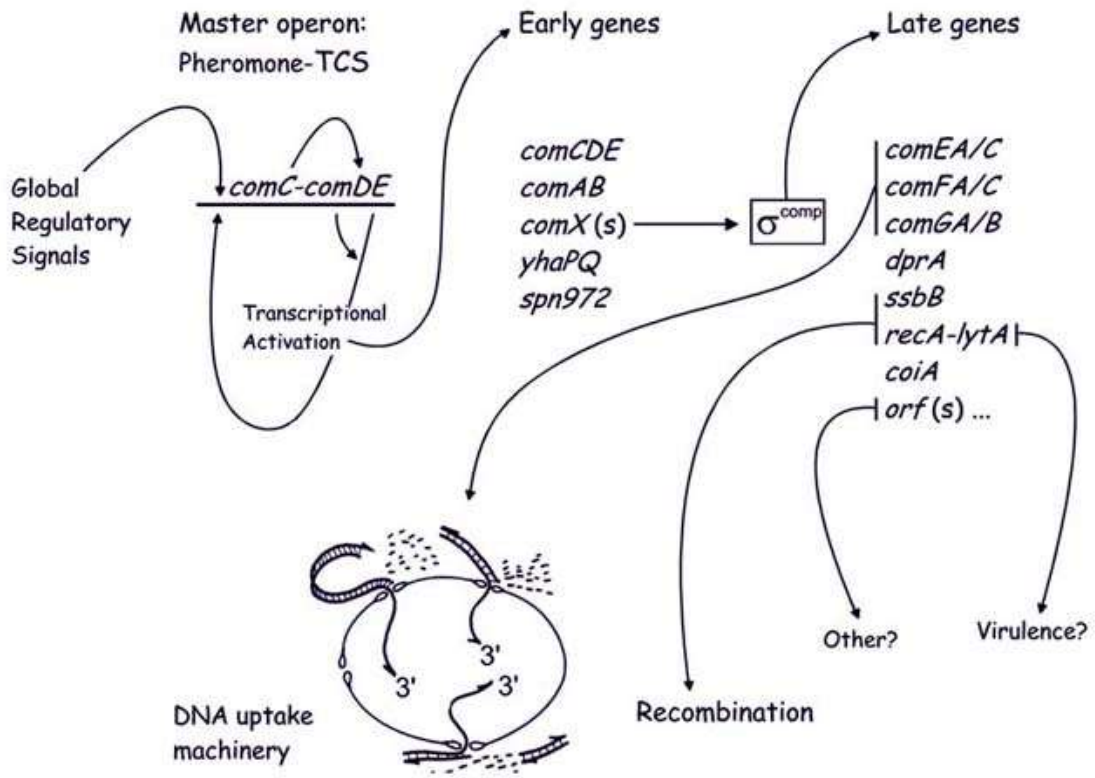
Réseaux de régulation

Master 1 MABS

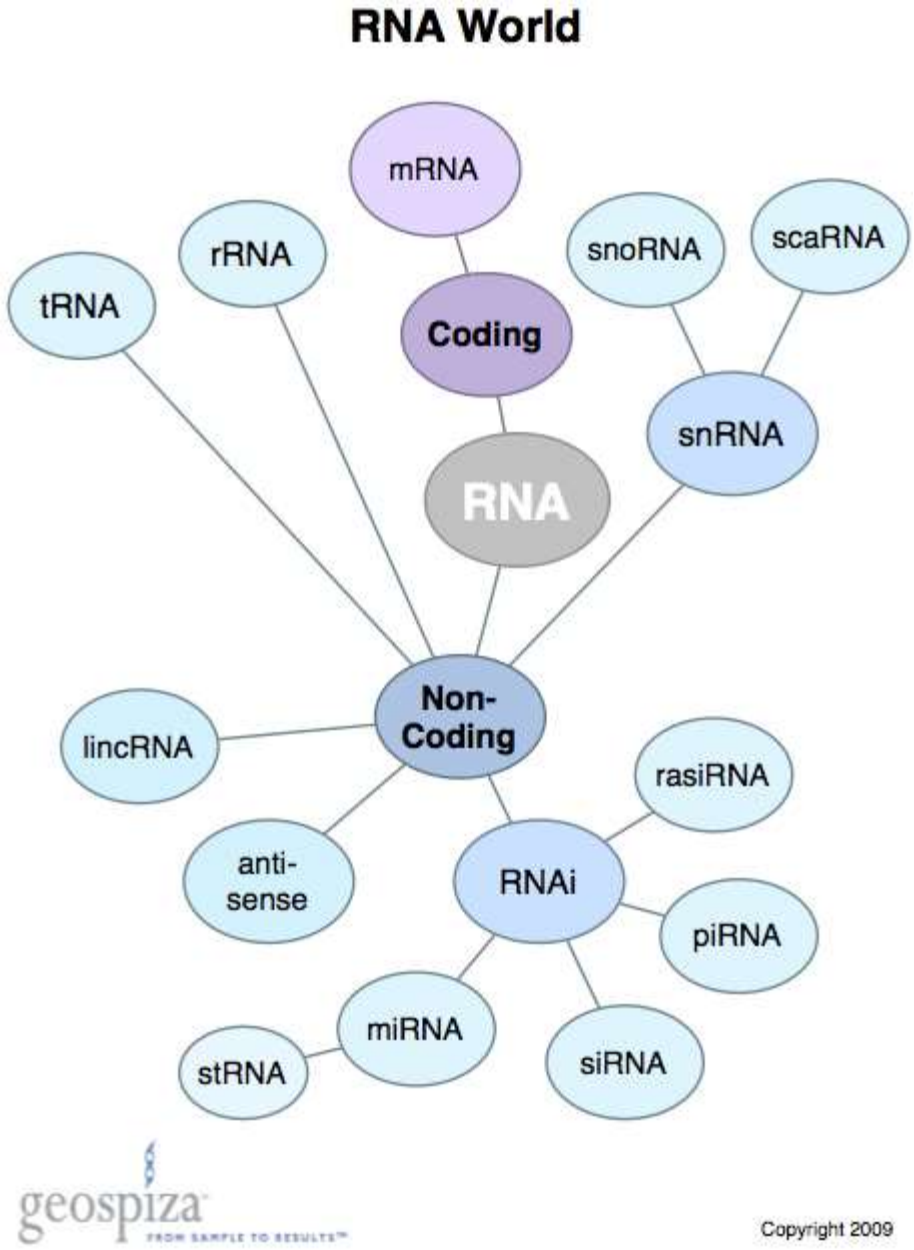


source : Claverys and Havarstein 2002

Exression du génome et relation génotype-phénotype

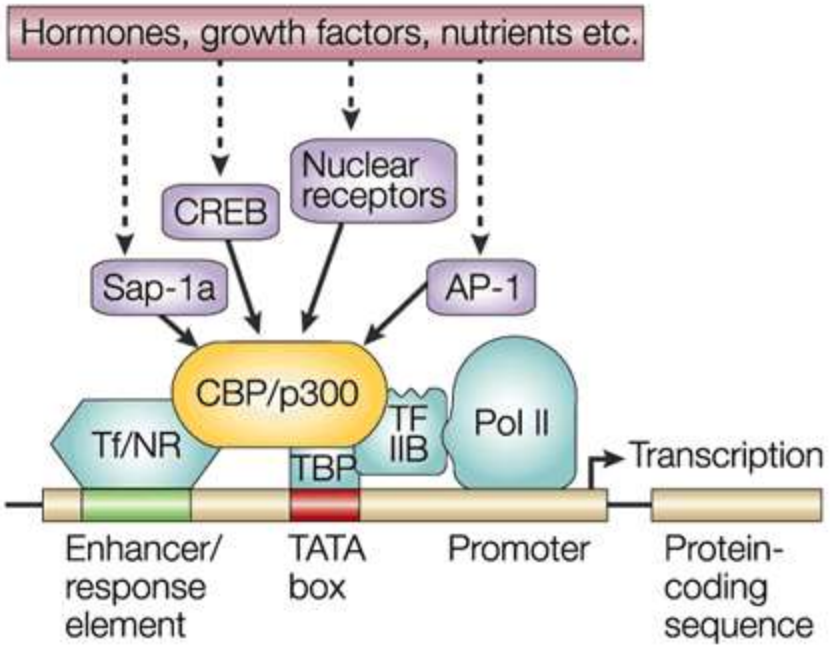
- Compréhension de propriétés du vivant conférées par les modalités d'expression du génome
- Expression
  - ARN codant et non codant
- Relation génotype/phénotype
  - et épigénétique : étude des changements, héritables au cours des divisions cellulaires, qui affectent la fonction des gènomes sans altération de la séqence ADN.
- Phénomènes complexes :
  - différenciation
  - développement
  - vieillissement
  - adaptation
  - effet de l'environnement

ARN



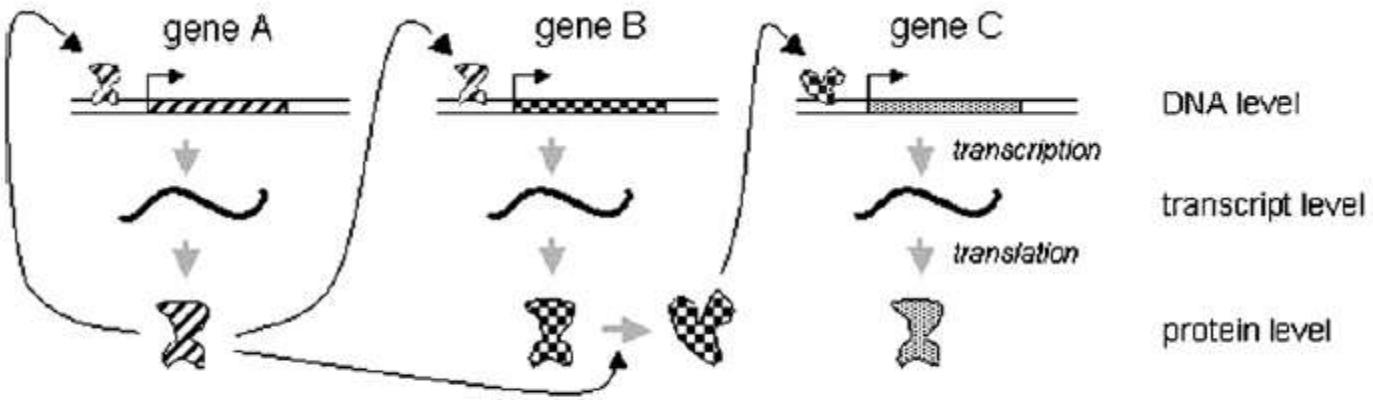
source : http://finchtalk.geospiza.com/2009/05/small-rnas-get-smaller.html

Facteurs de transcription

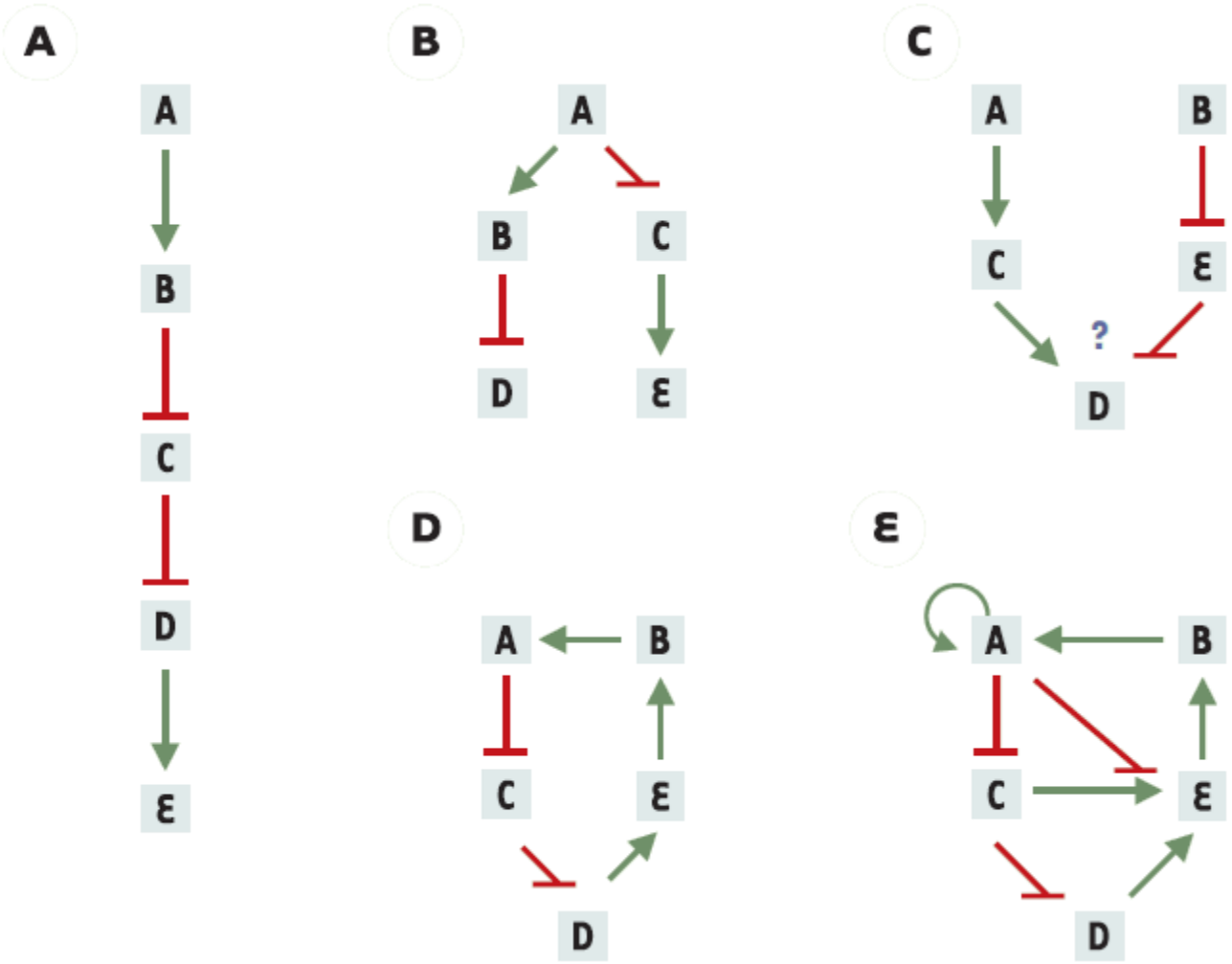


Nature Reviews | Molecular Cell Biology

Représentations et modèles

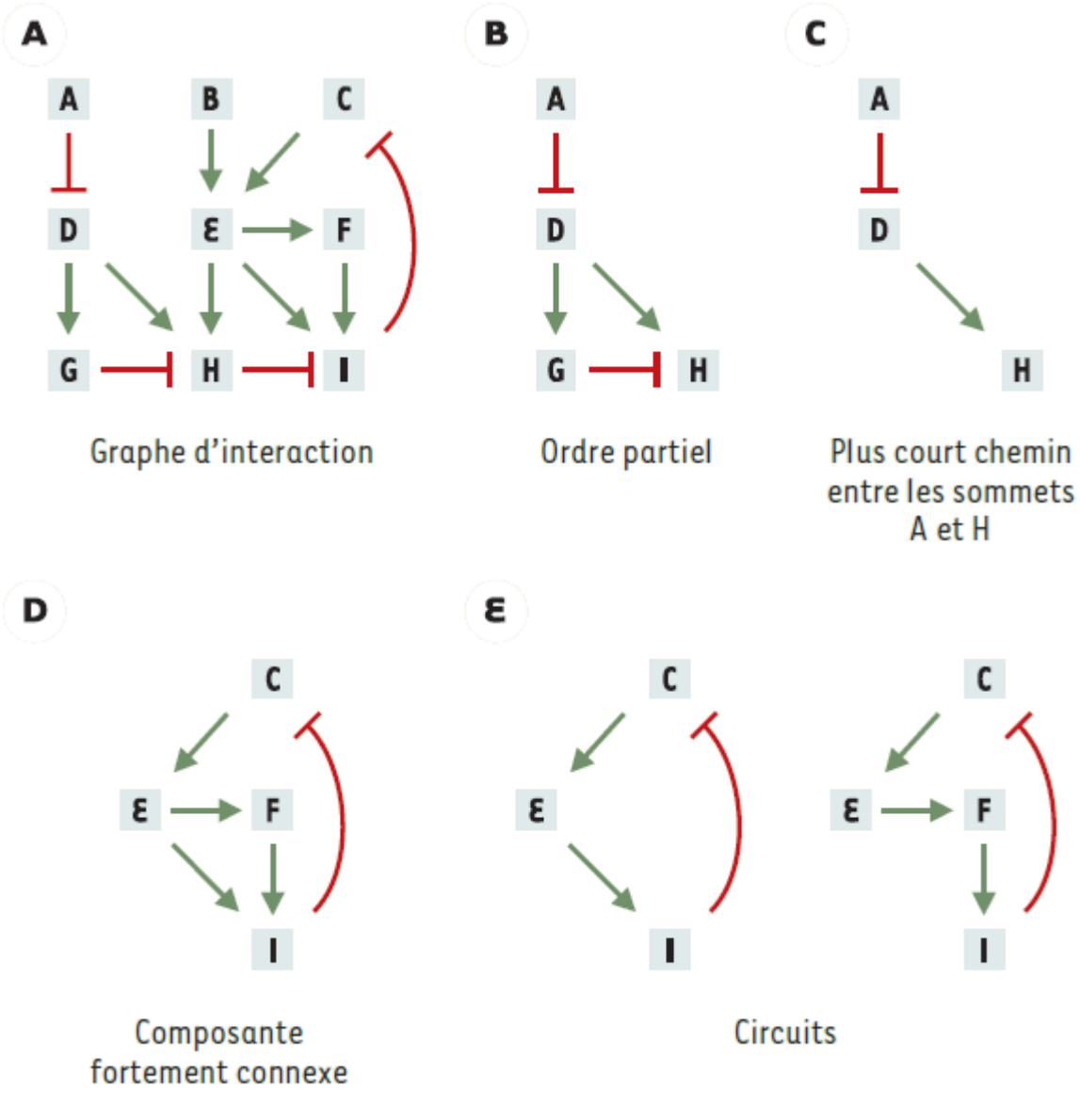


Représentations et modèles



source : De Jong et Thieffry 2002

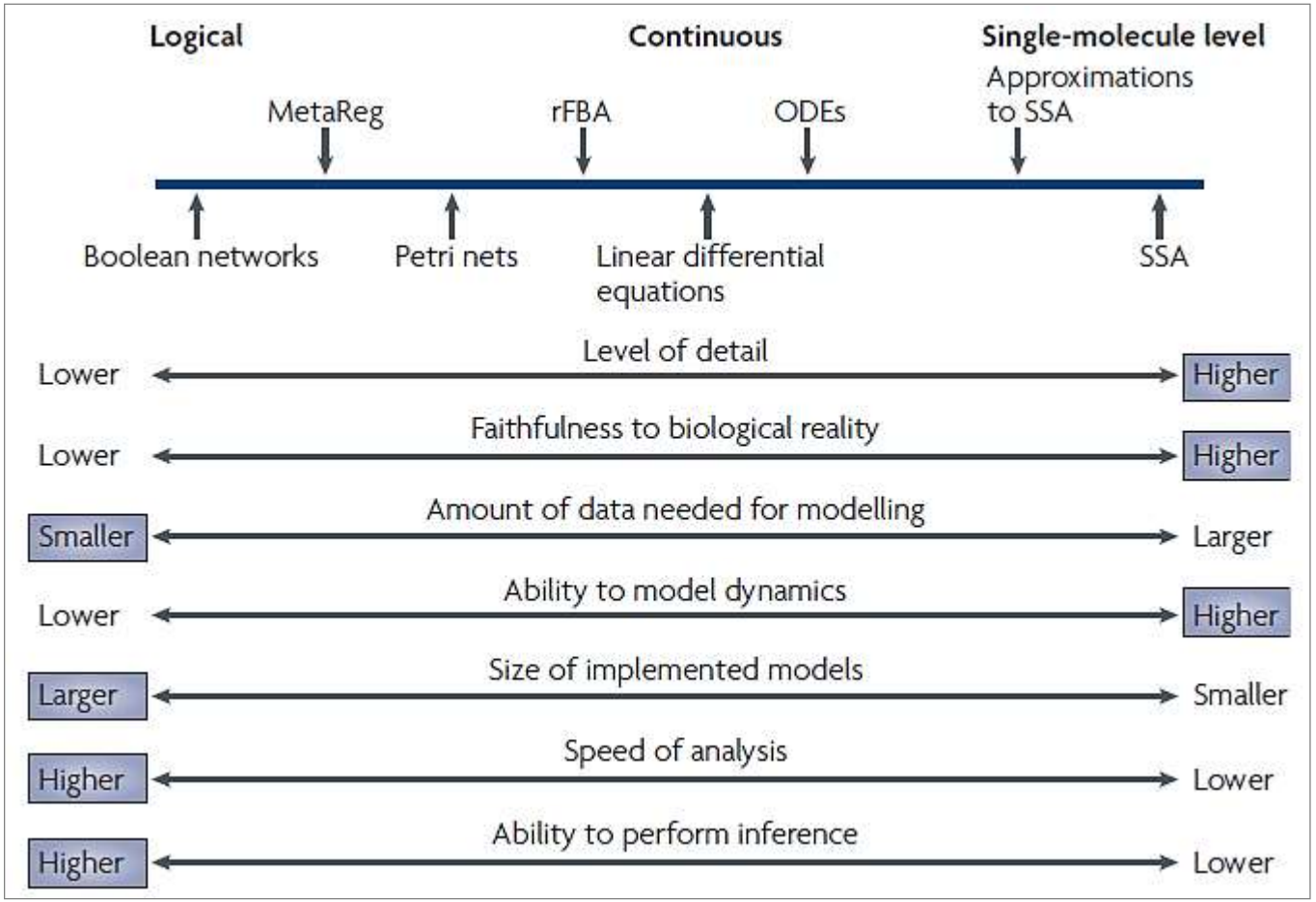
Décomposition en modules



source : De Jong et Thieffry 2002

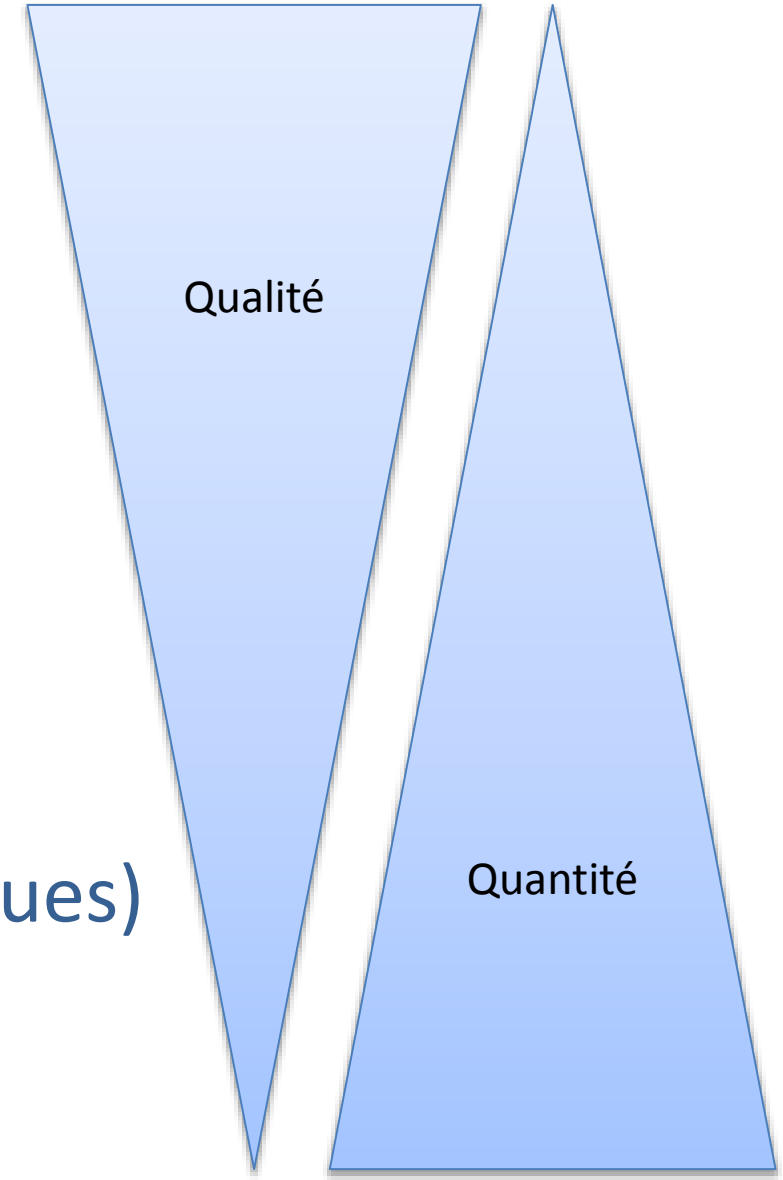
Données et connaissances utilisées pour la reconstruction

- Données d'expression
  - perturbations sur le phénomène d'intérêt
  - ensemble de conditions (>60) capturant un maximum de variations
- Données ChIP et ChIP-seq
  - localisation des sites de fixation des facteurs de transcription
- Facteurs de transcription connus
  - ex : RegulonDB, TransFac, JASPAR
- Prédiction de régions régulatrices
  - co-expression observée puis recherche des promoteurs
- Littérature et annotations

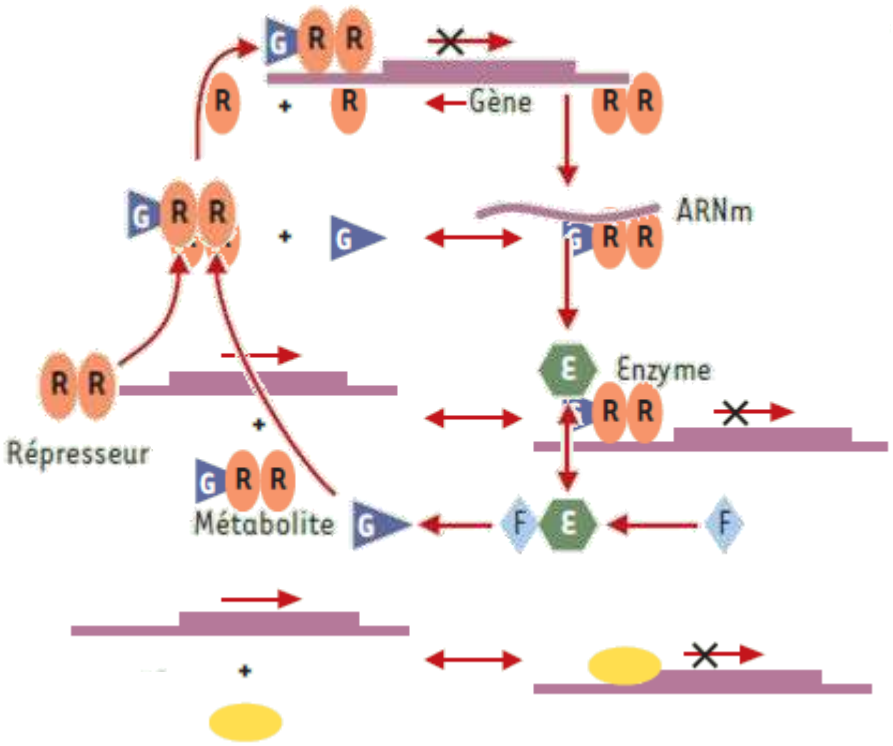


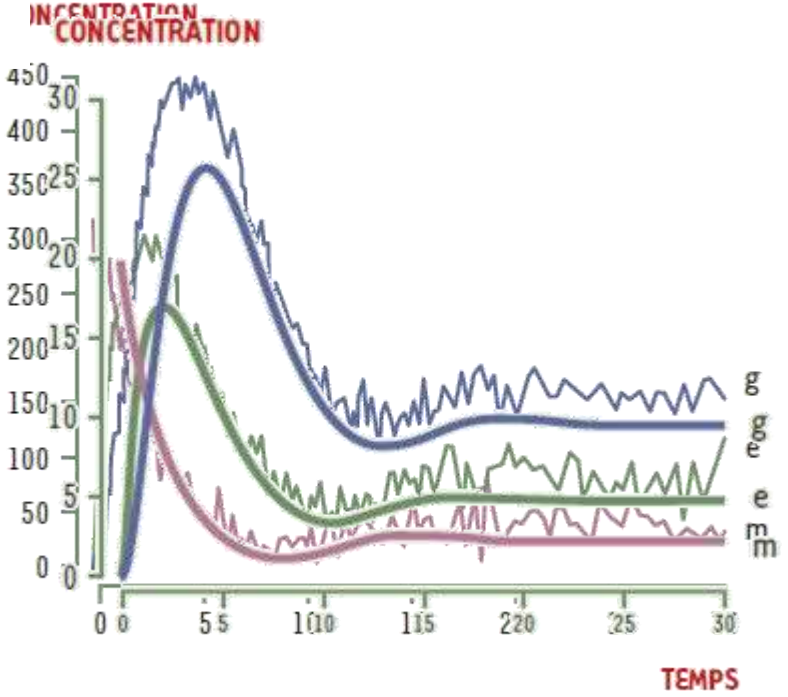
Karlebach and Shamir, 2008 *Nature reviews*

- Simulations stochastiques
- Systèmes d’équations
- Réseau de petri
- Réseaux booléens
- Réseaux bayésiens (dynamiques)
- Réseaux d’influence



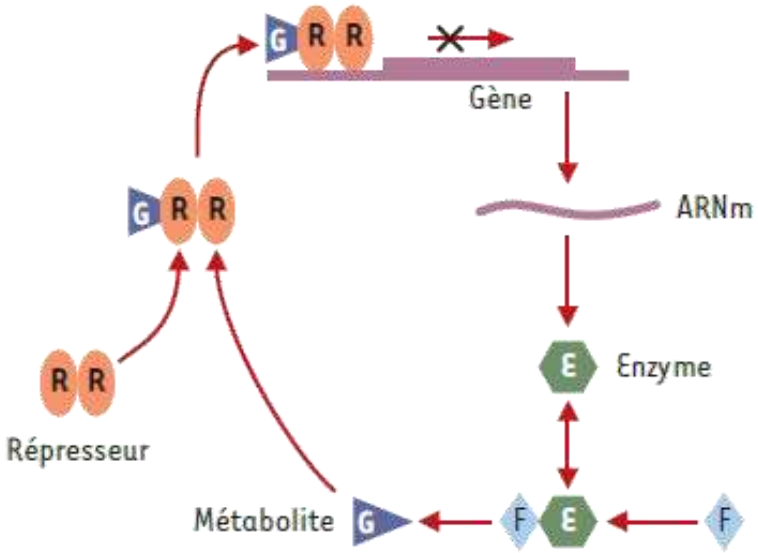
- Concentrations très faibles
- Probabilités associées aux réactions/transitions
- En général, trop complexe pour une solution analytique → simulations

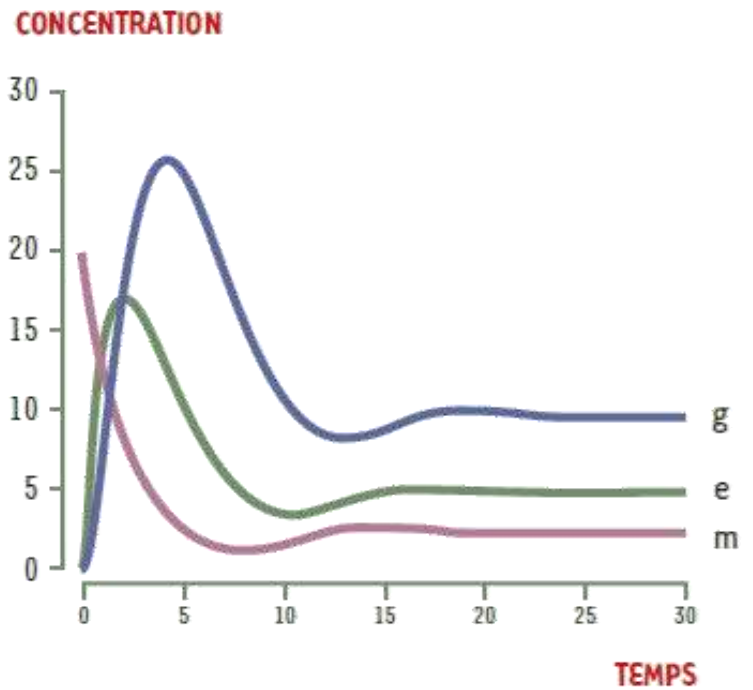




source : De Jong et Thieffry 2002

- Concentrations/activités des molécules
- Vitesse de production, dégradation
- Constante de seuil, association, dissociation, coopération

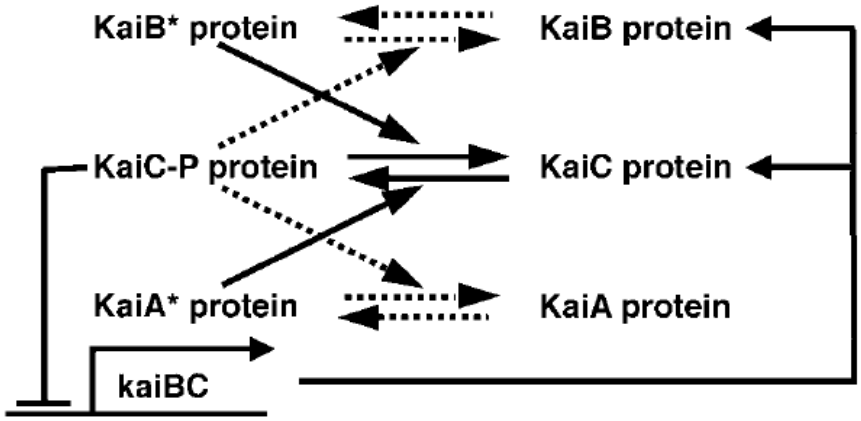




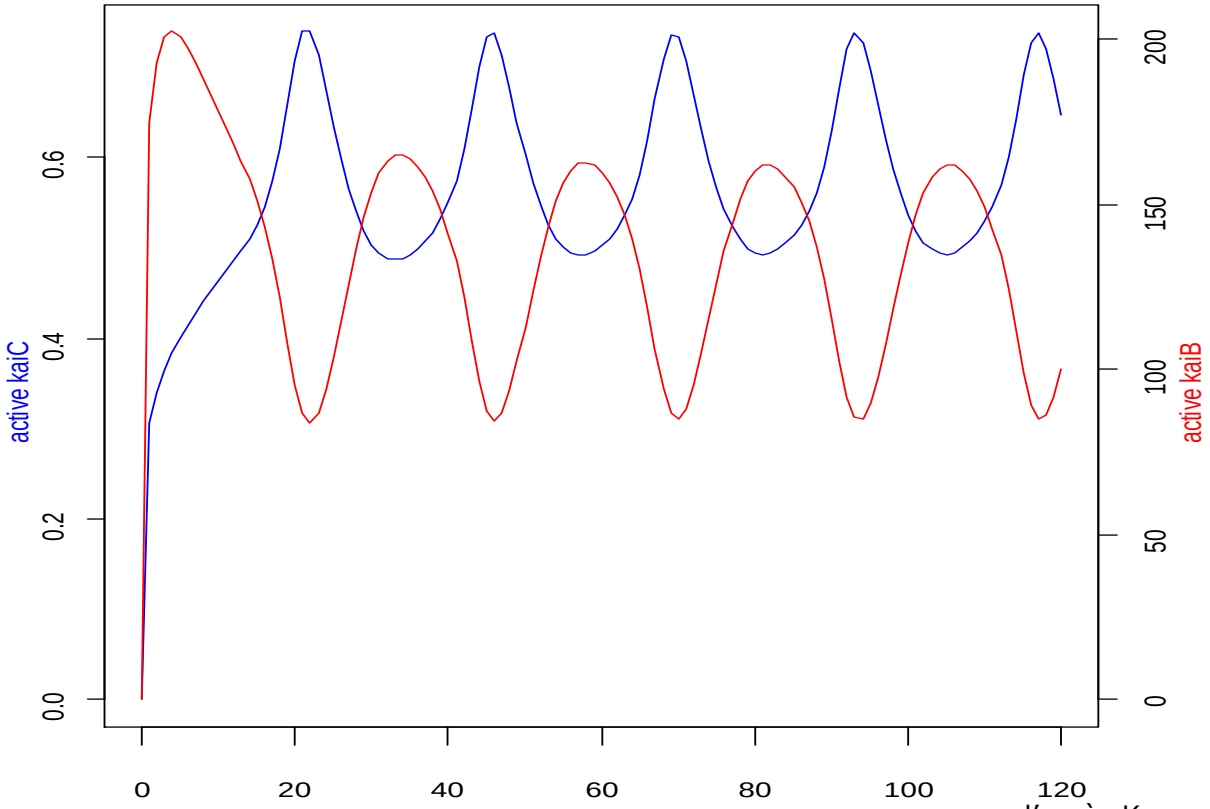
$$\frac{dm}{dt} = K_1 \frac{\theta^n}{\theta^n + g^n} - \gamma_1 m \qquad \frac{de}{dt} = K_1 m - \gamma_2 e \qquad \frac{dg}{dt} = K_3 e - \gamma_3 g$$

source : De Jong et Thieffry 2002

Exemple : oscillations circadiennes chez les cyanobactéries

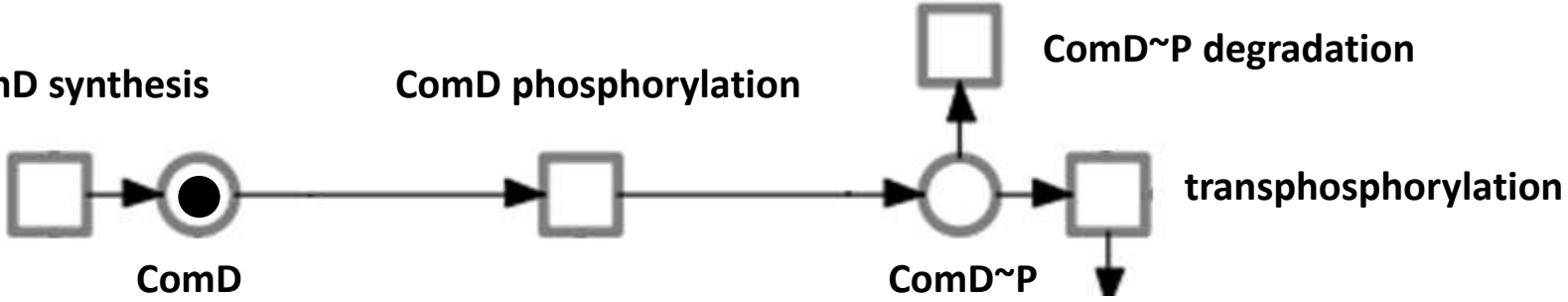


$$dx/dt = pa(C_0 s - x) - bx(y + f),$$
$$dy/dt = g(B_0 s - y) - k_1 y x^n/(q^n + x^n).$$
$$dB/dt = \varepsilon_1\{B_0\lambda/(1 + h_1 x^m) - \mu B\},$$
$$dC/dt = \varepsilon_1\{C_0\lambda/(1 + h_1 x^m) - \mu C\}.$$

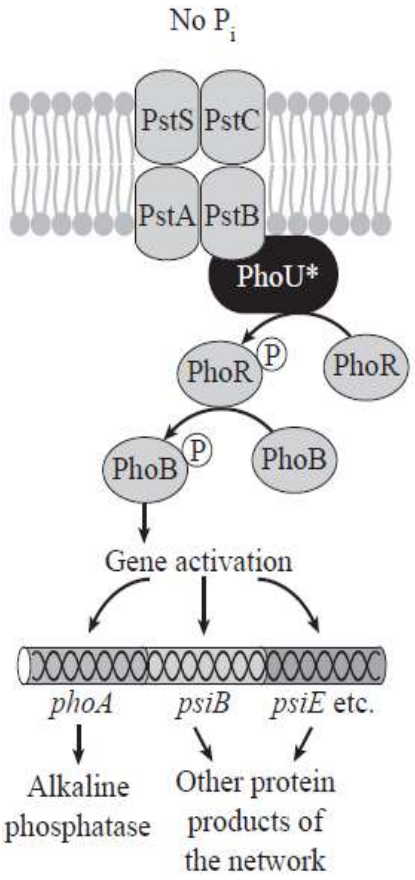


d'après Kurosawa *et al.* 2006

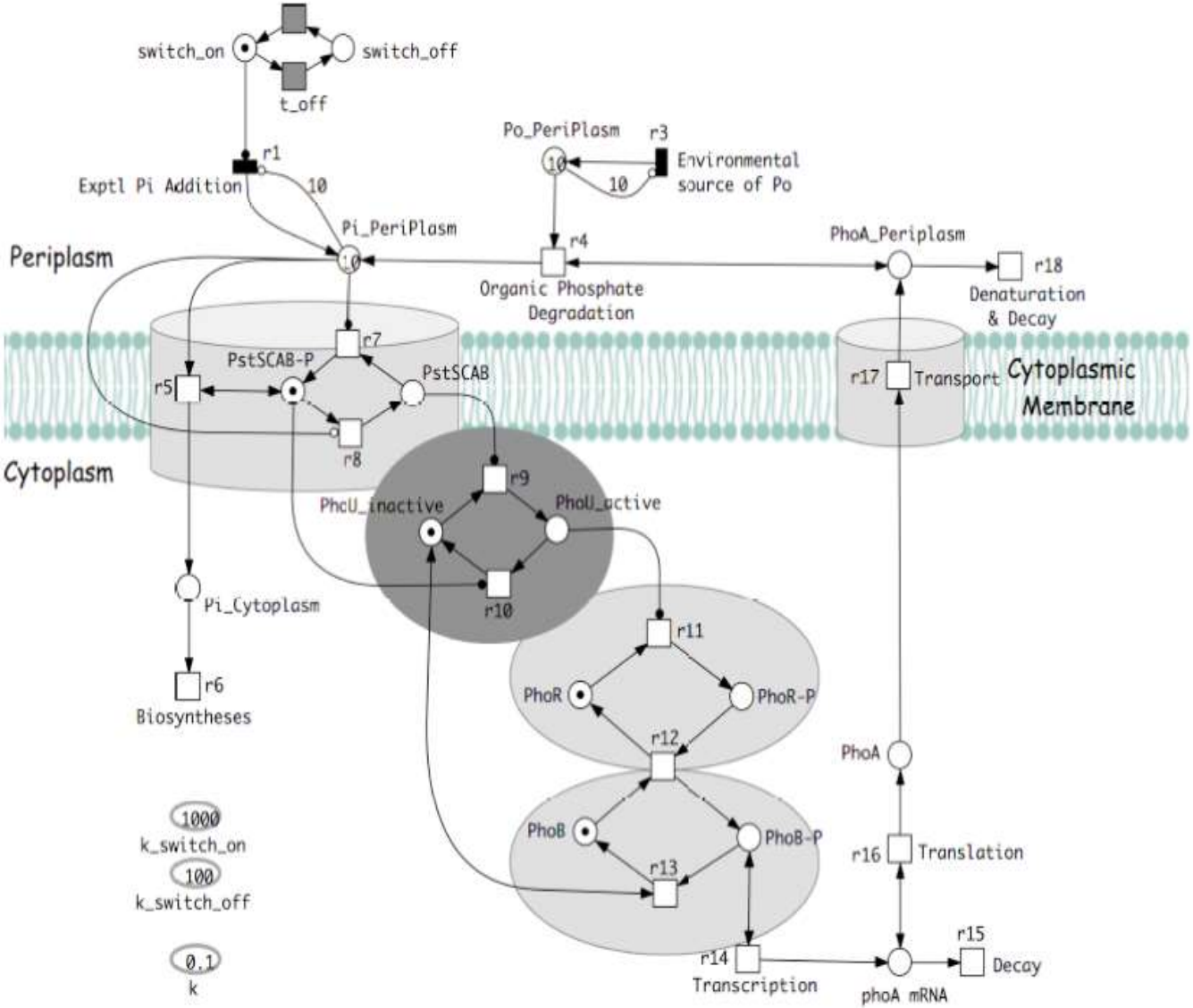
- Graphe biparti
  - Places : composés ou espèces moléculaires
  - Transitions : réactions
  - Jetons
- Différentes classes : colorés, stochastiques
- Permet
  - Analyse, propriétés du réseau
  - Simulations, prédiction



Réseau de petri



Neidhardt *et al.* 199(



Durzinsky *et al.*, 2011

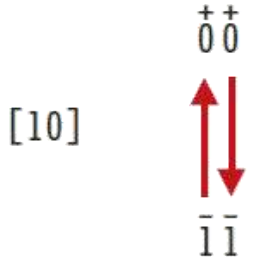
Synchrone

$$\begin{cases} x_{t+1} = \overline{y_t} \\ y_{t+1} = \overline{x_t} \end{cases}$$

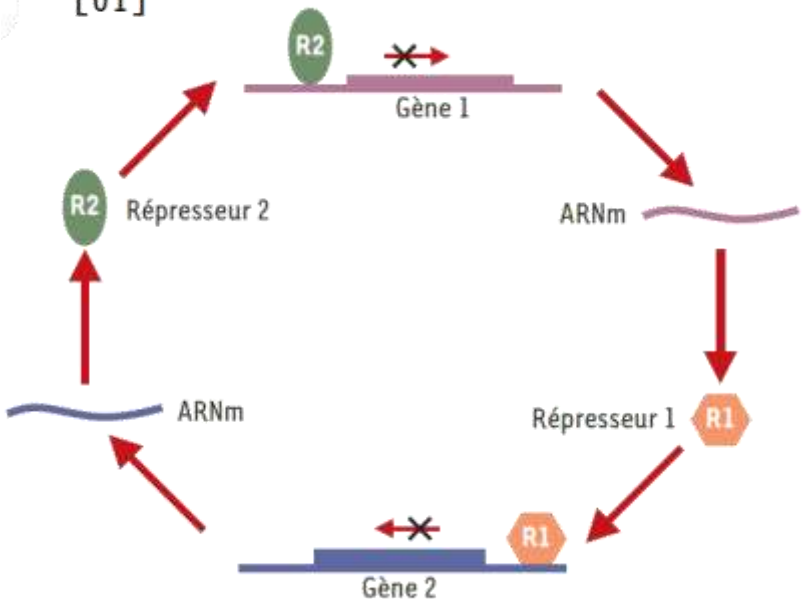
équations

$(xy)_t$	$(xy)_{t+1}$
$\overset{+}{0}\overset{+}{0}$	11
[01]	01
[10]	10
$\overline{1}\,\overline{1}$	00

table des états



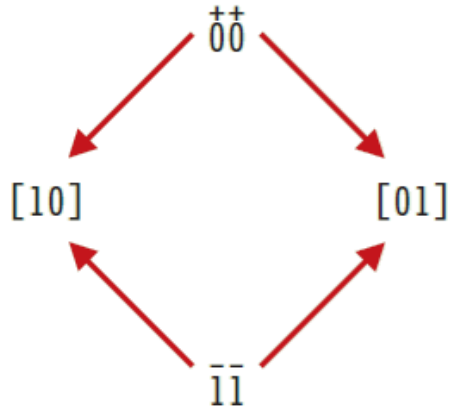
graphe des séquences d'états



Asynchrone

$$\begin{cases} X = \overline{y} \\ y = \overline{x} \end{cases}$$

xy	XY
$\overset{+}{0}\overset{+}{0}$	11
[01]	01
[10]	10
$\overline{1}\,\overline{1}$	00

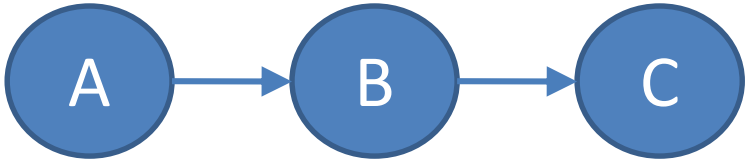


source : De Jong et Thieffry 2002

- Théorème de Bayes

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

  - A = (a1, a2, …, an)
  - B = (b1, b2, …, bn)
  - Pas faisable avec des milliers de variables
  - Hypothèse de l'indépendance **ou**
  - prise en compte partielle par modèle réseau
    - modélisation de l'influence d'un facteur de transcription sur ces cibles



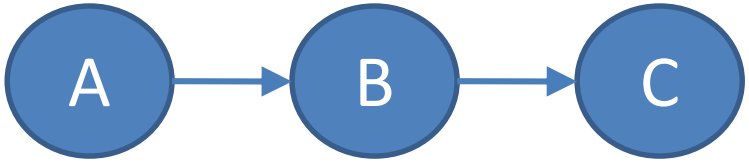
- Niveau d'expression de B dépend de celui de A (influence)
  - P(B/A)
- Le niveau d'expression de C
  - dépend de celui de B
  - est indépendant de celui de A sachant B

	cond1	cond2	cond3	cond4	cond5
A	off	low	high	low	low
B	off	low	high	high	low
C	off	off	low	high	low

observations

P(B/A)

B\A	off (1)	low (3)	high (1)
off	1/1	0/3	0/1
low	0/1	2/3	0/1
high	0/1	1/3	1/1

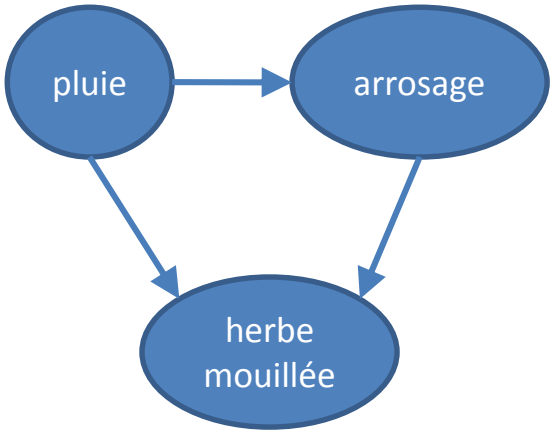


- Niveau d'expression de B dépend de celui de A (influence)
  - P(B/A)
- Le niveau d'expression de C
  - dépend de celui de B
  - est indépendant de celui de A sachant B

- Avantages :
  - capture l'aspect stochastique de la régulation
  - posibilité d'intégrer des régulations connues
  - peu de sur-apprentissage et robustesse
  - quantitatif (niveau d'expression) ou qualitatif (on/off)
- Inconvénient : Pas de cycle donc pas de boucle d'auto-régulation

Résaux bayésiens : illustration

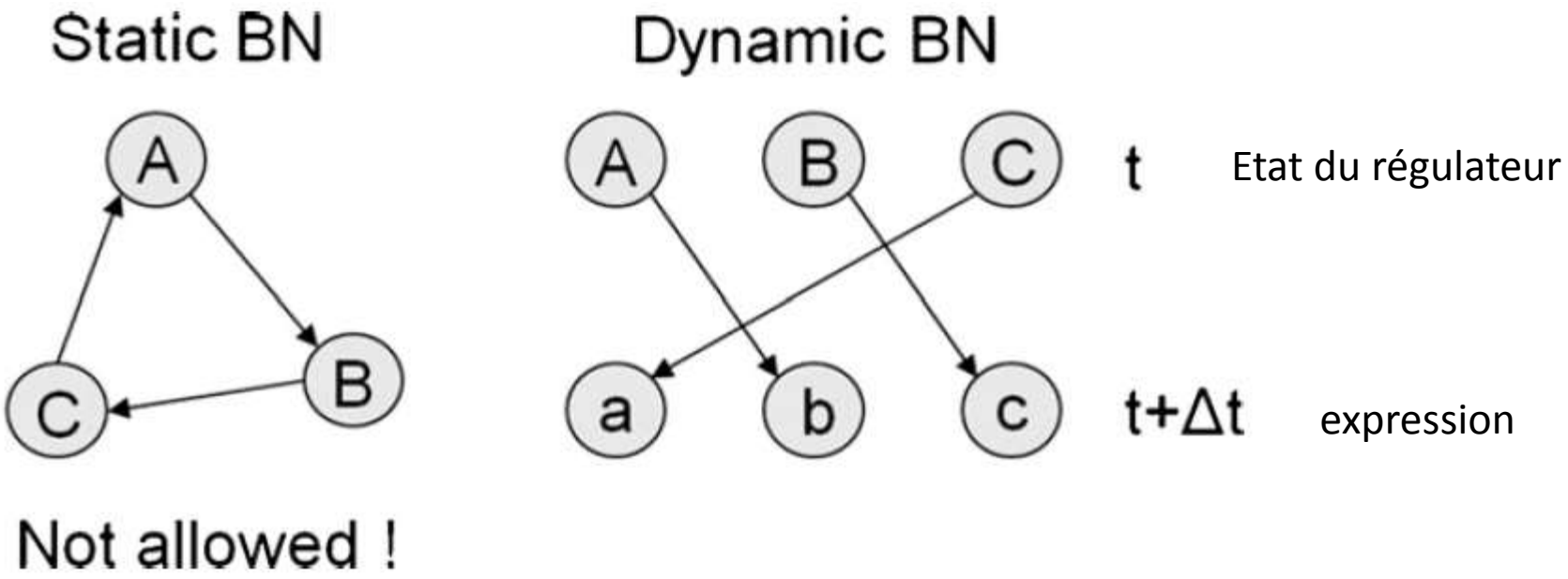
T	F
0.2	0.8



		arrosage / pluie	
		T	F
pluie	T	0.01	0.99
	F	0.4	0.6

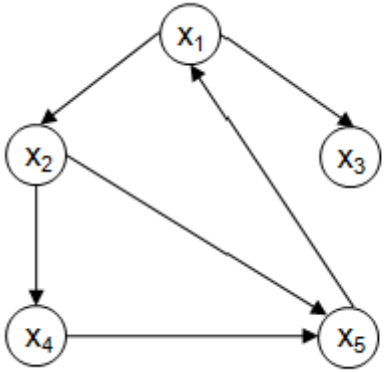
		mouilée	
arrosage	pluie	T	F
F	F	0	1
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

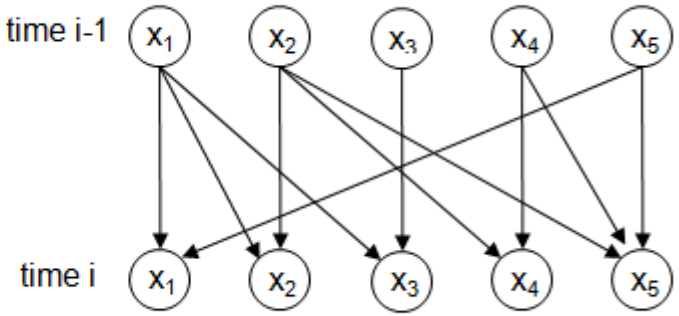
- Information temporelle (temps discret)
  - variable indicée par son pas de temps
  - distribution des probabilités d'une variable dépend de l'état de ses prédécesseurs au pas de temps précédent

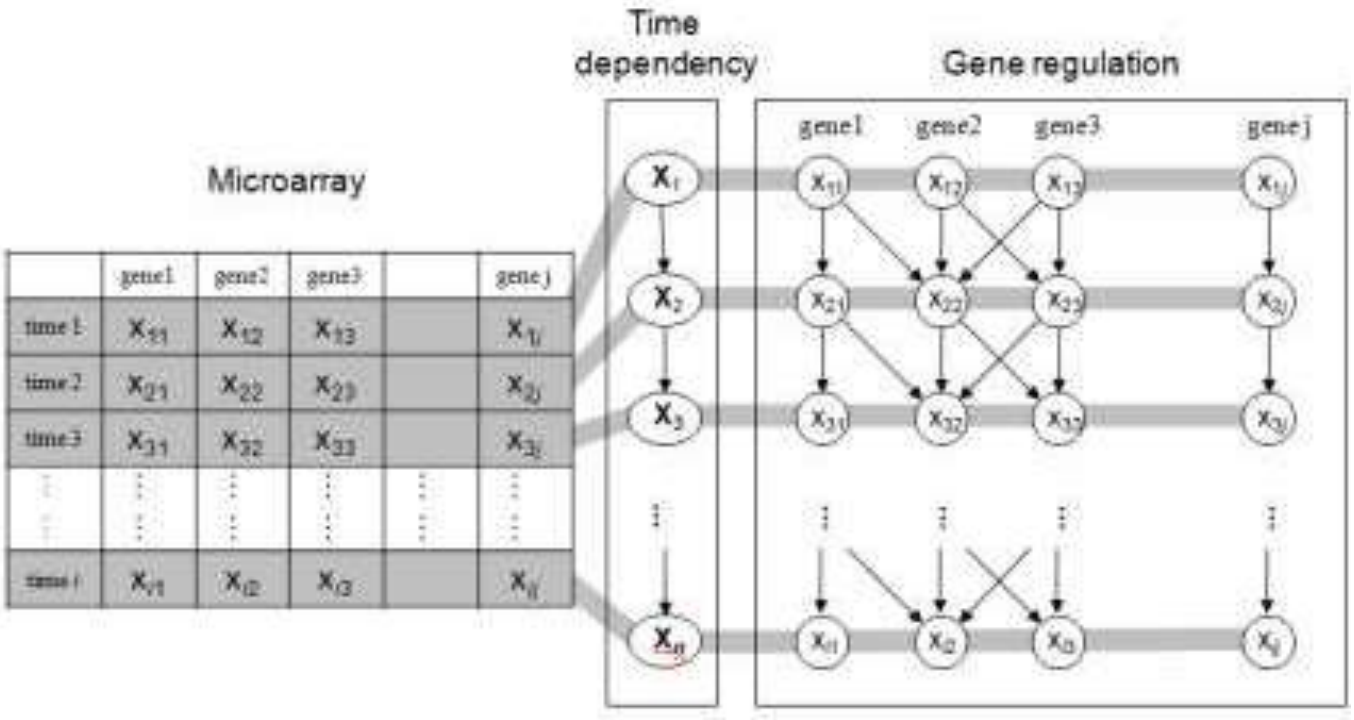


source : Hecket *et al.* 2009

Réseaux bayésiens dynamiques







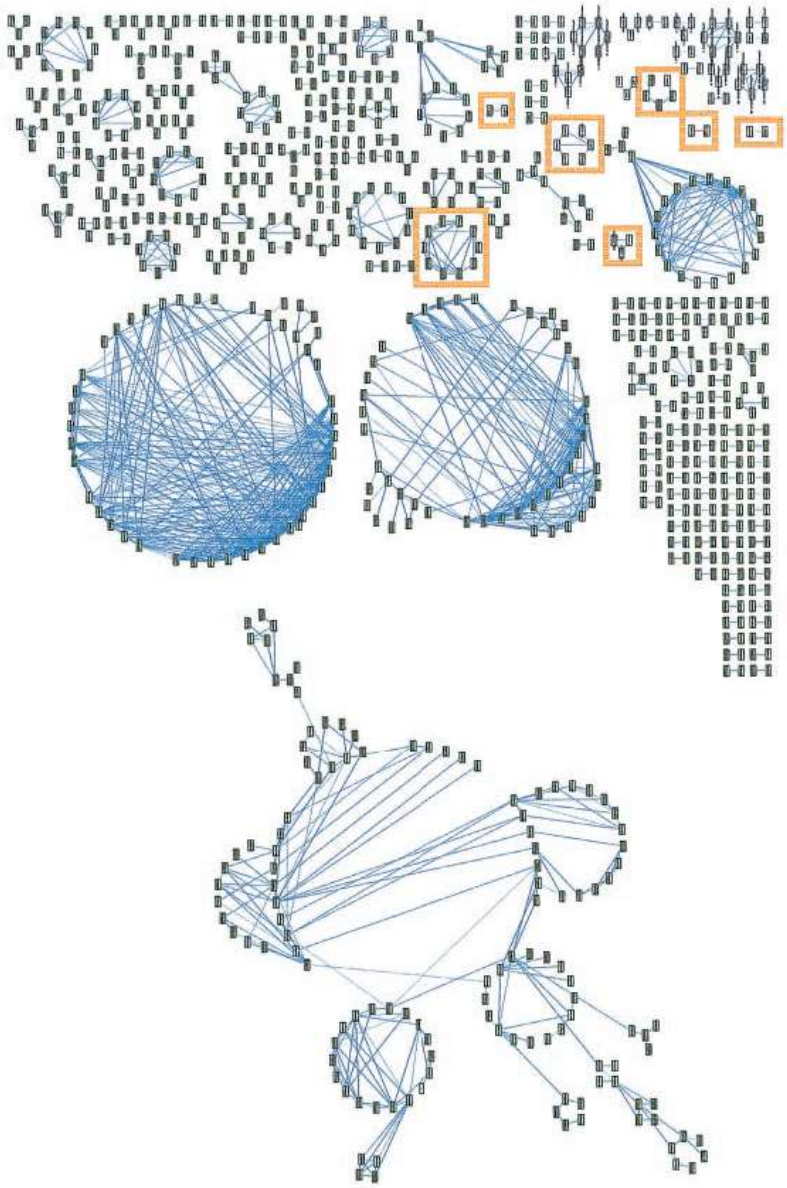
source : Chai and Mohd, 2011

- Adéquation entre observations et modèle
- Généralement en 3 étapes :
  - inférence de(s) la structure
  - détermination des paramètres
  - sélection du meilleur modèle
- Optimisation
  - de la structure (2^n-1 connexions pour chaque sommet, BIC)
    - feature selection : considérer uniquement les gènes différentiellement exprimés
    - feature mapping : agréger les ensembles tels que les opérons ou les gènes co-exprimés impliqués dans un même processus biologique
  - des paramètres : fonction de score
    - moindres carrés
    - maximum de vraissemblance
- Contraintes et incorporation de connaissances

- Mesure de similarité entre profils
  - Basée sur la corrélation entre l'activité de 2 éléments

$$\hat{r}^2 = \frac{r}{abs(r)} r^2$$

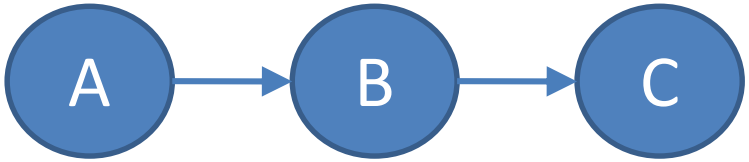
  - seuil pour l'inférence d'un lien entre les éléments



source : Butte *et al.* 2000

- Information Mutuelle

$$I(X;Y) = \sum_{i,j} P(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$



observations

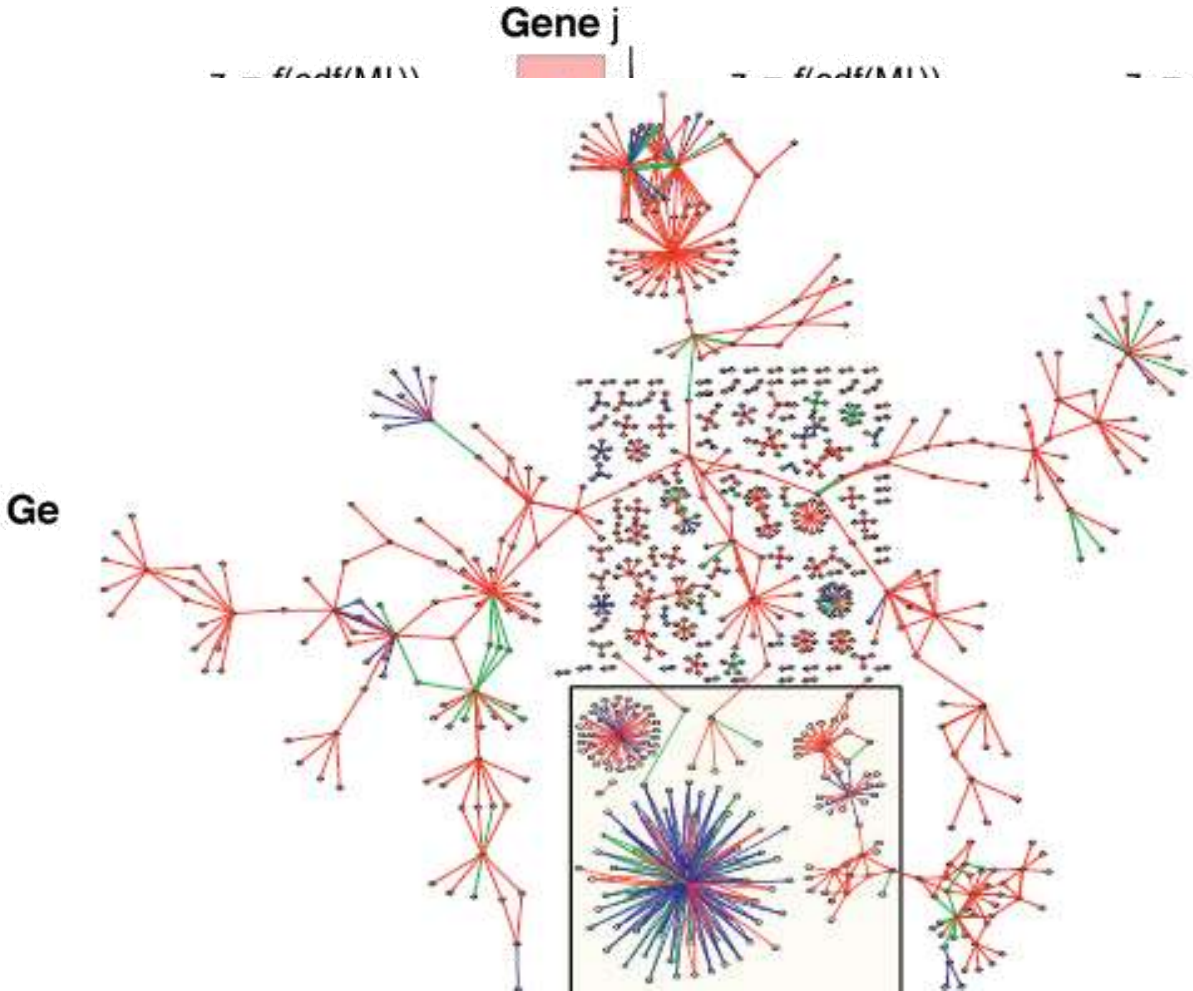
	cond1	cond2	cond3	cond4	cond5
A	high	low	high	low	low
B	high	low	high	high	low

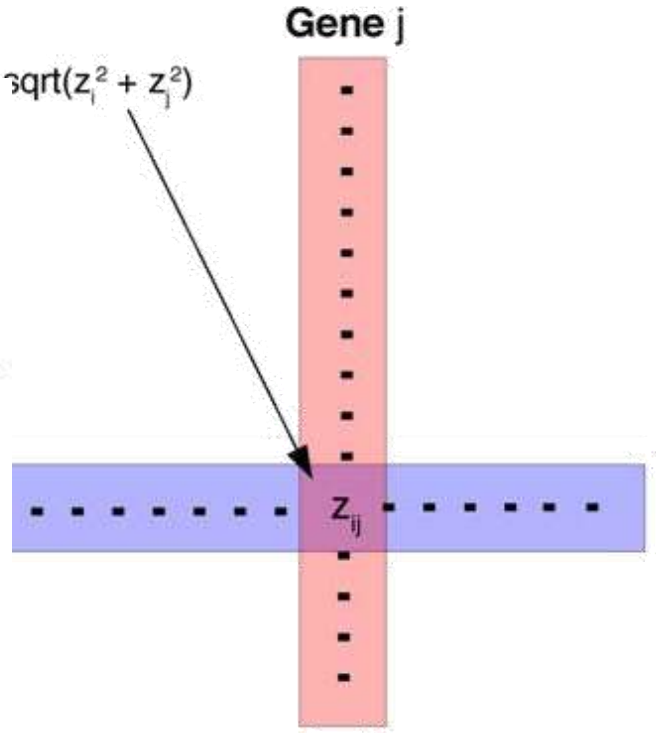
	B = low	B = high	total
A = low	2 / 5	1 / 5	3 / 5
A = high	0	2 / 5	2 / 5
total	2 / 5	3 / 5	1

```
   ab      ab    a  b      aB      aB    a  B      Ab    AB     AB   A   B
> .4* log( .4 / (.6*.4) ) + .2*log( .2 / (.6*.6) ) + 0 + .4*log( .4 / (.4*.6) )
[1] 0.2911032
```

- Information Mutuelle
- Context Likelihood Ration (CLR)

$$I(X;Y) = \sum_{i,j} P(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$





source : Faith *et al.* 2007

BMC Bioinformatics

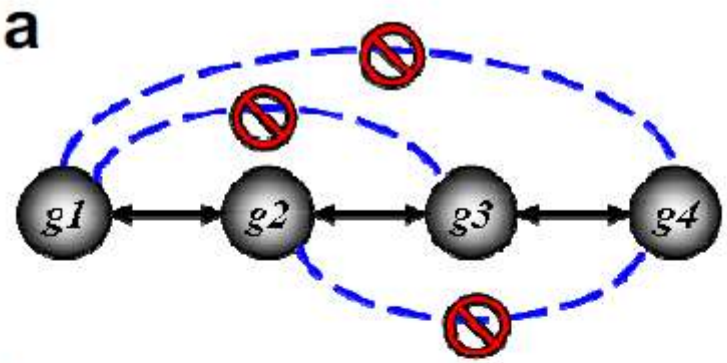
BioMed Central

Proceedings

Open Access

ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context

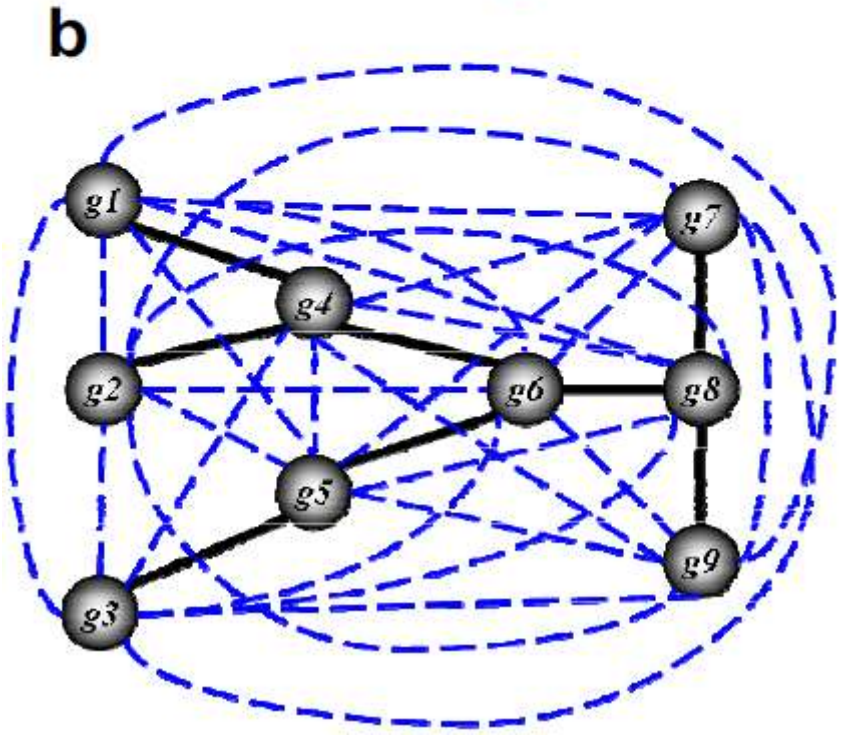
Adam A Margolin[1,2], Ilya Nemenman[2], Katia Basso[3], Chris Wiggins[2,4], Gustavo Stolovitzky[5], Riccardo Dalla Favera[3] and Andrea Califano*[1,2]



Data Processing Inequality

The DPI (Figure 2) [14] states that if genes g_1 and g_3 interact only through a third gene, g_2, (i.e., if the interaction network is $g_1 \leftrightarrow \ldots \leftrightarrow g_2 \leftrightarrow \ldots \leftrightarrow g_3$ and no alternative path exists between g_1 and g_3), then

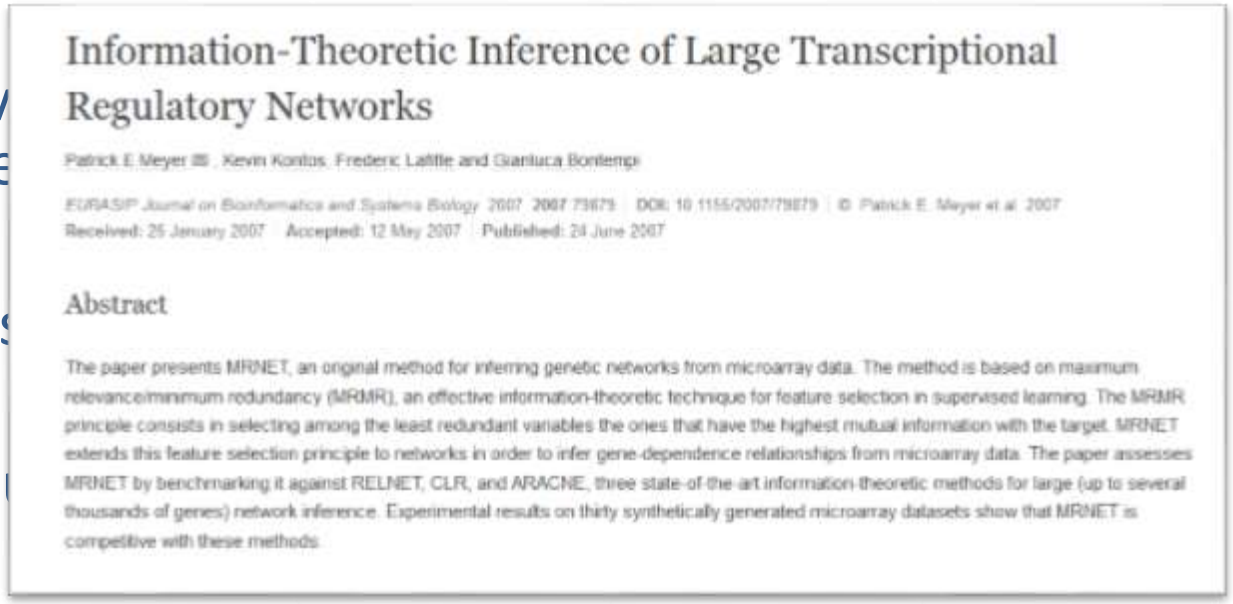
$$I(g_1, g_3) \leq \min [I(g_1, g_2); I(g_2, g_3)]. \quad (3)$$



MRNET : Maximum Relevance - Minimum Redundancy

- Apprentissage superv
  fonction de V variable

- La première variable s
  X_i telle que $I(X_i, Y)$
  est maximum (maxim



Information-Theoretic Inference of Large Transcriptional Regulatory Networks

Patrick E Meyer ✉, Kevin Kontos, Frederic Lafitte and Gianluca Bontempi

EURASIP Journal on Bioinformatics and Systems Biology 2007 2007 79879 | DOI: 10.1155/2007/79879 | © Patrick E. Meyer et al. 2007

Received: 25 January 2007 | Accepted: 12 May 2007 | Published: 24 June 2007

Abstract

The paper presents MRNET, an original method for inferring genetic networks from microarray data. The method is based on maximum relevance/minimum redundancy (MRMR), an effective information-theoretic technique for feature selection in supervised learning. The MRMR principle consists in selecting among the least redundant variables the ones that have the highest mutual information with the target. MRNET extends this feature selection principle to networks in order to infer gene-dependence relationships from microarray data. The paper assesses MRNET by benchmarking it against RELNET, CLR, and ARACNE, three state-of-the-art information-theoretic methods for large (up to several thousands of genes) network inference. Experimental results on thirty synthetically generated microarray datasets show that MRNET is competitive with these methods.

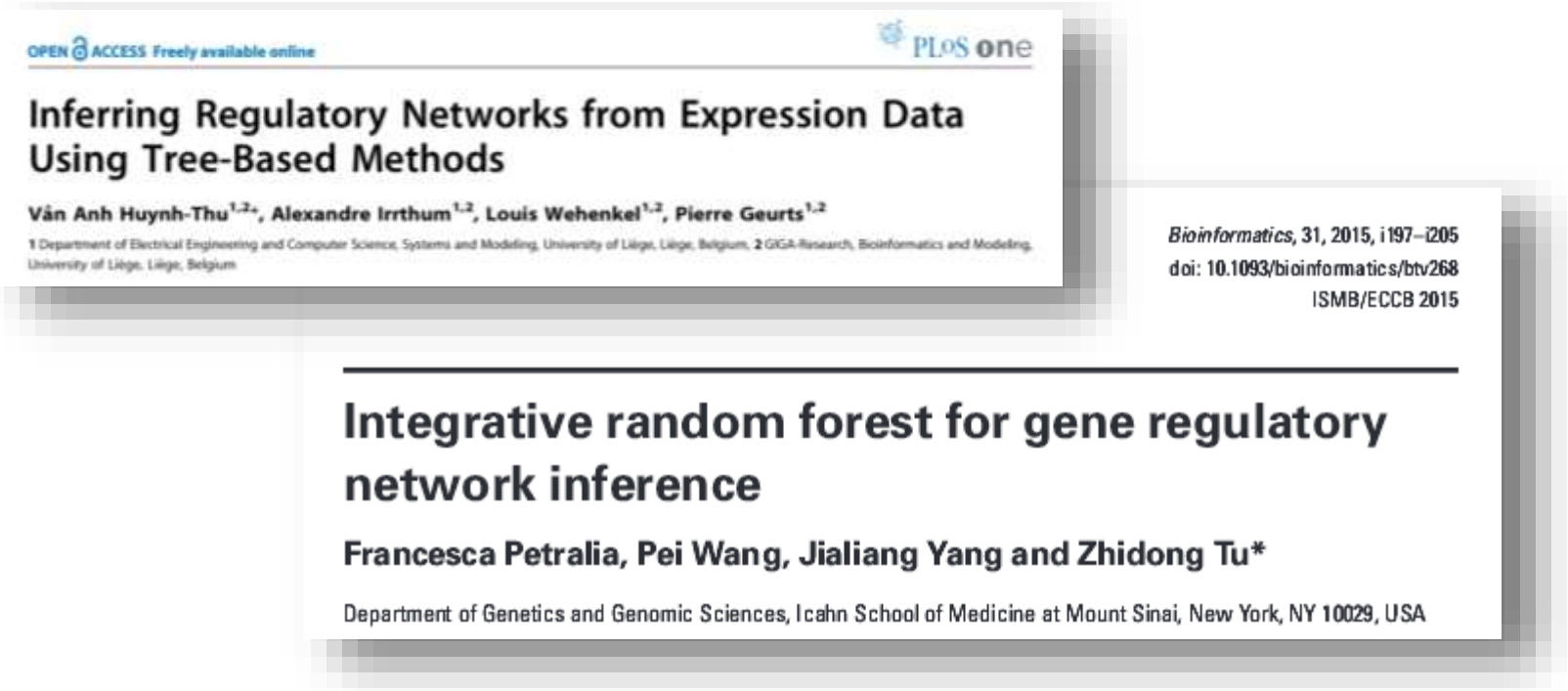
- La seconde variable sélectionnée est telle que $I(X_j, Y) - I(X_j, X_i)$ est maximum (minimum redundancy).

- Les suivantes telles que

$$I(X_k, Y) - \frac{1}{|S|} \sum_{X_i \in S} I(X_k, X_i)$$

est maximum

- Apprentissage supervisé : on cherche à prédire une sortie Y en fonction de V variables en entrée.
- Utilisation de forêts aléatoires (GENIE3) avec un tirage aléatoire des variables en entrée qui dépend d'autres sources de données (iRafNet)



OPEN ACCESS Freely available online

PLoS one

Inferring Regulatory Networks from Expression Data Using Tree-Based Methods

Vân Anh Huynh-Thu[1,2]*, Alexandre Irrthum[1,2], Louis Wehenkel[1,2], Pierre Geurts[1,2]

1 Department of Electrical Engineering and Computer Science, Systems and Modeling, University of Liège, Liège, Belgium, 2 GIGA-Research, Bioinformatics and Modeling, University of Liège, Liège, Belgium

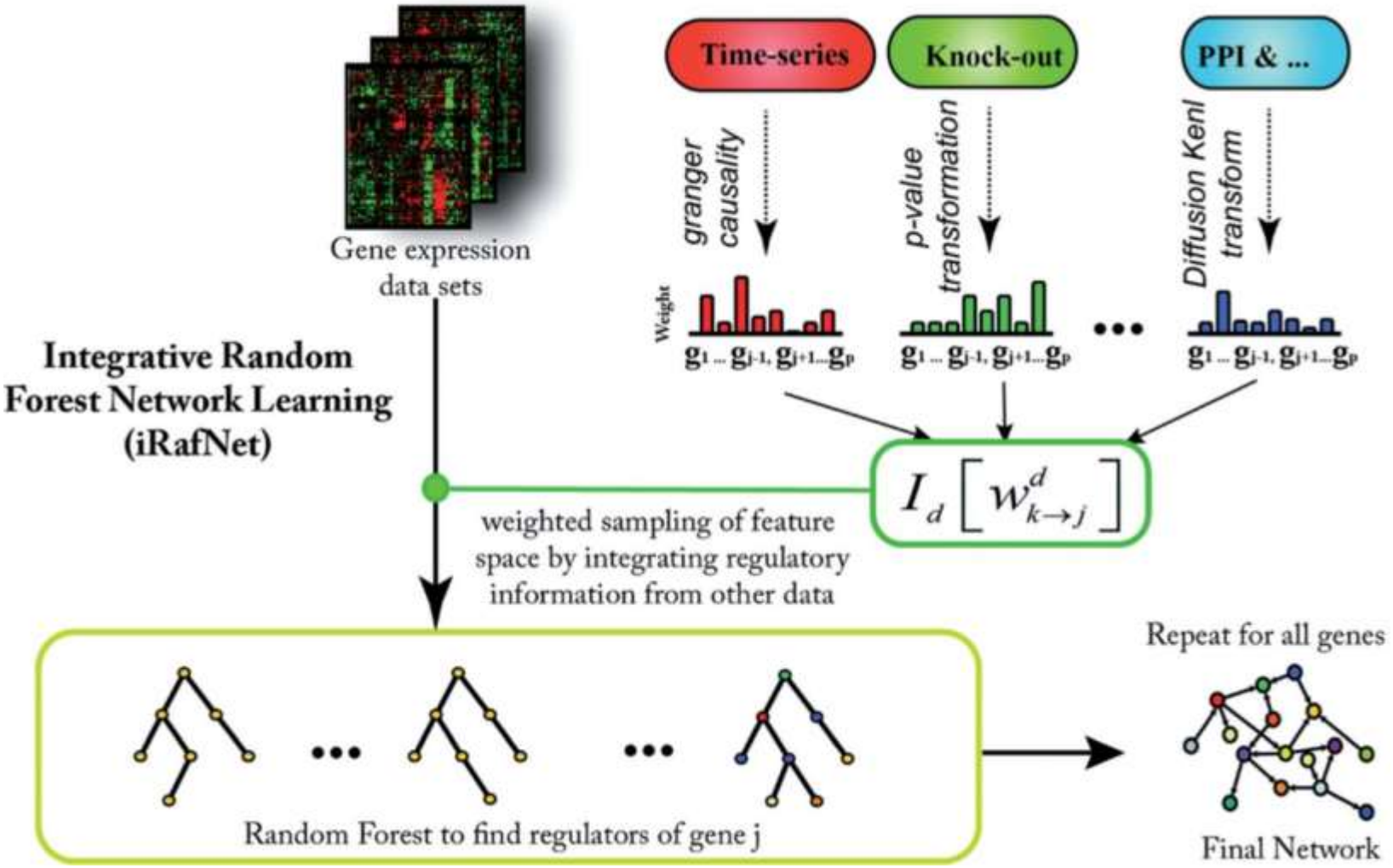
Bioinformatics, 31, 2015, i197–i205
doi: 10.1093/bioinformatics/btv268
ISMB/ECCB 2015

Integrative random forest for gene regulatory network inference

Francesca Petralia, Pei Wang, Jialiang Yang and Zhidong Tu*

Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

iRafNet schematics.

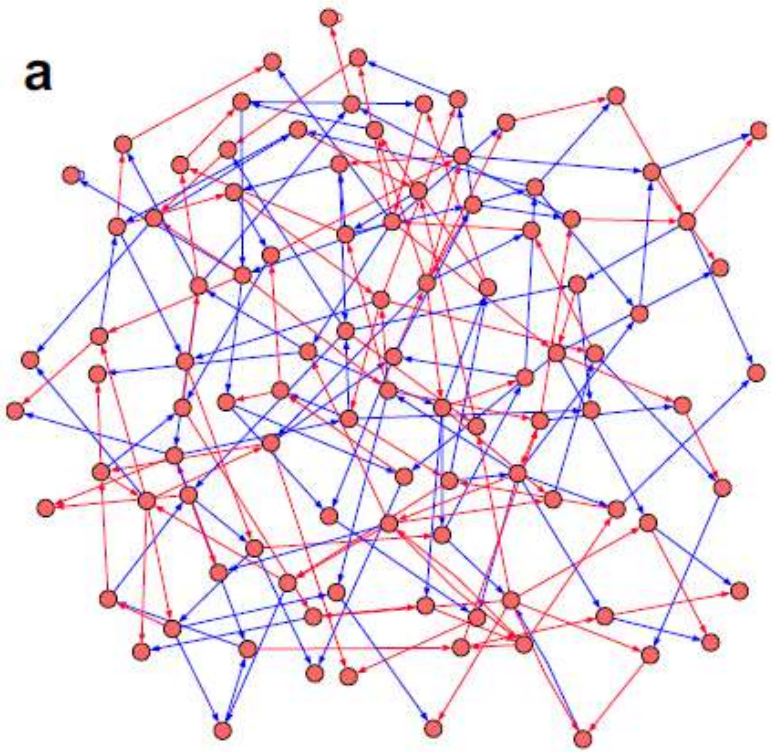


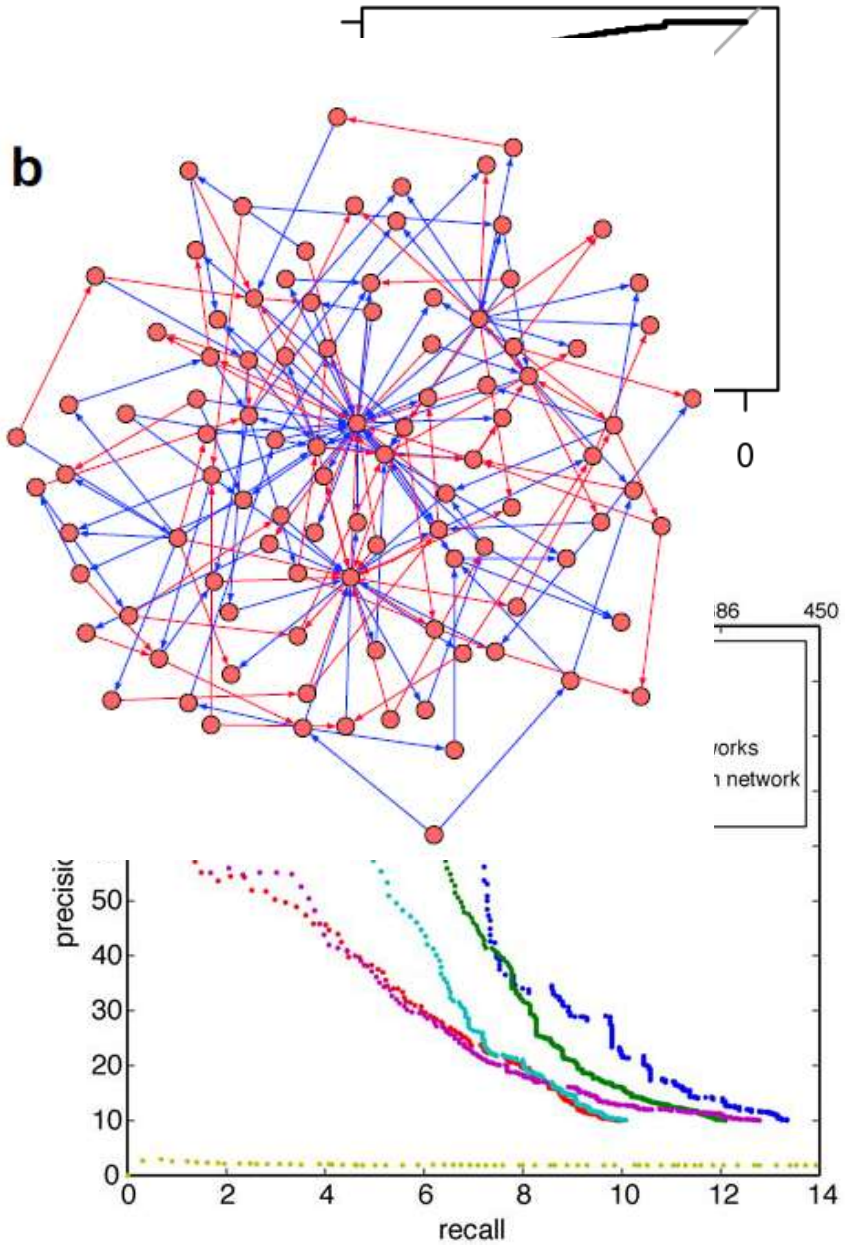
Francesca Petralia et al. Bioinformatics 2015;31:i197-i205

Bioinformatics

Evaluation et validation

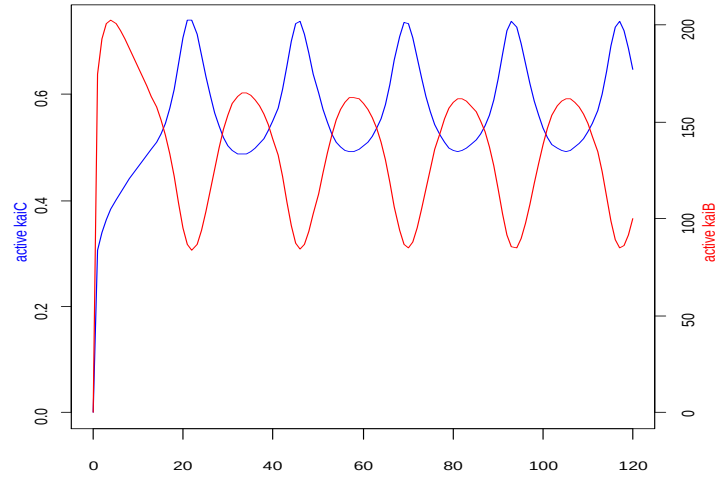
- Besoin d'un jeu de données de r
  e
- T
  T
  li
  - Precision = TP / (TP+FP)
  - Spécificité = TN / (TN+FP)
  - FPR = 1 – spécificité
  - FDR = 1 - précision



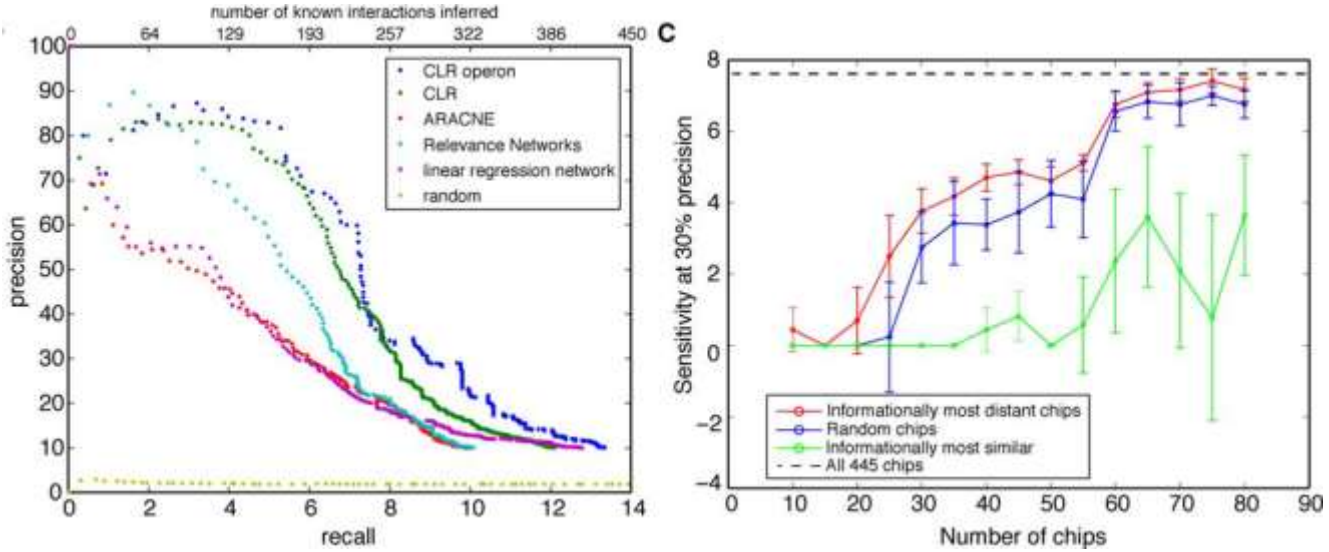


source : Faith *et al.* 2007

- Validation d'un modèle, d'un comportement observé



- Simulation et prédiction



source : Faith *et al.* 2007

Controls

A

LexA

	p-value	Motif location
recN	4.4e-09	
yebG	6.4e-10	
lexA	8.3e-10	
uvrA	1e-08	
sulA	1.7e-08	
dinI	2e-08	
dinP	4.4e-08	
recA	6.5e-08	
SCALE		1 25 50 75 100 125

The known motif is found in
8 out of 13 promoters

Putative novel regulons

C

YnaE

	p-value	Motif location
cspB	3.9e-10	
cspG	6.3e-09	
b1374_s	1.1e-08	
cspH	3.7e-08	
b1459	2.9e-07	
rhsE	6.1e-07	
SCALE		1 25 50 75 100 125

A conserved motif is found in
6 out of 8 promoters

source : Faith *et al.* 2007