

THÈSE

Pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITE DE NANTES

Ecole Polytechnique de l'Université de Nantes

Laboratoire d'Informatique de Nantes/Atlantique (LINA)

Ecole Doctorale Sciences et Technologies
de l'Information et Mathématiques (STIM)

Domaine de Recherche : **Informatique**

Présentée par

Hoai-Tuong NGUYEN

Etude différentielle de population de réseaux bayésiens:
Application en pathologies cancéreuses et réponses thérapeutiques

Directeur de thèse : M. Philippe LERAY

Co-direction : **M. Gérard RAMSTEIN, M. Yannick JACQUES**

Soutenue le 15 décembre 2011

JURY

M. Louis WEHENKEL	Professeur à l'Université de Liège	Rapporteur
M. Salem BENFERHAT	Professeur à l'Université d'Artois	Rapporteur
M. Philippe LERAY	Professeur à l'Université de Nantes	Examineur
M. Gérard RAMSTEIN	Maitre de Conférence à l'Université de Nantes	Examineur
M. Yannick JACQUES	Directeur de Recherche à l'INSERM 892	Examineur
M. (<i>à définir...</i>)	Professeur à l'Université de (<i>à définir...</i>)	Président

Table des matières

Table des figures	iii
Liste des tableaux	v
I Reconstruction de réseaux de régulation génétique	1
1 Régulation génétique	1
1.1 De l'organisme aux gènes	1
1.2 Régulation génétique : <i>activation/inhibition</i>	2
2 Réseaux de régulation génétique	3
2.1 Définition	3
2.2 Exemple	4
2.3 Problématiques	5
2.3.1 Complexité	5
2.3.2 Régulation mal connue	6
2.4 Transcriptome : techniques expérimentales et analyse des données	6
2.4.1 Puces à ADN	6
2.4.2 Traitement des données	8
2.4.3 Discrétisation	9
3 Reconstruction de réseaux de régulation génétique	10
3.1 Introduction	10
3.2 Méthodes de reconstruction de réseaux de régulation génétique	10
3.2.1 Description des entités biologiques	11
3.2.2 Topologie du réseau	11
3.2.3 La logique de contrôle	11
3.2.4 Le modèle dynamique	13
3.2.5 Réseaux booléens	13
3.2.6 Réseaux bayésiens	17

Table des matières

4	Conclusion	19
	Références bibliographiques	21

Table des figures

I.1	Schéma de la régulation génétique.	3
I.2	Exemple fictif de réseau de regulation transcriptomique. Les grands rectangles représentent la partie codante du génome, en amont desquels figurent les sites de liaison (region promotrice des gènes). Les cercles représentent les facteurs de transcription. D'après [Current approaches to gene regulatory network modelling, Thomas Schlitt and Alvis Brazma].	4
I.3	Biosynthèse des acides aminés valine, leucine et isoleucine.	5
I.4	Schéma de principe des expériences à base de puces à ADN. [BAR04]	7
I.5	Exemple d'une matrice de données d'expression des gènes issues de puces à ADN.	8
I.6	Exemple de la normalisation de données de puces.	9
I.7	Exemple de la discretisation de données de puces.	10
I.8	Modélisation de la logique de contrôle par une fonction booléenne. La réponse de y est définie par une fonction logique. Figure extraite de. [DRGB ⁺ 10]	12
I.9	Arbre de décision du gène CLN2 (d'après [SKB03]). Dans cet exemple, la régulation du gène est prédite à partir des gènes SWI5, CLN1 et CDC28. Les étiquettes sur les arcs indiquent les seuils de valeur d'expression définis à partir de données de puces à ADN.	12
I.10	Exemple d'un réseau obtenu par l'approche basée sur les réseaux de confiance. (A) Le réseau le plus large possible. (B) et (C) Deux petites branches dans (A) en version ZOOM. [BKK00]	14
I.11	Exemple d'un réseau dynamique pour représenter un RRG avec des données temporelles [Kau69]. (a) Matrice de transition de niveau de regulation dans 30 tranches temporelles. (b) Transition entre des tranches présentées dans (a)	15
I.12	Exemple d'un réseau booléen simple pour représenter un RRG [IEW02].	16
I.13	Exemple d'un réseau obtenu par l'approche basée sur les équations différentielles.[?]	17
I.14	Exemple d'un RB obtenu par la reconstruction du RRG présenté dans [Fri04].	18

Table des figures

Liste des tableaux

I.1	Exemple des génomes eucaryotes, avec la taille du génome, le nombre de gènes, le nombre des facteurs de transcription (FT) connus et le pourcentage d'ADN non codant.	6
-----	---	---

Liste des tableaux

Chapitre I

Reconstruction de réseaux de régulation génétique

To-to-list	
Tasks	Status
2.1	2.2

1 Régulation génétique

Nous présentons brièvement dans cette sous-section des notions de base en génétique que nous utilisons dans la suite de cette thèse. Une présentation approfondie est disponible dans [\[GWLC07\]](#).

1.1 De l'organisme aux gènes

Définition 1. *Organisme*

Un **organisme** est un ensemble d'éléments composant une structure fonctionnelle de l'être vivant.

Définition 2. *Cellule*

Une **cellule** est l'unité structurale, fonctionnelle et reproductrice constituant tout ou partie d'organisme.

Définition 3. *ADN*

L'**ADN** (acide désoxyribonucléique) est une molécule, présente dans toutes les cellules vivantes, qui renferme l'ensemble des informations nécessaires au développement et au fonctionnement d'un organisme.

Définition 4. *Gène*

Un **gène** est une séquence de l'ADN qui spécifie la synthèse d'une chaîne de polypeptides ou d'un acide ribonucléique (ARN) fonctionnel.

Définition 5. *ARN*

L'**ARN** (acide ribonucléique) est une molécule, présente dans toutes les cellules vivantes, qui est une copie d'une région de l'un des brins de l'ADN.

Définition 6. *Protéine*

Une '**protéine** est une macromolécule, présente dans toutes les cellules vivantes, qui est composée d'une ou plusieurs chaînes d'acides aminés.

Un organisme est une forme de vie individuelle comme une plante, un animal, une bactérie, protiste, ou un champignon. Les organismes sont divisés en deux grandes familles : les *prokaryotes*, organismes unicellulaires sans noyau et les *eucaryotes*, organismes dont les cellules ont un noyau qui renferme l'*information génétique*.

L'information génétique est portée par ADN. L'ADN joue un rôle indispensable dans la synthèse des protéines. Pour produire des protéines, la première étape consiste à transcrire l'ADN en ARN messager (mARN). La deuxième étape consiste ensuite à traduire mARN en protéine.

Dans la synthèse des protéines, les gènes sont considérés comme les "recettes" qui contiennent les instructions en "langage" particulier utilisant un alphabet de 4 lettres (*A, C, G et T*). L'ordre de ces lettres, c'est-à-dire la séquence des gènes, détermine la forme et la fonction de la protéine dans l'organisme.

1.2 Régulation génétique : *activation/inhibition*

Le comportement d'une cellule est déterminé par la concentration de protéines particulières (voir la Figure [I.1](#)). Ces protéines sont appelés des protéines régulatrices (ou simplement *régulateurs*). Ils sont des facteurs de transcription qui peuvent intervenir à des sites spécifiques, appelé *sites de fixation*, relativement liés aux régions régulatrices (ADN non-codant) et aux régions codants (qui sont traduites en protéines). Par l'existence du catalyseur (une substance qui augmente ou diminue la vitesse d'une réaction chimique), ces protéines *activent* ou *inhibent* l'expression d'un gène. Les régulateurs peuvent eux-mêmes être régulés, dans ce cas, ils participent à une voie (pathway) de régulation génétique.

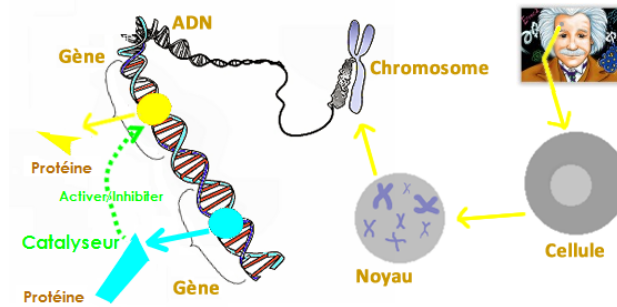


Figure I.1 – Schéma de la régulation génétique.

En général, un gène cible peut être régulé par une combinaison de facteurs de transcription, et un facteur de transcription peut réguler plusieurs gènes cibles. Pour modéliser cette régulation, les biologistes ont adopté raisonnablement le concept d'un réseau. Ce réseau s'appelle *réseau de régulations génétiques* (RRG).

2 Réseaux de régulation génétique

2.1 Définition

Les réseaux de régulation génétique font partie de trois types de réseaux biologiques : (1) *réseau de régulation de gènes* : deux gènes sont liés si l'expression d'un gène module l'expression d'un autre soit par l'activation soit par l'inhibition ; (2) *réseau d'interaction des protéines* : les protéines qui sont connectées à des interactions physiques ou métaboliques et les voies de signalisation de la cellule ; (3) *réseau métabolique* : produits métaboliques et les substrats qui participent à une réaction ;

Les interactions entre gènes dépendent de nombreux paramètres de différentes natures. C'est la raison pour laquelle la structure de réseaux de régulations génétiques est encore largement méconnue, malgré de nombreux travaux entrepris depuis le travail du prix Nobel 1965 des chercheurs A. Lwoff, F. Jacob et J. Monod [JM61] qui pour la première fois décrivent un ensemble de faits biologiques permettant d'imaginer une structure de régulation de l'expression de certains gènes de la bactérie *Escherichia coli*.

Définition 7. Réseau de régulation génétique

Le réseau de régulation génétique (RRG) est un graphe orienté et signé qui indique quelles sont les activations et les inhibitions présentes entre les gènes du réseau. Dans le RRG, chaque nœud est un gène et chaque lien entre deux nœuds représente une interaction génétique entre deux

gènes. Les interactions se traduisent par des influences sur le niveau d'expression des gènes. RRG est aussi appelé le réseau d'interaction entre les gènes.

2.2 Exemple

La biologie moléculaire fait donc fortement appel à la notion de réseaux, dans des contextes différents et impliquant des entités biologiques distinctes. L'un des modèles les plus étudiés est probablement le réseau de régulation transcriptomique se focalisant sur les relations existantes entre les gènes et les produits de gènes. Ce type de modélisation a pour but de définir la manière dont un gène est régulé en réponse à certains signaux. Dans les années 60, les biologistes ont montré que les gènes possèdent des séquences proches servant à leur régulation et que des protéines sont capables de se lier à ses séquences, permettant ainsi le contrôle de l'expression du gène selon deux modes, l'activation ou la répression. Comme ces protéines régulatrices sont elles-mêmes le produit de gènes, un réseau de régulation se forme, avec ses boucles de rétroaction positives et négatives. La figure I.2 illustre ce mécanisme de régulation. Cet exemple de réseau est bien évidemment une simplification de l'activité biologique réelle qui possède d'autres contraintes (de régulation post-transcriptionnelle par exemple).

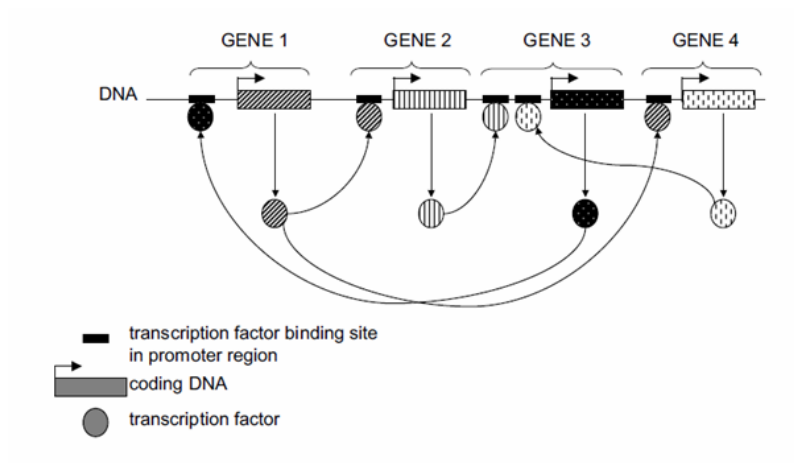


Figure I.2 – Exemple fictif de réseau de régulation transcriptomique. Les grands rectangles représentent la partie codante du génome, en amont desquels figurent les sites de liaison (region promotrice des gènes). Les cercles représentent les facteurs de transcription. D'après [Current approaches to gene regulatory network modelling, Thomas Schlitt and Alvis Brazma].

2.3 Problématiques

Les RRG permettent de contrôler le fonctionnement et le développement des organismes vivants via des relations entre gènes. L'étude sur des RRG pose des problématiques importantes et diverses.

2.3.1 Complexité

Les processus cellulaires sont généralement d'une grande complexité et impliquent de nombreuses molécules. Le métabolisme d'une cellule nécessite de multiples réactions biochimiques interagissant entre-elles : le produit d'une réaction entraîne à son tour une série de réactions fortement intriquée. La figure I.3 montre un exemple de réseau métabolique de la base KEGG [citation]. De même, les molécules de signalisation sont liées entre elles et communiquent sous la forme de cascades de signalisation spécifiques. La régulation des relations entre les gènes et leurs produits implique également un nombre important de molécules. La situation se complexifie encore si on considère le fait que ces réseaux sont dépendants entre eux : par exemple, un processus biologique au sein d'une cellule peut être affectée par des signaux extra-cellulaires.

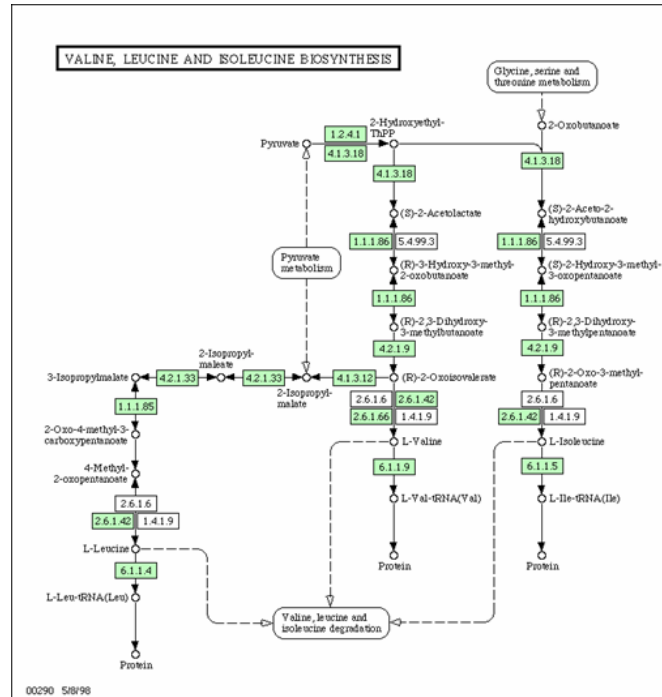


Figure I.3 – Biosynthèse des acides aminés valine, leucine et isoleucine.

Chapitre I. Reconstruction de réseaux de régulation génétique

Organisme	Taille du génome	Nbre de gènes	Nbre de FT	ADN non-codant
Colibacille	$\sim 4.6Mb$	~ 4300	~ 100	$\sim 10\%$
Levure	$\sim 13Mb$	~ 6200	~ 250	$\sim 30\%$
Homme	$\sim 3200Mb$	~ 30000	~ 1700	$\sim 98\%$

Tableau I.1 – Exemple des génomes eucaryotes, avec la taille du génome, le nombre de gènes, le nombre des facteurs de transcription (FT) connus et le pourcentage d'ADN non codant.

2.3.2 Régulation mal connue

Tous ces gènes ne peuvent être exprimés à la fois. En effet, pour caractériser un organisme vivant, chaque gène s'exprime dans un environnement spécifique, à des moments précis et pendant une durée limitée en fonction de son niveau d'expression. Toutefois, une proportion très importante de tous les génomes eucaryotes est composée de la classe d'ADN non-codant (l'ensemble des séquences du génome qui ne sont pas traduites en protéines) dont la fonction biologique est mal connue et souvent sous-estimée (voir Tableau I.1). En effet, seul un petit nombre de gènes fonctionnent comme des activateurs ou inhibiteurs, par conséquent leur identification est un problème important et difficile. De plus la plupart des organismes, la structure de ces réseaux est encore largement méconnue.

2.4 Transcriptome : techniques expérimentales et analyse des données

2.4.1 Puces à ADN

Définition 8. *Puces à ADN* [?]

Une puce à ADN (aussi appelées puces à gènes, biopuces, ou en anglais "DNA chip, DNA-microarray, biochip") est un ensemble de molécules d'ADN fixées en rangées ordonnées sur une petite surface qui peut être du verre, du silicium ou du plastique. C'est une technique qui permet d'analyser et de quantifier simultanément l'expression de plusieurs milliers de gènes.

Principe :

Les puces à ADN [?] reposent sur le mécanisme d'hybridation de la double hélice d'ADN. Ce sont de lames en verre ou silicium de petite taille ($25 \times 75mm$) sur lesquels on peut synthétiser de milliers de gènes dont des séquences d'ADN. Ensuite, elles se sont hybridées avec

l'ensemble des transcrits issus d'une cellule cible (normale) pour marquer leur niveau d'expression grâce à la leur fluorescente. L'hybridation est une interaction des deux chaînes de séquences complémentaires (liaisons hydrogènes)

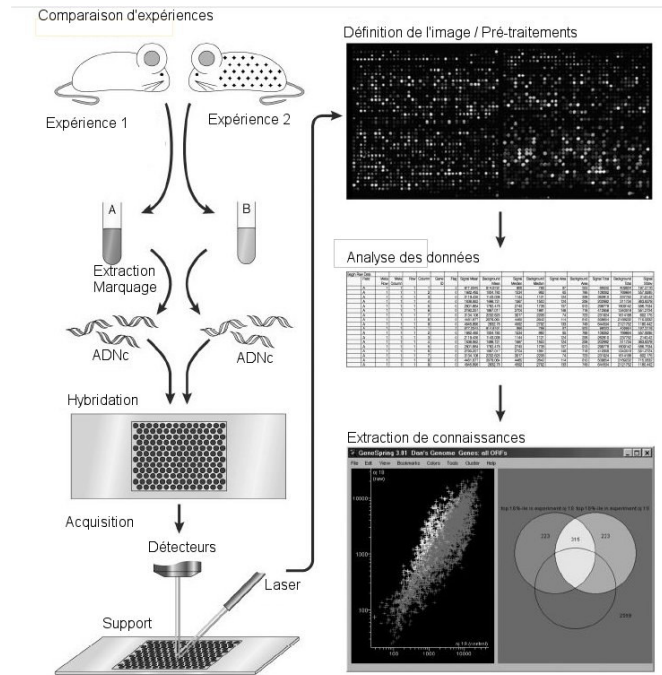


Figure I.4 – Schéma de principe des expériences à base de puces à ADN. [BAR04]

Les niveaux d'expression des gènes sont présentés sous forme d'ARN messagers qui, après traduction, vont permettre la production des protéines indispensables au fonctionnement de la cellule normale ou tumorale. Après la traitement d'image (scan), C'est une matrice où les lignes représentent les gènes, et les colonnes représentent les échantillons (expériences).

Problématiques :

Dans les puces à ADN, les changements des niveaux d'expression des gènes différents échantillons fournissent des informations qui permet des techniques d'ingénierie inverse ("*reverse engineering*" en anglais) pour reconstruire le réseau de régulation génétique. Pour les méthodes d'apprentissage à partir de données, l'utilisation de données de puces à ADN pose deux problématiques suivantes :

- *Premièrement*, il est nécessaire d'avoir suffisamment de données d'observations. la Pour-tant, dans les données de puces à ADN ne, le nombre de gènes est très élevé (30.000 gènes) et excédant toujours celui des échantillons (10 – 100 échantillons).

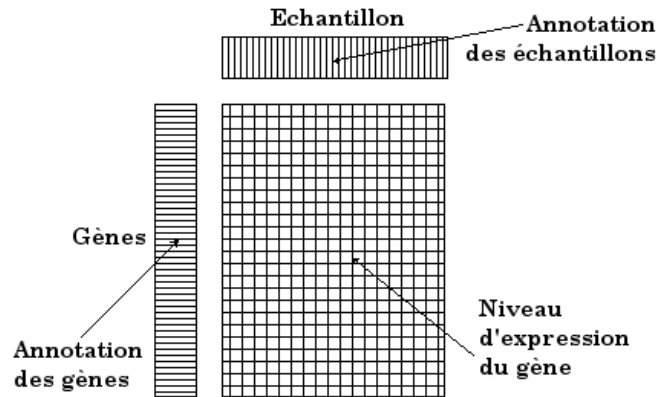


Figure I.5 – Exemple d’une matrice de données d’expression des gènes issues de puces à ADN.

- *Deuxièmement*, comme toutes méthodes de recueil de données, le problème du bruit de données est inévitable.

Nous présentons dans la sous-section ci-après les méthodes de traitement de données de puces à ADN.

2.4.2 Traitement des données

Le traitement de ces données d’expression est nécessaire avant de les utiliser dans l’algorithme d’apprentissage. Les différents problèmes à considérer sont les suivants :

- **Normalisation** : Les données d’expression souvent ne sont pas parfaites : valeurs manquantes, imprécises, non-homogènes. Donc, on parle généralement du "bruit" de donnée. Le traitement du bruit dans les données n’est pas un problème facile, surtout dans le contexte de la pauvreté de données¹. Parce qu’il n’est pas facile de distinguer ce qui est le résultat d’une erreur ou d’une différence non significative d’une observation aléatoire. Le résultat du traitement u bruit influe plus ou moins sur les résultats d’un modèle d’apprentissage. Donc les outils statistiques (traitement de bruit de fond, normalisation, etc.) sont apparues presque immédiatement après la naissance de la technologie de puces à ADN. Grâce à l’amélioration considérable des techniques de puces à ADN et les différentes techniques de traitement d’image et statistique proposées, nous avons besoin d’utiliser ce qui existe dans la littérature [Qua02] (voir la Figure I.6).
- **Réduction de dimensions** : Les données de puces à ADN sont présentées par les ma-

1. le nombre de gènes est très élevé et excédant toujours celui des échantillons

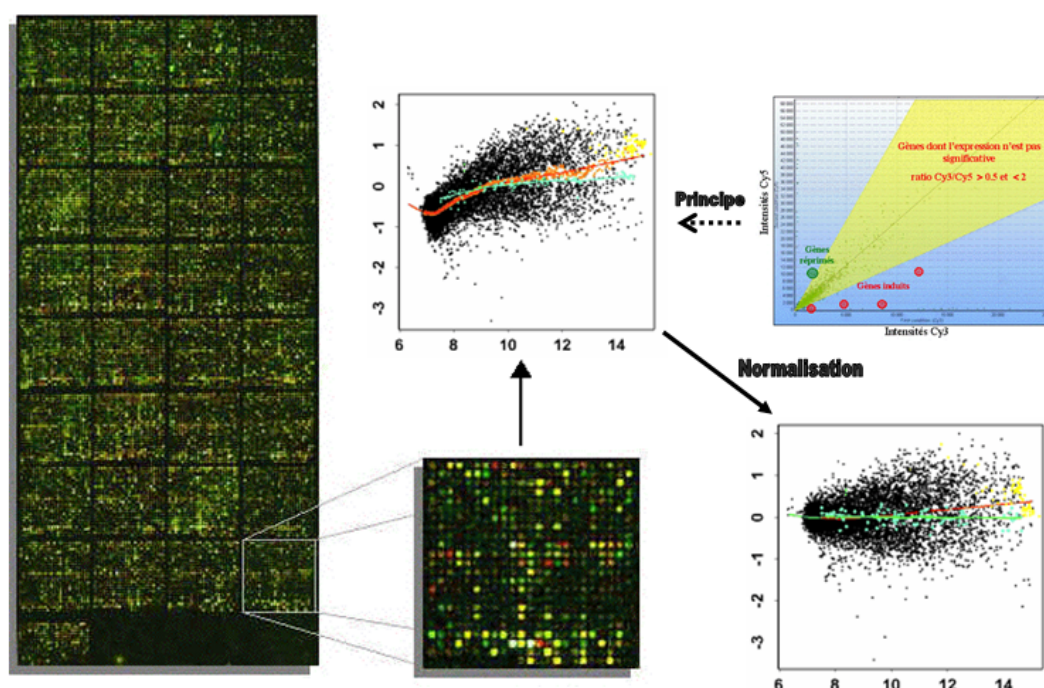


Figure I.6 – Exemple de la normalisation de données de puces.

trices. Donc, la taille de données peut être mesurée selon deux dimensions, le nombre de variables (ligne) et le nombre d'exemples (colonne). Pour la plupart des bases de données disponibles, les données d'expression sont très dissymétriques, le nombre de gènes est beaucoup plus élevé que celui des échantillons. Donc, il faudrait une réduction des dimensions pour sélectionner un sous-ensemble optimal de gènes. Deux grands types d'approche sont possibles pour la sélection d'attributs [LC08] : (1) les techniques *forward* où l'on démarre sans attribut et on ajoute au fur et à mesure les attributs qui seront estimés comme étant les plus pertinents ; (2) les techniques *backward* qui procèdent exactement à l'inverse : on démarre avec tous les attributs et on essaye d'éliminer ceux qui ne servent pas pour classer.

2.4.3 Discretisation


Définition 9. *Discretisation*

Certains algorithmes ne fonctionnent qu'avec des données catégorielles et surtout des données booléennes. Ainsi on devrait les transformer en variables discrètes par les méthodes de

discrétisation. Les méthodes de discrétisation sont nombreuses dans la littérature [DKS95]. Elles sont souvent adaptées aux contextes de données particuliers (données temporelles) ou le modèle de régulation utilisé par 2 valeurs (0 : *normal* et 1 : *sur-exprimé*) ou 3 valeurs (-1 : *sous-exprimé*, 0 : *normal* et 1 : *sur-exprimé*) .

	Expt. 1	Expt. 2	...
Gène 1	-2.32	-1.69	...
Gène 2	2.71	2.09	...
Gène 3	-1.21	-0.35	...
Gène 4	0.4	1.2	...
...

Discrétisation 3 niveaux:
1 = *sur-exprimé*
0 = *normale*
-1 = *sous-exprimé*



	Expt. 1	Expt. 2	...
Gène 1	-1	-1	...
Gène 2	1	1	...
Gène 3	-1	0	...
Gène 4	0	1	...
...

Figure I.7 – Exemple de la discretisation de données de puces.

3 Reconstruction de réseaux de régulation génétique

3.1 Introduction

3.2 Méthodes de reconstruction de réseaux de régulation génétique

D'après [SB07], il est possible d'établir une taxinomie des réseaux en fonction des critères suivants :

- *la description des entités biologiques* que l'on cherche à modéliser (par exemple, les facteurs de transcription, les sites de liaison,
- *la topologie du réseau*. Ce critère implique la définition du type d'interaction entre les entités biologiques,
- *la logique de contrôle*. Ce critère détermine comment se combinent les effets des noeuds sources sur un noeud cible du réseau,
- *le modèle dynamique*. Ce critère correspond à la modélisation temporelle du réseau et à la prédiction de la réponse des entités biologiques à certains stimuli.

3.2.1 Description des entités biologiques

Pour décrire les interactions entre des entités biologiques, le formalisme mathématique est choisi lorsqu'on a besoin de représenter précisément l'ensemble des réactions biochimiques impliquées dans la régulation de l'expression des gènes. Si on veut un modèle intuitif et plus interprétable, l'utilisation de représentation graphique est plus avantageuse.

La classe de réseaux qui nous intéresse se modélise sous la forme d'un graphe dont les noeuds sont des gènes ou des protéines et dont les arcs représentent une relation biologique.

3.2.2 Topologie du réseau

On trouve dans la littérature de nombreuses variétés de réseaux biologiques. Parmi ceux-ci, on peut citer :

- *les réseaux transcriptionnels* : les arcs sont dirigés ; le gène source est un facteur de transcription ; le gène cible est un gène activé par le gène source,
- *les réseaux d'interaction* : Les noeuds sont des protéines et les arcs non dirigés représentent des liaisons entre protéines (par exemple : la transduction de signal),
- *les réseaux de mutation* : les noeuds sont des gènes et l'arc dirigé indique une mutation entre une séquence parente et une séquence fille,
- *les réseaux de co-citation* : un arc non dirigé est étiqueté selon la fréquence de colocation des deux gènes dans la littérature scientifique.

3.2.3 La logique de contrôle

Réseau transcriptionnel

La topologie du réseau permet d'appréhender les relations entre entités biologiques, mais elle n'indique pas le comportement d'un noeud cible par rapport à ses noeuds sources. Concernant le réseau transcriptionnel, le modèle le plus simple compare la machinerie biologique à un circuit logique : le gène cible est activé selon une fonction booléenne, comme l'illustre la figure [I.8](#).

Ce modèle suppose une représentation des entités biologiques par des états discrets : par exemple, on peut associer une valeur booléenne au fait qu'un gène est exprimé. Une modélisation continue permet d'affiner le modèle.

Arbres de décision

Dans [\[SKB03\]](#), une modélisation par arbres de décision a été proposée. Dans ce modèle, l'expression d'un gène est déterminée par les valeurs d'expression de certains gènes, comme le

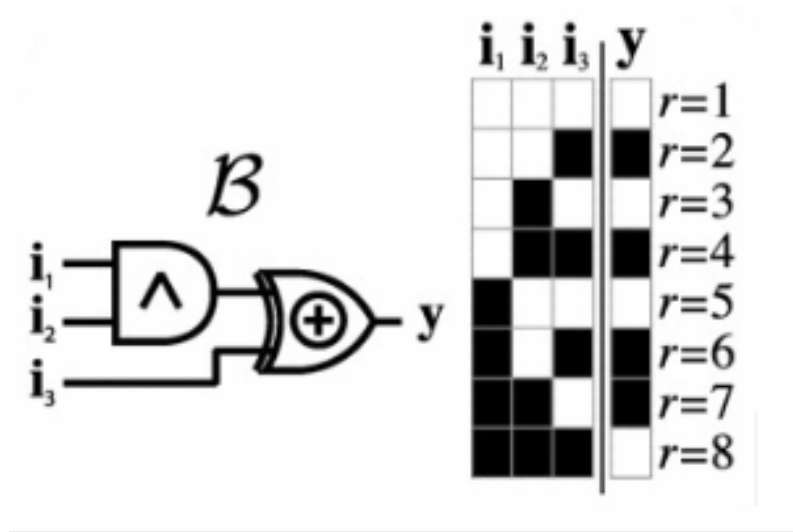


Figure I.8 – Modélisation de la logique de contrôle par une fonction booléenne. La réponse de y est définie par une fonction logique. Figure extraite de. [DRGB⁺10]

montre la figure I.9.

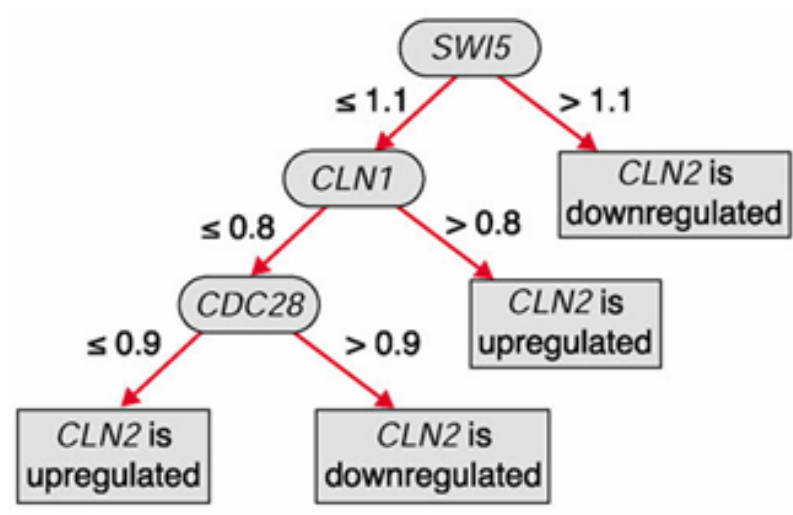


Figure I.9 – Arbre de décision du gène CLN2 (d'après [SKB03]). Dans cet exemple, la régulation du gène est prédite à partir des gènes SWI5, CLN1 et CDC28. Les étiquettes sur les arcs indiquent les seuils de valeur d'expression définis à partir de données de puces à ADN.

Réseaux de confiance

Les réseaux de confiance [BKK00, CBGB04] calculent l'information mutuelle entre les niveaux d'expression pour chaque couple de gènes, puis génère un réseau de type d'interaction gène-gène. L'hypothèse est *"si une association avec l'information mutuelle élevée signifie qu'un gène est non-aléatoirement associé à un autre ; alors les deux sont liés biologiquement"* ou *"deux gènes sont reliés si et seulement si ils sont co-exprimés"* (c'est la raison pour laquelle cette approche est aussi appelé "réseaux de co-expression").

Le principe est que deux gènes sont reliés si et seulement si ils sont co-exprimés. Butte et Kohane [BKK00] ont développé une méthode basée sur l'information mutuelle. A partir d'un test de permutation, ils considèrent que deux gènes sont co-exprimés si et seulement si leur information mutuelle est supérieure à l'information mutuelle maximale obtenue dans les données permutées. Ils montrent que les composantes connexes du graphe obtenu, appelées *relevance networks*, contiennent des gènes (co-régulés) avec des fonctions biologiques proches. Pour calculer l'information mutuelle, Butte and Kohane a discrétisé des valeurs de niveaux d'expressions. Carter et al. [CBGB04] se sont intéressés à la topologie de réseaux de co-expression des gènes. Ils utilisent la corrélation de Pearson pour mesurer la relation entre chaque paire de gènes et éliminent les interactions ayant une faible significativité.

Cette méthode prend en compte l'association biologiques des gènes. Toutefois, la discrétisation risque une perte d'information.

3.2.4 Le modèle dynamique

Ce modèle décrit comment évolue le réseau dans le temps. Il peut être absent de certains réseaux, purement statiques. Par exemple, l'arbre de décision de la figure X définit la régulation d'un gène en fonction de gènes informatifs, indépendamment de la notion de temps. De nombreux modèles dynamiques ont été proposés [Kau69, OGP02, BFG⁺04, KIM04, NRF04, RWSH07, ZC05].

3.2.5 Réseaux booléens

Stuart Kauffman [Kau69] a été l'un des premiers à modéliser la dynamique d'un réseau, en se basant sur le concept de réseau booléen. Dans ce modèle, on définit l'activité de chaque gène par une valeur binaire et on fait évoluer le réseau booléen par des pas de temps discret. Cette approche est synchrone : tous les gènes changent d'état simultanément.

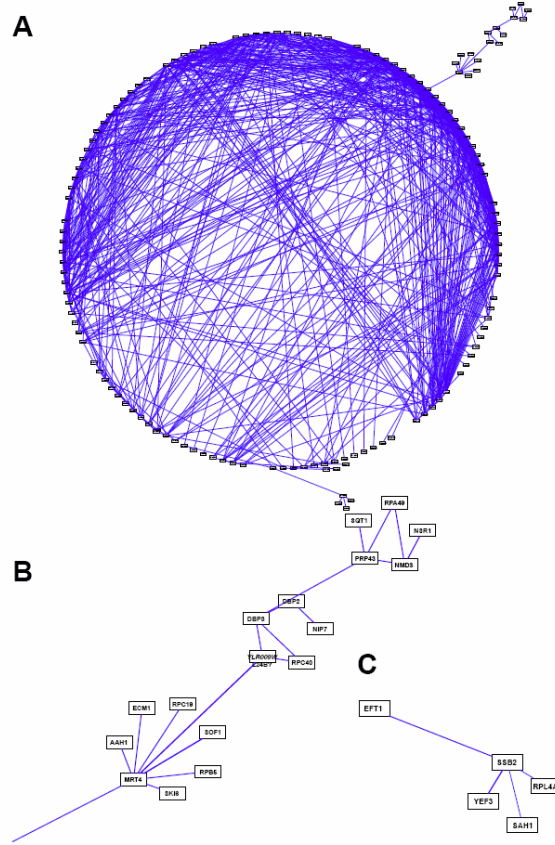


Figure I.10 – Exemple d’un réseau obtenu par l’approche basée sur les réseaux de confiance. (A) Le réseau le plus large possible. (B) et (C) Deux petites branches dans (A) en version ZOOM. [BKK00]

L’idée des réseaux booléens est inspiré ensuite par [AMK99, LSYH03, IEW02, RKLP⁺08] pour la modélisation des RRG est simple. On modélise un gène par une variable booléenne. Un gène est donc soit *sur-exprimé* (transcrit), soit *sous-exprimé* (non transcrit). On représente ensuite les influences (régulations) *positives* (activation) ou *négatives* (inhibition) d’un gène sur les autres par des fonctions booléennes qui détermine l’état d’un gène en fonction de l’état des certains autres gènes.

Plus précisément, c’est un graphe orienté $G = (V, F)$ où $V = v_1, v_2, \dots, v_n$ est un ensemble de noeuds et F est un ensemble de fonctions booléennes qui définit une topologie d’arêtes. n est appelé taille ou dimension du réseau. Un noeud représente un gène. A chaque noeud v , on associe une valeur booléenne $x(v)$ qui représente le niveau d’expression du gène correspondant. La valeur de x sera 0 si le gène est sur-exprimé (ou transcrit) et 1 si le gène est sous-exprimé

I.3 Reconstruction de réseaux de régulation génétique

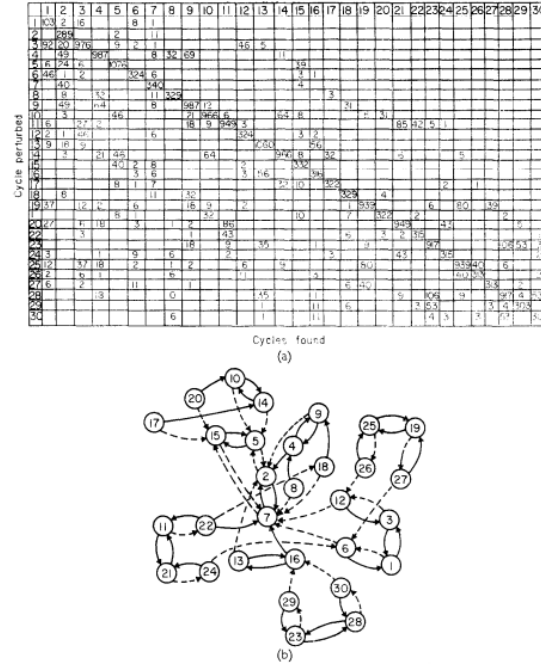


Figure I.11 – Exemple d’un réseau dynamique pour représenter un RRG avec des données temporelles [Kau69]. (a) Matrice de transition de niveau de régulation dans 30 tranches temporelles. (b) Transition entre des tranches présentées dans (a)

(ou non-transcrit).

L’avantage est la simplicité de la méthode. Cependant, d’après [LSYH03] cette méthode souvent accepte implicitement une hypothèse la le bruit-libre (noisy-free) de la mesure de données d’expression. Cela ne semble pas aux contextes des applications réelles.

Réseaux de Pétri

Les réseaux de Pétri [CRRT04, CKL05, MDNM00, SBSW07] sont des graphes bipartis utilisés pour la modélisation et le raisonnement sur les systèmes concurrentiels et distribués. Ils sont utiliser dans le contexte de RRG pour identifier la structure et les comportements dynamiques des RRG.

Cette méthodes prend en compte l’aspect dynamique de relations dans RRG. Cependant, c’est une approche basée sur l’expert et empirique, donc il semble infaisable avec des RRG de grande taille.

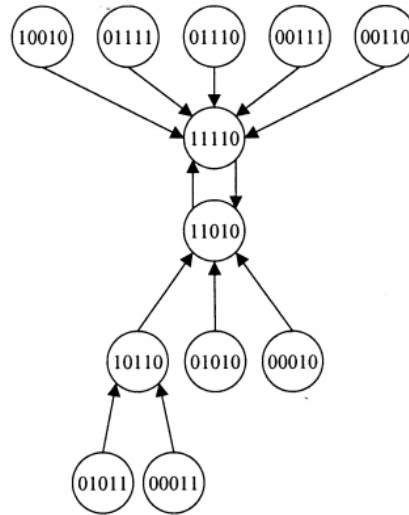


Figure I.12 – Exemple d’un réseau booléen simple pour représenter un RRG [IEW02].

Equations différentielles

L’approche la plus commune en termes de modélisation dynamique repose sur une formalisation physique des phénomènes biologiques. Un modèle utilisant les équations différentielles [CCK, HIM03] permet de prédire les valeurs de concentration des entités biologiques telles que les protéines, les molécules de signalisation, les ARN messagers.

Cette méthode considère un RRG comme un système d’équations différentielles. Le taux de changement d’une concentration particulière est donné par le fonction d’influence d’autres concentration RNA. Les concentrations de ces composants sont exprimées par des valeurs réelles positives évoluant dans le temps de manière continue. La variation de ces valeurs est décrite par une équation différentielle ayant comme paramètres les concentrations des molécules régulant l’entité étudiée. Ce type de modélisation nécessite la connaissance précise des concentrations des composants moléculaires ainsi que de leur cinétique, information malheureusement souvent difficile à obtenir.

Las approches basées sur les équations différentielles relient la valeur de chaque variable à la valeur de toutes les autres variables sous la forme d’une équation (voir la Figure ??). Ces équations peuvent être linéaires, non linéaires, et/ou des équations différentielles. Les interactions putatives sont identifiées par la résolution d’un ensemble d’équations incluant des poids comme paramètres. Ces pondérations représentent l’influence de chaque variable sur les

I.3 Reconstruction de réseaux de régulation génétique

autres. Généralement, seuls quelques poids dans l'équation pour une seule variable diffèrent considérablement de zéro et sont donc considérés comme des influenceurs putatifs.

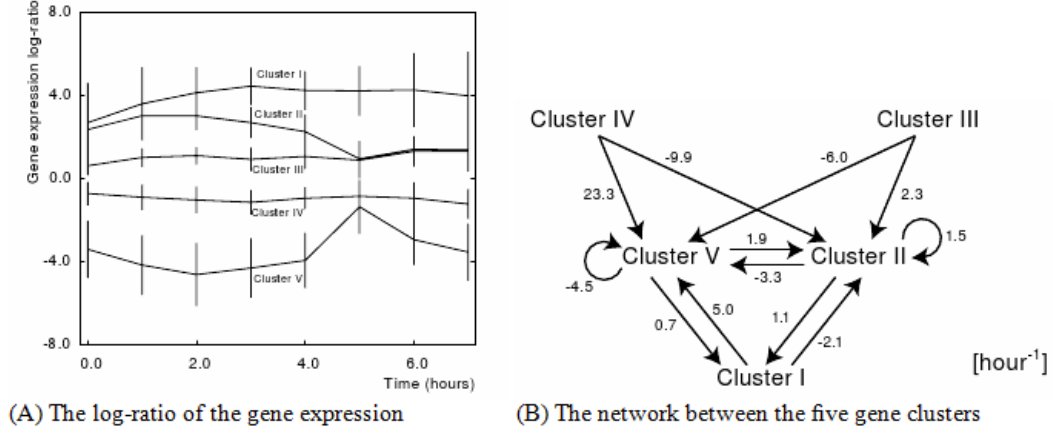


Figure I.13 – Exemple d'un réseau obtenu par l'approche basée sur les équations différentielles.[?]

L'avantage de cette approche est la précision du calcul. Pourtant, elle est limitée à une fonction simple à cause de la complexité de paramètres dans le cas où le nombre de variables est grand. En plus, cette méthode demande des données temporelles pour apprendre des paramètres [CCK].

Par ailleurs, cette modélisation ne tient pas compte des fluctuations aléatoires du système biologique qui peuvent induire des effets non négligeables sur les concentrations moléculaires. Des auteurs ont suggéré d'utiliser des modèles stochastiques [PMR11] pour modéliser la dynamique des réseaux. Ce type de modèle définit la probabilité qu'une variable atteigne un état donné en fonction des états des molécules à un instant donné. Pour ce type de modélisation, les réseaux Bayésiens sont un des outils le plus répandus qui peuvent capturer les relations linéaires, non-linéaires, ou stochastiques...

3.2.6 Réseaux bayésiens

Les réseaux Bayésiens (RB) sont utilisés dans de nombreux domaines comme des outils de modélisation. En effet, ils sont un des représentants les plus connus des modèles graphiques probabilistes [NWL⁺07]. Ils rendent particulièrement intéressants la représentation de systèmes complexes. Pour la modélisation de réseaux de régulation génétique : (i) la partie graphique de RB donne un outil visuel qui indique non seulement les régulations entre les gènes, mais aussi la causalité de ces régulations ; (ii) la partie probabiliste permet de quantifier ces régulations.

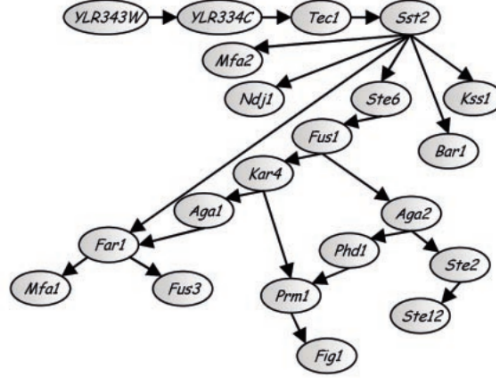


Figure I.14 – Exemple d'un RB obtenu par la reconstruction du RRG présenté dans [Fri04].

C'est une technique de modélisation de systèmes comportant un grand nombre de variables, au moyen de relations mesurées par la fonction de distribution des probabilités conditionnelles (paramètres) et présentées sur un graphe acyclique orienté (structure). L'apprentissage de RB à partir de données et de connaissances a priori offre un cadre théorique et méthodologique pour relier les problématiques de modélisation, d'identification, de simulation du fonctionnement, de l'interaction entre des gènes au sein d'un organisme. On peut citer quelques propositions de modélisation de réseaux de régulation génétique par RB depuis ces dernières années : [? Fri04, AFGdB08].

Cette présentation des différentes modélisations de réseaux montre une grande variété de modèles, qui diffèrent selon les catégories suivantes :

- approche statique/dynamique
- approche discrète/continue
- approche déterministe/stochastique

Remarque 1. Les RB possède une importante classe d'outils d'apprentissage de la structure à partir de données. Cet avantage présente en particulier des forts intérêts par rapport aux autres méthodes qui n'ont pas suffisamment de moyens de calcul ou de qualification des relations de variables. Le détail de l'apprentissage de la structure de RB est présenté dans la partie suivante. Les RB apportent un cadre formel pour la modélisation de réseaux génétiques. Ce mémoire portant sur cette approche, la section suivante expose plus particulièrement celle-ci.

4 Conclusion

Nous avons présenté dans cette partie de différentes approches pour la modélisation de réseaux de régulation génétique, pour l'apprentissage de la structure de réseaux Bayésiens, un des modèles graphiques probabilistes les plus performants, et pour l'étude différentielle de réseaux Basésiens. Les réseaux Bayésiens dégagent un fort intérêt dans la modélisation de réseaux de régulation génétique. L'approche évolutionnaire permet d'obtenir un ensemble de meilleures structures de réseaux Bayésiens. Les approches existantes dans la littérature pour l'étude différentielle de réseaux Bayésiens présentent des avantages mais aussi des inconvénients qui demandent une recherche approfondie pour résoudre les problèmes posés. Les contributions présentées dans la suite de cette thèse sont différentes solutions théoriques et expérimentales pour ces problèmes.

Quelles que soient les approches utilisées pour modéliser les réseaux de régulation génétique à partir de biopuces, la difficulté principale consiste à trouver la meilleure structure qui possède à la fois la *robustesse* (maximisation de précision de l'inférence avec une bonne précision sur toute quantité/qualité de données) et la *simplicité* (minimisation de relations entre variables).

Références bibliographiques

- [AFGdB08] C. Auliac, V. Frouin, X. Gidrol, and F. d’Alche Buc. Evolutionary approaches for the reverse-engineering of gene regulatory networks : A study on a biologically realistic dataset. *BMC Bioinformatics*, 9(1) :91, 2008. [18](#)
- [AMK99] T Akutsu, S Miyano, and S Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pacific Symposium On Biocomputing*, 28(4) :17–28, 1999. [14](#)
- [BAR04] Vincent BARRA. *Modélisation, classification et fusion de données biomédicales*. PhD thesis, Université Blaise Pascal - Clermont-Ferrand II, 2004. [iii](#), [7](#)
- [BFG⁺04] Matthew J. Beal, Francesco Falciani, Zoubin Ghahramani, Claudia Rangel, and David L. Wild. A bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, 21(3) :349–356, 2004. [13](#)
- [BKK00] Atul J. Butte, Isaac S. Kohane, and I. S. Kohane. Mutual information relevance networks : Functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 5 :415–426, 2000. [iii](#), [13](#), [14](#)
- [CBGB04] Scott L. Carter, Christian M. Brechbuhler, Michael Griffin, and Andrew T. Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14) :2242–2250, September 2004. [13](#)
- [CCK] Zeke S. H. Chan, Lesley Collins, and Nikola K. Kasabov. Bayesian learning of sparse gene regulatory networks. *Biosystems*. [16](#), [17](#)
- [CKL05] J.P. Comet, H. Klaudel, and S. Liauzu. Modeling multivalued genetic regulatory networks using high level petri nets. In *G. Ciardo and P. Darondeau (eds), Proc. of the Int. Conf. on the Application and Theory of Petri Nets, Lecture Notes in Computer Science 3536*, pages 208–227. Springer–Verlag, 2005. [15](#)
- [CRRT04] Claudine Chaouiya, Elisabeth Remy, Paul Ruet, and Denis Thieffry. Qualitative modelling of genetic networks : From logical regulatory graphs to standard petri nets. In Jordi Cortadella and Wolfgang Reisig, editors, *Applications and Theory of Petri Nets 2004*, volume 3099 of *Lecture Notes in Computer Science*, pages 137–156. Springer Berlin – Heidelberg, 2004. [15](#)

Références bibliographiques

- [DKS95] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In *International Conference on Machine Learning*, pages 194–202, 1995. [10](#)
- [DRGB⁺10] Jeroen De Ridder, Alice Gerrits, Jan Bot, Gerald De Haan, Marcel Reinders, and Lodewyk Wessels. Inferring combinatorial association logic networks in multimodal genome-wide screens. *Bioinformatics*, 26(12) :i149–i157, 2010. [iii](#), [12](#)
- [Fri04] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303 :799–805, 2004. [iii](#), [18](#)
- [GWLC07] J. F. Griffiths, S. R. Wessler, R. C. Lewontin, and S. B. Carroll. *An Introduction to Genetic Analysis*. 2007. [1](#)
- [HIM03] Michiel De Hoon, Seiya Imoto, and Satoru Miyano. Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations. In *Pac. Symp. Biocomput*, pages 17–28, 2003. [16](#)
- [IEW02] S. Ilya, R.D. Edward, and Z. Wei. From boolean to probabilistic boolean networks as models of genetic regulatory networks. In *Proceedings of The IEEE*, volume 90, pages 1778–1792, 2002. [iii](#), [14](#), [16](#)
- [JM61] F. JACOB and J. MONOD. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3 :318–356, 1961. [3](#)
- [Kau69] S. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3) :437–467, March 1969. [iii](#), [13](#), [15](#)
- [KIM04] Sunyong Kim, Seiya Imoto, and Satoru Miyano. Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems*, 75(1-3) :57 – 65, 2004. `<ce :title>Computational Systems Biology</ce :title>`. [13](#)
- [LC08] Kim-Anh (2008) Lê Cao. *Outils statistiques pour la sélection de variables et l'intégration de données "omiques"*. PhD thesis, Institut National des Sciences Appliquées de Toulouse., 2008. [9](#)
- [LSYH03] Harri Lahdesmaki, Ilya Shmulevich, and Olli Yli-Harja. On learning gene regulatory networks under the boolean network model. *Machine Learning*, 52 :147–167, 2003. 10.1023/A :1023905711304. [14](#), [15](#)
- [MDNM00] H. Matsuno, A. Doi, M. Nagasaki, and S. Miyano. Hybrid petri net representation of gene regulatory network. In *Pacific Symp. on Biocomputing*, 2000. [15](#)
- [NRF04] I. Nachman, A. Regev, and N. Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20(suppl 1) :i248–i256, 2004. [13](#)

- [NWL⁺07] P. Naïm, P. Wuillemin, P. Leray, O. Pourret, and A. Becker. Les réseaux bayésiens. *3e Édition Eyrolles*, 2007. [17](#)
- [OGP02] Irene M. Ong, Jeremy D. Glasner, and David Page. Modelling regulatory pathways in e. coli from time series expression profiles. *Bioinformatics*, 18(suppl 1) :S241–S248, 2002. [13](#)
- [PMR11] Loïc Paulevé, Morgan Magnin, and Olivier Roux. Refining dynamics of gene regulatory networks in a stochastic π -calculus framework. In Corrado Priami, Ralph-Johan Back, Ion Petre, and Erik de Vink, editors, *Transactions on Computational Systems Biology XIII*, volume 6575 of *Lecture Notes in Computer Science*, pages 171–191. Springer Berlin / Heidelberg, 2011. [17](#)
- [Qua02] J. Quackenbush. Microarray data normalization and transformation. *Nature*, 32 :123–148, 2002. [8](#)
- [RKLP⁺08] A. S. Ribeiro, S.A. Kauffman, J. Lloyd-Price, B. Samuelsson, and J. E. S. Socolar. Mutual information in random boolean models of regulatory networks. *Phys. Rev. E*, 77(1) :011901, Jan 2008. [14](#)
- [RWSH07] Henning Redestig, Daniel Weicht, Joachim Selbig, and Matthew Hannah. Transcription factor target prediction using multiple short expression time series from arabidopsis thaliana. *BMC Bioinformatics*, 8(1) :454, 2007. [13](#)
- [SB07] Thomas Schlitt and Alvis Brazma. Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 8(Suppl 6) :S9, 2007. [10](#)
- [SBSW07] L. Jason Steggles, Richard Banks, Oliver Shaw, and Anil Wipat. Qualitatively modelling and analysing genetic regulatory networks : a petri net approach. *Bioinformatics*, 23(3) :336–343, 2007. [15](#)
- [SKB03] Lev A. Soinov, Maria A. Krestyaninova, and Alvis Brazma. Towards reconstruction of gene networks from expression data by supervised learning. *Genome biology*, 4(1), 2003. [iii](#), [11](#), [12](#)
- [ZC05] Min Zou and Suzanne D. Conzen. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1) :71–79, 2005. [13](#)