Université d'Evry Val d'Essonne

Anné Universitaire 2008 / 2009

Master 2 Génie Biologique et Informatique

Gabriel Chandesris

Projet Bibliographique

Quantitative inference of dynamic regulatory pathways via microarray data

Wen-Chieh Chang, Chang-Wei Li and Bor-Sen Chen

Prédiction quantitative des voies de régulation dynamiques par l'intermédiaire des données de microarray

Tutrice universitaire : `Florence d'Alché-Buc`

Projet bibliographique - soutenance le 21 novembre 2008

Table des matières

Abstract

The prediction, or inference, of gene interaction networks is part of the current work in biology, to understand the metabolic work inside cells. The objective of this project is to present a bibliographic work in this field by Chang and his staff until 2005 and presented in the article *Quantitative inference of dynamic regulatory pathways via microarray data* (BMC Bioinformatics, 2005).

The mathematical approach of Chang is explained, with its results and improvements of the prediction algorithm. Other approaches and methodologies for prediction of gene interaction networks are presented in this paper, both post and earlier approaches are presented in the context of developments in this research area, some explaination about mathematical tools and representation are given (graphs, bayesian networks, statistical analysis and dynamic analysis).

Résumé

La prédiction, ou inférence, des réseaux d'interaction géniques fait partie des travaux actuels en biologie, notamment dans la compréhension des mécanimes métaboliques de fonctionnement des cellules. L'objectif de ce projet bibliographique est de présenter les travaux effectués dans ce domaine par Chang et ses collaborateurs jusqu'en 2005 et présentés dans l'article *Quantitative inference of dynamic regulatory pathways via microarray data* (BMC Bioinformatics, 2005).

L'approche mathématique de Chang est explicitée, ainsi que ses résultats et les améliorations possibles de l'algorithme de prédiction. D'autres approches et méthodologies de prédiction des réseaux d'interaction géniques sont présentées dans ce mémoire, tant les approches antérieures que postérieures dans un contexte d'évolution des recherches dans ce domaine et d'explicitation des outils mathématiques et de représentation nécessaire à ce type de prédiction (graphes, réseaux bayésiens, analyse statistique et analyse dynamique).

1 Introduction

La recherche du fonctionnement du métabolisme cellulaire est l'une des principales voies de recherche actuelle en biologie, notamment par l'analyse et la compréhension du fonctionnement des voies métaboliques, mais aussi par la compréhension des mécanismes de régulation au niveau des gènes. L'analyse des voies de régulation des gènes permet de mieux comprendre certains aspects de régulation cellulaire, tant métaboliques que génétique, au cours du cycle cellulaire.

Dans l'article *Quantitative inference of dynamic regulatory pathways via microarray data* (BMC Bioinformatics, 2005), Chang et ses collaborateurs [7] présentent un algorithme pour analyser les voies de régulation à partir de données de puces à ADN (qui seront appelées microarray par la suite). L'objectif ici est de présenter leurs travaux, ainsi que d'autres travaux reliés à ceux-ci, sur lesquels l'équipe de Chang a pu comparer ses résultats ou d'autres équipes qui ont utilisés des méthodes analogues.

Dans ce rapport, le contexte biologique est d'abord explicité afin de donner des éléments sur l'exploration des voies métaboliques et les réseaux de signalisation cellulaire, les éléments de mesures pour cette exploration et l'approche utilisée ; ensuite les différents outils existants sont doécrits dans un état de l'art autour de la méthode utilisée par Chang *et al.* [7].

---

2 Contexte biologique

2.1 Voies et réseaux de signalisation cellulaires

Les interactions intra-cellulaires entre les gènes sont une voie de signalisation à la base de la recherche en génomique fonctionnelle. Ceci concerne également la façon dont les gènes se régulent et s'influencent entre eux via des accroches transcriptionnelles ou des interactions physiques. Cela constitue l'un des principaux axes de recherche en biologie cellulaire et sur le fonctionnement des organismes.

Dans un cadre de systématique et de modélisation dynamique, une recherche peut être effectuée afin de comprendre le fonctionnement des voies et réseaux de régulation [31], afin de mieux comprendre le système de régulation intracellulaire et intercellulaire de régulation des gènes via une recherche des gènes exprimés (séquences d'ADN codant les protéines présentes au cours de la dynamique [37]).

2.2 Mesure synchronisée

On peut analyser le fonctionnement cellulaire à un instant donné de l'expression génique (transcriptome) avec les microarray dans des conditions expérimentales bien définies [29]. Une analyse temporelle peut être effectuée de cette façon, par une succession d'analyses à des temps différents dans les mêmes conditions expérimentales.

C'est une mesure quantitative efficace pour prédire l'évolution et le fonctionnement d'un réseau de régulation telle que décrite par Wen-Chieh Chang et ses collaborateurs [7]. Pour cela différentes données issues de banques de données de stockage des données de microarray ont été utlisées, comme Stanford MicroArray Database [30], Gene Expression Omnibus [13] et ArrayExpress [4].

2.3 Approche dynamique systématique

Cette approche permet l'exploration des relations causales entre gènes, afin d'étudier les voies de régulation et de signalisation au sein d'un génome. C'est cette méthodologie qui est principalement utilisée par Chang *et al.* en 2004 [7]. Le principe de l'approche dynamique est d'utiliser la mesure synchronisée citée plus haut (un ensemble de données de microarray échelonnées dans le temps).

---

3 État de l'art : méthodologies et approches

La grande quantité de données produites en biologie, aunsi que les méthodes d'analyses adéquates nécessitent l'utilisation d'un outil automatique et rapide (informatique). Différentes voies de recherches existent pour la prédiction de voies de régulation, utilisant différents aspects de modélisation [5, 6, 10, 12, 21, 27], notamment pour la recherche sur des données de microarray : les méthodes diffèrent essentiellement par les méthodes de modélisation et de calcul utilisées pour retrouver et prédire les réseaux de régulation.

3.1 Modes de recherches proches sur les réseaux de régulation

3.1.1 Dynamique et principes des réseaux métaboliques

Chin et ses collaborateurs [10] ont étudié les principes de conception des réseaux de régulation et de leur dynamique. Cette étude a été réalisée sur des voies métaboliques communes à différentes espèces, comme par exemple les voies de biosynthèse d'acides aminés (leucine, valine, isoleucine) chez *Escherichia coli* et *Saccharomyces Cerevisiae*.

Différents éléments ont été trouvés pour relier la production de protéines, les facteurs de transcription inclus dans ces voies métaboliques et établir un lien entre métabolisme et expression des gènes. Un modèle théorique a été établi en même temps que la mesure de l'abondance de protéines chez plusieurs espèces, afin de pouvoir trouver des éléments généraux de la dynamique de l'expression génétique dans différentes voies métaboliques [20, 26].

Les résultats présentés par Chin *et al.* [10] regroupent notamment des indications concernant certains principes de modélisation des voies métaboliques et de leur régulation, pour la recherche et l'analyse de données pour établir ces modèles.

- Utilisation d'un système automatisé de mesure temporelle de l'abondance des protéines.
- Mesure de la réponse différentielle du gène en aval et en amont du point de contrôle.
- Distinguer les différents profils dynamiques causés par la régulation différentielle.
- Modéliser la dynamique d'induction des gènes.
- Modéliser l'induction des gènes en amont.
- Prédire les réponses dynamiques aux perturbations.

L'analyse du lien entre la dynamique et l'architecture en réseau du métabolisme a permis de trouver différents systèmes de régulation, liés à la présence (ou à l'absence) de certaines molécules pour activer ou inhiber les voies métaboliques étudiées. La relative simplicité de ces voies a mené à la construction d'un modèle permettant des prédictions, confirmées expérimentalement.

3.1.2 Réseaux génétiques et données de microarray

Djebbari et Quackenbush [12] ont travaillé sur la problématique d'analyse de données en grande quantité, issues de microarray, pour construire des réseaux génétiques. L'objectif étant de relier les gènes entre eux plutôt que des données seulement issues de traduction (protéines). La méthodologie des réseaux bayesiens est dérivée de l'utilisation des graphes (détaillée plus loin [21]), avec un complément d'utilisation de probabilités conditionnelles de liens entre noeuds.

Tout d'abord utilisée par Friedman *et al.* [15], cette méthodologie a conduit à construire un modèle prédictif du système métabolique de la levure. L'application de la méthode des réseaux bayésiens est relativement complexe en génomique, mais l'application à des connaissances spécifiques permet de résoudre certaines topologies de réseaux, malgré l'introduction de biais. De plus, une analyse temporelle de données de microarray permet de mieux organiser les données et trouver un modèle optimal.

Différentes étapes sont distinguées afin de construire les réseaux de gènes par réseaux bayésiens : tout d'abord il s'agit d'établir une structure en réseau probable à partir de données publiées, puis d'établir une structure en réseau à partir de données d'analyse expérimentale complémentaire, ensuite le réseau bayésien est construit et les arcs du graphe sont orientés, enfin une étape de *bootstrap* (ou re-échantillonage) est effectuée pour confirmer la structure du réseau à partir des données (ou de partition des données).

La validation et la confirmation de cette méthodologie s'effectuent par l'utilisation d'une base de données de relations et d'interactions connues, comme KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [23], malgré les limitations aux découvertes et recherches faites jusqu'à présent. D'autres limites de la méthodologie d'utilisation des réseaux bayésiens existent sur la structuration du modèle, biais de l'étude des gènes et non des liens physiques ; malgré cela, certaines idées de conception de modèle sont utilisées comme approches de construction de réseaux génétiques.

3.2 Différentes approches d'analyses de données

Différentes approches d'analyses de données de microarray existaient déjà auparavant, notamment dans la description des différentes méthodes d'analyses de données et de prédiction de réseau de régulation génétique. On peut citer notamment des approches d'analyses par scores informatifs [5] ou analyse statistique [27], et des méthodes de prédiction [6] avec par exemple des applications de la théorie des graphe à la biologie [21].

3.2.1 Analyse des données complexes de transcription par des scores informatifs

Bussemaker et ses collaborateurs [5] ont analysé les réponses transcriptionnelles complexes en utilisant des scores de niveau de parcours, ces scores étant basés sur des informations connues des gènes : fonction codée, protéine traduite ou interaction entre des facteurs de régulations et la séquence de contrôle de l'expression.

Deux approches ont été considérées :
- La première méthode utilise une comparaison entre le niveau d'expression d'un ensemble de gènes par rapport à l'ensemble du génome.
- La seconde méthode part d'une estimation d'un réseau de régulation sur l'ensemble du génome créé à partir de données connues (séquence, interaction) et utilise une analyse de régression pour estimer les différentes voies de régulation génétiques.

Dans ces approches, différentes utilisations de données sur l'expression différentielle ont été faites, tant sur des données qualitatives (appartenance ou non d'un gène à un ensemble) que quantitatives (log-ratio d'expression sur ChIP-chip [18]) ; différentes méthodes pour déterminer l'appartenance à une voie de régulation ou à une fonction existent :

- Sur-représentation d'un ensemble de gènes prédéfini.
- Attribution de scores de niveaux d'expression d'un ensemble de gènes.
- Au-delà des ensembles de gènes : analyse de régression.

L'utilisation des deux approches citées [5] permet de réduire la taille du problème de l'analyse et d'obtenir des groupes de gènes selon les conditions d'expression et selon des profils d'expression analysables.

3.2.2 Réseaux de régulation cellulaires

Carter [6] présente des méthodes d'analyses et les objectifs associés pour comprendre et modéliser les données d'interactions en réseaux d'interaction pour mieux comprendre les systèmes biologiques. Différentes données (microarray de protéines, séquençage de génomes...) sont analysables afin de trouver des motifs de reconnaissances (interactions ADN-protéines), notamment chez la levure *Saccharomyces cerevisiae* pour détecter la suppression de souches mutantes ou encore le nématode *Caenorhabditis elegans* dans le cadre de modèles sur l'apoptose.

D'autres approches existent et étudient de façon abstraite (méthodes Bayesiennes) les ensembles d'interactions afin de regrouper les éléments et rechercher des motifs. L'étude des interactions génétiques a aussi été faite afin de permettre un recoupement de réseau de régulation entre différentes espèces et de phénotypes observés, notamment du fait de l'existence de nombreux mutants chez certaines espèces (levure, nématode, ou la mouche *Drosophilia melanogaster*, dont les données et annotations sont accessibles).

Les implications pour la modélisation en biologie sont de permettre une analyse et un recoupement des informations, notamment par l'utilisation d'outils d'acquisition et de représentation des données. Quelques éléments d'intérêts étaient notés comme une évolution possible [6] pour le développement de modèles : la prédiction et la détection d'interactions moléculaires, un consensus sur un ensemble de standard de détection de phénotypes et d'ontologies, des mesures à haut débit quantitatives en lien avec des perturbations génétiques connues...

3.2.3 Biologie moléculaire et graphes

Huber, Carey et leurs collaborateurs [21] explicitent l'usage des graphes en biologie moléculaire, notamment pour l'analyse de données dans ce domaine ils en font une application pour l'intégration de données d'interactions protéine-protéine et les réseaux de co-expression. Un graphe étant défini comme un ensemble de noeuds et d'arcs, les noeuds étant les éléments d'intérêt, ici des gènes ou des protéines, et les arcs, les relations entre éléments d'intérêt.

L'utilisation d'un graphe permet de définir un réseau de façon compréhensible et d'une représentation aisée (si celui-ci n'est pas excessivement complexe). Un ensemble de connaissances peut ainsi être représenté. Ceci est applicable pour la régulation, la transduction de signal, les réseaux métaboliques ou l'ontologie.

Différents aspects de l'utilisation des graphes sont présentés, notamment l'exemple de Gene Ontology [1] (pour regrouper les données sous des termes spécifiques), ou Bioconductor [16] (pour l'analyse et la compréhension de données en biologie), et également des éléments de calcul d'adjacence entre différents éléments issus de l'analyse de données de transcription et d'interaction.

L'utilisation des graphes permet de relier des données de biologie moléculaires mais également entre les protéines et les gènes d'intérêts par des données connues, d'annotation notamment. Ceci est effectué dans un objectif de compréhension, notamment par une visualisation, ou la recherche de solutions optimales et de suppression du bruit de fond des données d'observation.

3.2.4 Aspects statistiques de la régulation des gènes

Michaël Lässig [27] introduit un ensemble de méthodes d'analyses statistiques pour connaitre les fonctions génomiques des gènes au sein des réseaux de gènes. Dans ce cadre, il a déterminé un certain nombre d'aspects théoriques de la régulation de la transcription, selon certains principes :

1 La transcription est un **processus biophysique**. Ce processus est une interaction ADN-protéines en certains loci, proches du gène régulé.
2 Sachant que les protéines trouvent leurs sites fonctionnels, il est possible de prédire ces interactions en construisant le réseau de régulation et en identifiant les loci fonctionnels par des approches statistiques ou des algorithmes, par des **méthodes bioinformatiques**.
3 Les réseaux de régulation sont devenus une part importante de l'analyse de **la différentiation des organismes et de l'évolution en biologie** [32, 34], du fait des connaissances sur la régulation entre les gènes.

Les différents facteurs qui interviennent dans ces éléments de régulation sont nombreux, mais un certain nombre de recherches en génomique ont permis de les modéliser autour d'un ensemble d'idées convergentes dans différentes disciplines, autour des trois aspects donnés ci-dessus.

- Concernant l'**approche biophysique de la régulation**, différents éléments ont été retenus comme :
  * l'énergie de liaison entre les facteurs de transcription et l'ADN ainsi que les aspects thermodynamiques,
  * la distribution de cette énergie sur le génome et la cinétique des facteurs de transcription,
  * la sensibilité de la régulation, liée à certains déterminants génomiques et leur évolution [17].

- Concernant l'**approche bioinformatique**, les méthodes retenues reposent notamment sur l'utilisation de modèles de Markov sur les séquences, de modèles probabilistes (sites fonctionnels), de modèles Bayesiens (loci génomiques) et d'analyse des séquences par programmation dynamique.

- Concernant l'**approche sur l'évolution**, les différentes méthodologies utilisées sont les suivantes :
  * la détermination de la dynamique de population (de façon déterministe ou stochastique),
  * l'inclusion du processus de mutation et d'équilibre d'évolution dans la dynamique de population,
  * d'autres effets inclus dans une dynamique stochastique des populations, comme le changement d'état ou l'équilibre entre différents mutants, voire un effet de sélection,
  * la mesure de la sélection et de la conservation des sites de liaison (un effet stationnaire ou adaptatif est possible sur ces sites [3]),
  * une corrélation avec la fréquence des nucléotides.

L'ensemble de ces différentes techniques et méthodes permet de construire une dynamique des génomes, notamment au travers des interactions évolutives, les interactions au cours du temps et les interaction entre gènes.

4 Prédiction de la régulation d'un ensemble de gènes

Le mécanisme de régulation d'un gène cible est étudié en remontant la chaîne causale [7], par une analyse statistique et différentielle de données obtenues à partir de microarray. À partir d'un profil d'expression et un modèle dynamique du gène cible, une estimation de la régulation en amont est effectuée. En étudiant cette régulation en amont (fonction mathématique) les gènes responsables de la régulation en amont sont trouvés avec leurs possibilités de régulation ainsi que leur mode d'activation ; ceci permet de les relier entre eux au sein d'une même voie de régulation.

4.1 Profil d'expression et modèle dynamique

L'analyse des données de microarray s'effectue sur un ensemble ordonné temporellement, ce qui permet d'obtenir une dynamique sur le modèle obtenu (action temporisée de la régulation d'un gène sur un autre).

La systématique utilisée ici est d'utiliser cette succession de données pour connaître des informations d'activités de transcriptions de gènes au cours du temps. Ainsi, à partir d'une information sur une activité transcriptionnelle d'un gène à un moment M_T, on peut déterminer les gènes précédemment actifs au moment M_{T-1}, et éventuellement trouver un lien de causalité entre l'activité d'un gène A et l'activation ou l'inactivation d'un gène B. Dans ce dernier cas, l'activation d'un gène A en M_{T-1} entraîne une baisse de transcription du gène B en M_T.

La prédiction est affinée par un calcul prédictif sur plusieurs ensembles de données concernant les mêmes activités, mais concernant une autre expérimentation sur une même espèce, les voies de régulations étant les mêmes.

4.2 Régulation en amont, analyse itérative

De façon itérative, chacun des gènes situés en amont est considéré comme un gène cible pour trouver leur régulation en amont. La construction des voies de régulation (réseau de régulation) est ainsi possible pour l'ensemble du génome. C'est à dire que l'on effectue tout d'abord la prédiction à une étape donnée M_T sur un gène cible C, afin d'obtenir des gènes à l'origine de la régulation du gène C, ces gènes peuvent être notés dans un ensemble B_C ; l'étape suivante concernera la prédiction des gènes régulateurs en amont de chaque gène dans B_C (chacun des gènes trouvés devient un gène cible).

La prédiction sur quelques voies de régulation (par l'utilisation de profils quantitatifs) est également effectuée pour confirmer quelques voies peu ou pas connues dans un réseau de régulation [7].

4.3 Validation de l'analyse temporelle des données de microarray

Le lien de causalité est confirmé par l'analyse de différentes données expérimentales, ordonnées dans le temps. Une validation de la méthode est effectuée en arrangeant ces données de façon aléatoire (perte de l'ordonnancement dans le temps), afin de vérifier que la dynamique causale est liée uniquement au déroulement dans le temps et non au hasard.

---

5 Algorithme, données expérimentales, résultats et analyse

La méthode d'analyse dans cette approche quantitative est un algorithme centré sur les capacités de régulation des gènes identifiés en amont d'un gène cible, et leur délai de régulation des gènes situés en aval. Ainsi un ensemble de voies de régulation peut être construit à partir de données de microarray. Dans cette étude, deux voies de signalisation sont reconstituées (voies de régulation circadiennes chez *Arabidopsis thaliana* et la voie de régulation métabolique du passage de la fermentation à la respiration chez *Saccharomyces cerevisiae*). Ceci afin d'évaluer la performance de la méthode proposée.

Les données expérimentales utilisées sont celles de ces deux organismes modèles bien connus et étudiés, pour lesquels des données de microarray sont disponibles et analysables pour l'approche utilisée par Chang et ses collaborateurs pour la prédiction quantitative [7]. Des modèles sont construits autour de ces données afin de prédire correctement les voies de régulation dynamiques.

5.1 Méthode de construction des modèles

L'algorithme proposé par Chang et ses collaborateurs comprend essentiellement quatre étapes :

1 Description du modèle dynamique avec équation différentielle et établissement du profil d'expression pour chaque gène.
2 Extraction des fonctions de régulation en amont, depuis le profil d'expression du gène cible, avec une méthode d'estimation optimale.
3 Estimation de la fonction de régulation, pour aider à chercher le signal de régulation corrélé, et reconstitution itérative de l'ensemble de la voie de régulation en amont.
4 Application de filtres biologiques (connaissances et éléments biologiques connus) pour affiner la construction des voies de régulation du signal, et améliorer pertinence de la solution proposée...

5.1.1 Description du système dynamique du modèle

Le modèle de régulation de régulation du signal a été déterminé comme une équation différentielle du second ordre est utilisée pour décrire le profil du gène à la position i en un moment t :

$$\ddot{X}_i(t) + a_i\dot{X}_i(t)_i + b_iX_i(t) = G_i(t) + \varepsilon_i(t) \tag{1}$$

Les paramètres a_i et b_i caractérisent les propriétés de la dynamique interne de la transcription du gène i (oscillation et dégradation). Le bruit lié aux données courantes est indiqué par $\varepsilon_i(t)$. L'élément central de la reconstitution de la voie de régulation en amont du gène i est dans $G_i(t)$, équation (2), l'analyse directe des données de microarray n'étant pas facile, une décomposition de Fourier est utilisée ainsi qu'une dérivation de $G_i(t)$ dans l'équation (3).

$$G_i(t) = \sum_{n=0}^{N}[\alpha_n \cos(n\omega t) + \beta_n \sin(n\omega t)] \tag{2}$$

$$\widehat{G}_i(t) = \sum_{n=0}^{N}[\widehat{\alpha_n} \cos(n\omega t) + \widehat{\beta_n} \sin(n\omega t)] \tag{3}$$

L'utilisation d'une fonction de régulation d'entrée est reliée aux interactions avec les gènes situés en amont (liaison physique ou transcription combinée). Ceci permet de retracer les gènes régulateurs correspondants, dans un premier temps à partir de la fonction $\widehat{G}_i(t)$ (3) du gène cible. L'algorithme complet n'est pas détaillé ici, l'objectif essentiel étant de trouver l'équation (4).

$$\widehat{G}_i(t) = \varepsilon_{i0} + \sum_{j \in R_i} c_{ij}\widetilde{X_j}(t - \tau_j) + e_i(t) \tag{4}$$

5.1.2 Prédiction des voies de régulation

Cette prédiction s'effectue par l'utilisation de la fonction de régulation identifiée comme étant $\widehat{G}_i(t)$ (4), cette fonction est interprétée comme étant la connection de régulation (transcription combinée ou interaction par liaison physique).

Les données de niveau d'expression d'ARNm sont disponibles sur les microarray, ces données permettent de remonter la voie de régulation en amont. Le niveau d'expression des transcripts d'ARNm est considéré comme proportionnel à la quantité de protéines produites dans la cellule [7], même si cela est une approximation du fait d'une activité post-transcriptionnelle importante susceptible de diminuer la traduction (microARN notamment).

De plus, différents éléments amènent à modifier certains éléments de prédiction, notamment de limiter l'effet des gènes de régulation au gène cible, où un effet de saturation est perçu et est modélisé par une transformation sigmoïde de $X_i(t)$ (équation (5)) de la part du gène régulateur j, avec γ le taux de transition, M_j l'expression moyenne du profile du gène j et τ_j le temps de transmission du signal.

$$\widetilde{X_j}(t-\tau_j) = \frac{1}{1+e^{-\gamma(X_j(t-\tau_j)-M_j)}} \tag{5}$$

Le temps τ_j de transduction du signal est calculé à partir d'une corrélation entre $\widetilde{X_j}$ du gène régulateur$j$ et la fonction de régulation $\widehat{G}_i(t)$ du gène cible i et d'un certain nombre de critères de maximisation. L'utilisation des connaissances biologiques sur ces étapes de calcul permet d'éliminer de l'ensemble R_j les gènes non impliqués comme facteurs de transcription, la phosphorylation de protéines ou l'activité enzymatique sur le gène cible i, car dans ce cas il s'agit de gènes co-exprimés.

Enfin, le Critère Informatif d'Aikake (*Aikaike Information Criteria - AIC* (6)) est utilisé pour déterminer le nombre optimal Ri de gènes régulateurs en amont, cette détermination s'effectue par la minimisation du critère AIC. Ce critère permet en effet de de déterminer la complexité d'un système par le calcul de sa variance résiduelle estimée $2\log\sigma$ qui diminue quand R_i augmente.

$$AIC = 2\log\sigma + \frac{2R_i}{m} \tag{6}$$

De façon générale, le déroulement de l'algorithme de prédiction des voies de régulation construit et utilisé par Chang et ses collaborateurs [7] est illustré dans la figure 1. Les étapes essentielles sont les suivantes :

1. Récupération d'un ensemble de données de microarray.
2. Application d'un filtre (interpolation).
3. Estimation de la dynamique du modèle et des temps de transduction.
4. Application des filtres biologiques et de AIC.
5. Reconstruction des voies modélisées.
6. Validation.

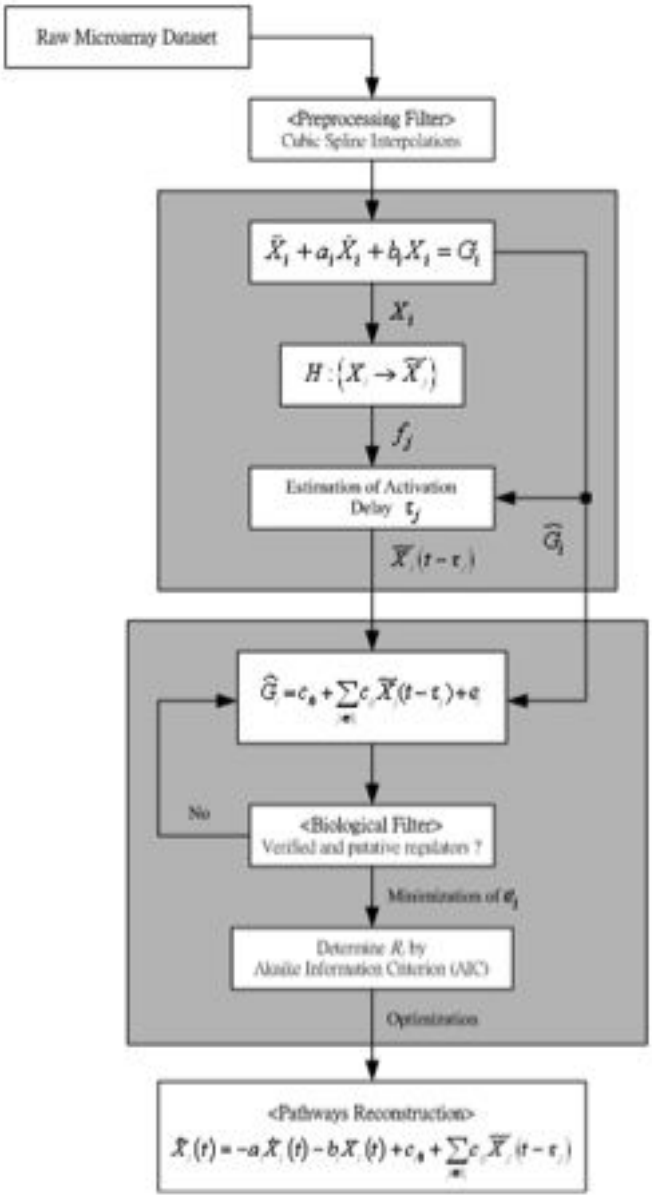

FIG. 1 – Déroulement général de l'algorithme de prédiction [7]

5.2 Modèle d'étude des rythmes circadiens de *A. thaliana*

Les données utilisées pour ce modèle sont celles établies par Harmer *et al.* dans leur étude sur le cycle circadien de *A. thaliana* [19]. En effet cette équipe a réalisé un certain nombre de puces à ADN pour déterminer des motifs d'expressions transcriptionnelles chez *A. thaliana* par intervalles de 4 heures. L'utilisation d'une analyse statistique a permis d'identifier 453 gènes impliqués dans le cycle circadien de *A. thaliana*, dont un certain nombre n'avaient pas été fonctionnellement identifiés, d'autres avaient été confirmés expérimentalement et annotés comme impliqués dans le cycle circadien.

Les gènes ainsi trouvés font partie de voies métaboliques dans la photosynthèse (ligand de la chlorophylle, cytochromes et phytochromes...) et de la coordination de voies métaboliques des lipides, des sucres, des protéines et des acides aminés. De plus, certains éléments du processus de développement végétal sont régulés par le biais de certains promoteurs utilisés à certaines étapes du cycle circadien. L'approche de Chang *et al.* a également utilisé des données connues sur les photorécepteurs de *A. thaliana* [14] pour affiner l'analyse des voies de régulation.

Le modèle de dynamique de régulation a permis d'obtenir un réseau de régulation (Figure 2), le jaune clair indique les crytochromes, le bleu clair les phytochromes, la couleur orange les gènes de l'horloge biologique, le vert clair certains gènes impliqués dans la dépendance physiologique à la lumière, et le gris pour des gènes importants non pris comme cibles. L'activité répressive est indiquée avec les lignes bleues (terminées par un tiret) et l'induction avec les lignes rouges (terminées par une flèche). La lecture est la même pour le réseau de régulation obtenu après validation par analyse temporelle aléatoire (Figure 3)

5.3 Modèle d'étude du métabolisme de *S. cerevisiae*

Les données utilisées pour ce modèle sont celle de DeRisi *et al.* dans leur étude sur le changement métabolique de *S. cerevisiae* [11]. Les expérimentations effectuées donnent un certain nombre d'informations sur les régulations allostériques des activités enzymatiques, la modification des protéines et la régulation transcriptionnelle.

Dans la voie de régulation du changment de voie métabolique (*S. cerevisiae*), la relation classique entre gènes enzymatiques est détectée correctement. Les voies de régulation les plus importantes sont identifiées (Figure 4), cependant certaines voies de régulation liées à la fermentation (bleu clair) ne peuvent être vraiment identifiées qu'au regard des relations de la néoglucogénèse (jaune clair). La validation (par mélange temporel aléatoire des données de microarray) de ce réseau métabolique (Figure 5) entraîne sa modification complète.

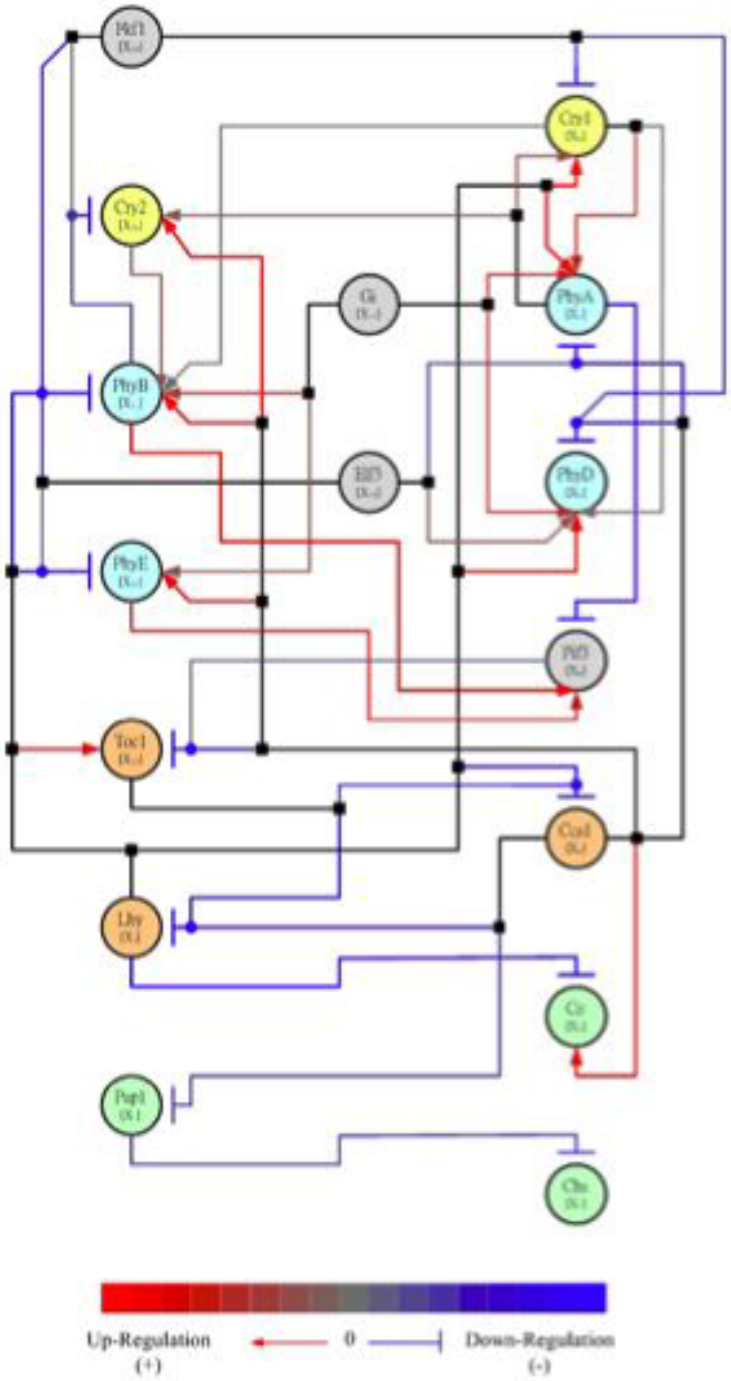

FIG. 2 – Régulation circadienne de *A. thaliana* [7]

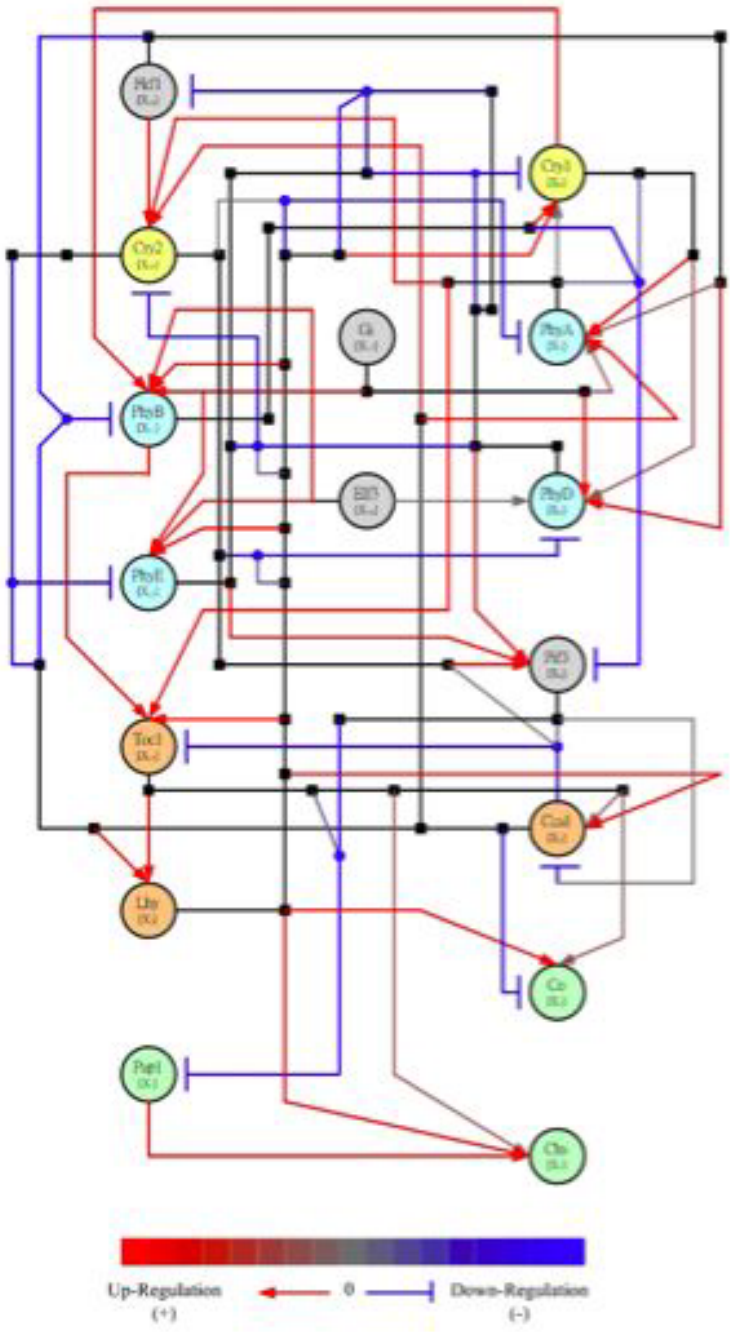

FIG. 3 – Régulation circadienne de *A. thaliana* après validation [7]

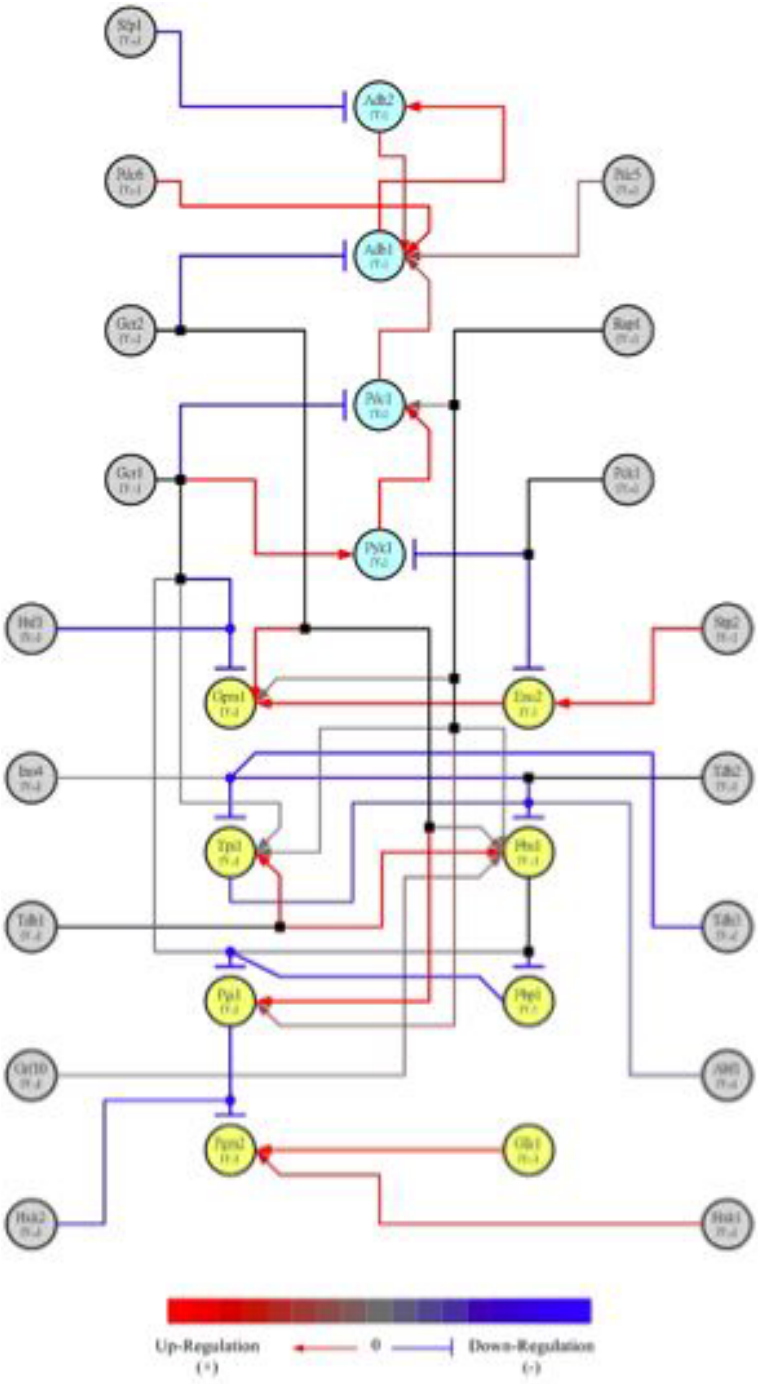

FIG. 4 – Régulation métabolique de *S. cerevisiae* [7]

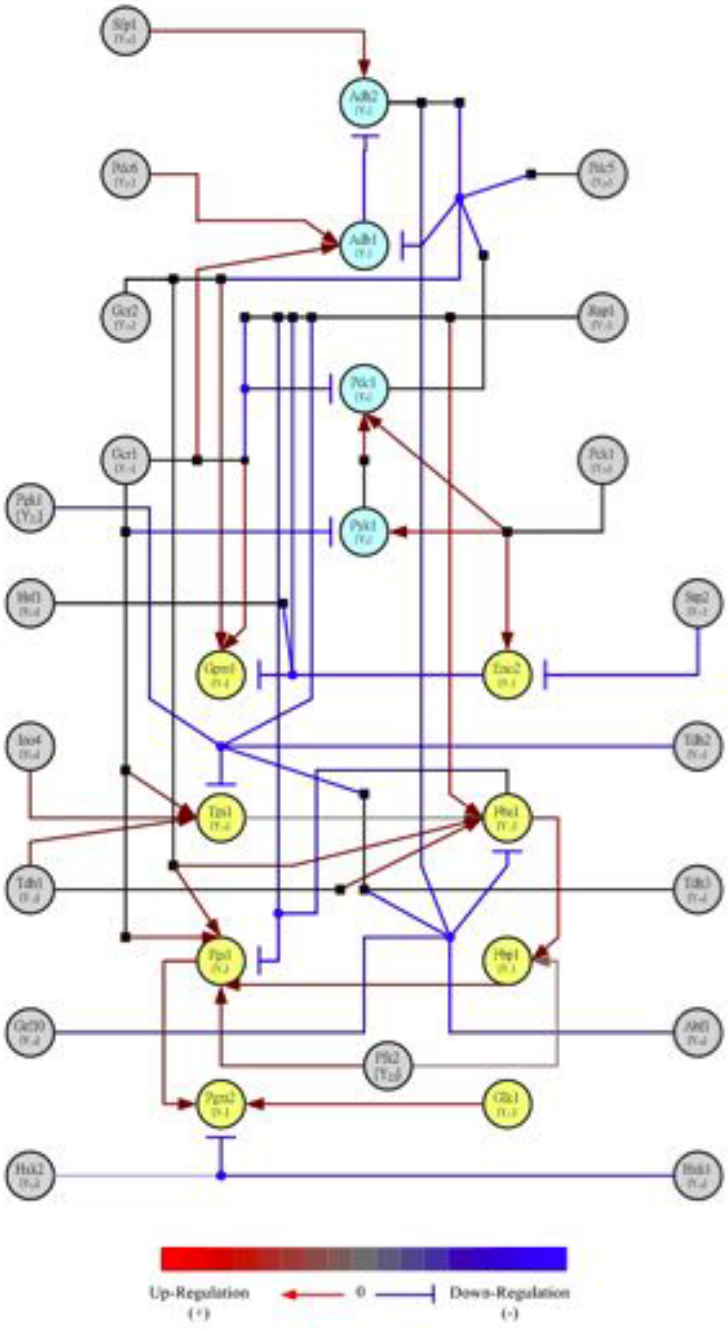

FIG. 5 – Régulation métabolique de *S. cerevisiae* après validation [7]

6 Conclusion et perspectives

6.1 Conclusion

6.1.1 Analyse des résultats : poursuite des recherches

Pour les deux voies étudiées, les résultats confirment les éléments déjà connus et complètent les connaissances pour le système de régulation circadien considéré, mais la recherche sur la voie de changements métaboliques ne recoupe pas correctement les éléments validés par ailleurs. Une analyse plus complète nécessite une méthode plus précise, notamment dans la partie de préparation des données : il peut être intéressant d'utiliser un plus grand nombre de données, voire d'utiliser plusieurs ensembles parallèles de données. De plus, les profils d'activation devraient être corrélés aux données protéomiques afin de permettre une meilleure interprétation et une meilleure analyse.

Un des objectifs d'utilisation de cette méthode est d'orienter la reconstruction des réseaux de régulation, pour éventuellement les confirmer expérimentalement, au besoin en utilisant les données protéomiques (traduction) et non plus seulement d'expression génomique (transcription).

6.1.2 Intérêt d'une telle approche

L'intérêt d'une telle méthodologie est de répondre au besoin croissant d'analyse des données de microarray par l'annotation de ces données dans un objectif de compléter les connaissances sur le fonctionnement du métabolisme et la compréhension des génomes ; de plus, l'utilisation d'organisme modèles bien étudiés, tels que *Arabidopsis thaliana* et *Saccharomyces cerevisiae*, permet d'obtenir un grand nombre de données d'études et de valider rapidement le modèle obtenu, par la comparaison avec d'autres modèles ou une série d'expérimentations dans le même sens.

L'analyse de données de microarray par une approche dynamique permet d'obtenir **un point de vue fonctionnel sur l'ensemble du génome** et de relier les profils d'expression des gènes via un système d'équations différentielles temporelles, avec un sens biologique. L'utilisation d'un critère temporel pour trier les données et déterminer la régulation entre gènes permet de modéliser sans directement tenir compte des interactions multi-spécifiques (contrôle transcriptionnel, phosphorylation, régulation enzymatique).

De plus, cette approche a plusieurs avantages par rapport aux approches bayesiennes ou booléennes, car elle relie de façon quantitative les différents éléments sans se limiter à un nombre d'états définis. **L'utilisation d'un calcul itératif permet de considérer chaque gène isolément au sein du réseau de régulation**, ceci permet de construire et de visualiser ce réseau étape par étape.

6.2 Perspectives

6.2.1 Recherches parallèles analogues

D'autres approches analogues ont été poursuivies et continuées de façon parallèle à celle faite par Chang *et al.* [7], mais utilisant des méthodologies de calcul de prédiction différentes pour la reconstruction de réseaux de régulation de gènes. Par exemple une reconstruction construite sur un classement (Chen *et al.* [9]) ou une méthode de couplage pour une modélisation directionnelle de la dépendance entre gènes (Jong-Min Kim *et al.* [25]).

Concernant l'approche par l'utilisation des réseaux bayesiens, on peut notamment en citer trois : une intégration bayésienne de connaissances biologiques dans la reconstruction des réseaux de régulation génétiques (Dirk Husmeier *et al.* [22]), une approche de réseaux bayesiens dynamiques pour identifier les réseaux de régulation des gènes à partir de données de microarray temporisées (Zou *et al.* [38]) et un algorithme des moindres carrés d'ingénierie inverse sur la régulation des gènes (Chang Sik Kim [24]).

Ces approches utilisant les réseaux bayésiens rejoignent celle conçue par Chang dans leur sensibilité et la spécificité des scores attribués, ainsi que dans la conception des équations différentielles utilisées pour calculer et prédire les réseaux de régulation géniques étudiés. Ces algorithmes, conçus sur des éléments d'approche et d'échantillonnage des données différents, sont étudiés, utilisés et validés sur des données analogues, notamment les gènes impliqués dans le métabolisme de *Saccharomyces cerevisiae*. L'objectif étant là aussi de limiter l'espace de recherche à un réseau de gènes au sein du métabolisme, avant d'étendre la recherche à un génome entier ou un autre réseau de gènes.

6.2.2 Recherches effectutées dans le prolongement

Certains articles postérieurs à 2005 citent Chang *et al.* [7], à propos certains aspects de l'approche utilisée dans la prédiction des réseaux de régulation de gènes. Wu et ses collaborateurs reprennent en 2006 [35] et 2007 [36], l'intérêt d'une représentation du profil des facteurs de transcription et sa modélisation par une fonction sigmoïde, même si cela ne reflète pas toujours la réalité biologique (notamment dans la régulation post-transcriptionnelle) mais cela reflète bien l'ensemble des cas de régulation génique au sein des cellules.

En 2006, Wang *et al.* [33] indiquent l'utilisation des prédictions de Chang *et al.* [7]. Plus proche dans le temps, Novikov et Barillot en 2008 [28] précisent l'utilisation de l'interpolation des données lors de la prédiction et l'utilisation d'équations différentielles, de la même façon que Barenco *et al.* en 2006 [2] pour la prise en compte du temps d'action lors de l'estimation des paramètres et variables concernant les facteurs de transcription et les gènes régulateurs.

En 2006, Chang et ses collaborateurs avaient également repris leur travail sur l'identification des facteurs de transcription, mais via un modèle stochastique [8], dans une volonté d'améliorer les prédictions en prenant aussi en compte les intéractions et coopérations entre les facteurs de transcription et les gènes régulateurs.

---

7 Bibliographie

Références

[1] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin et Gavin Sherlock : Gene Ontology : tool for the unification of biology. *Nature Genetics*, 25:25–29, May 2000.

[2] Martino Barenco, Daniela Tomescu, Daniel Brewer, Robin Callard, Jaroslav Stark et Michael Hubank : Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25, 2006.

[3] Johannes Berg, Stana Willmann et Michael Lässig : Adaptive evolution of transcription factor binding sites. *BMC Evolutionnary Biology*, 4(1):42, October 2004.

[4] Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Jaak Vilo, Niran Abeygunawardena, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, Ahmet Oezcimen, Philippe Rocca-Serra et Susanna-Assunta Sansone : ArrayExpress - public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, 31(1):68–71, January 2003.

[5] Harmen J. Bussemaker, Lucas D. Ward et André Boorsma : Dissecting complex transcriptional responses using pathway-level scores based on prior information. *BMC Bioinformatics*, 8(S6):S6, September 2007.

[6] Gregory W. Carter : Inferring network interactions within a cell. *Briefings in Bioinformatics*, 6(4):380–389, December 2005.

[7] Wen-Chieh Chang, Chang-Wei Li et Bor-Sen Chen : Quantitative inference of dynamic regulatory pathways via microarray data. *BMC Bioinformatics*, 6:44, March 2005.

[8] Yu-Hsiang Chang, Yu-Chao Wang et Bor-Sen Chen : Identification of transcription factor cooperativity via stochastic system model. *Bioinformatics*, 22(18):2276–2282, September 2006.

[9] Guanrao Chen, Peter Larsen, Eyad Almasri et Yang Dai : Rank-based edge reconstruction for scale-free genetic regulatory networks. *BMC Bioinformatics*, 9:75, January 2008.

[10] Chan-San Chin, Victor Chubukov, Emmitt R. Jolly, Joe DeRisi et Li Hao : Dynamics and design principles of a basic regulatory architecture controlling metabolic pathways. *PLoS Biology*, 6(6):1343–1356, June 2008.

[11] Joseph L. DeRisi, Vishwanath R. Iyer et Patrick O. Brown : Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, October 1997.

[12] Amira Djebbari et John Quackenbush : Seeded bayesian networks : Constructing genetic networks from microarray data. *BMC System Biology*, 2:S7, July 2008.

[13] Ron Edgar, Michael Domrachev et Alex E. Lash : Gene Expression Omnibus : NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 30(1):207–210, January 2002.

[14] Christian Fankhauser et Dorothee Staiger : Photoreceptors in Arabidopsis thaliana : light perception, signal transduction and entrainment of the endogenous clock. *Planta*, 216(1):1–16, November 2002.

[15] Nir Friedman, Michal Linial, Iftach Nachman et Dana Pe'er : Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3–4):601–620, 2000.

[16] Robert C. Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y. H. Yang et Jianhua Zhang : Bioconductor : open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, September 2004.

[17] Ulrich Gerland, J. David Moroz et Terence Hwa : Physical constraints and functional characteristics of transcription factor-dna interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 99(19):12015–12020, September 2002.

[18] Christopher T. Harbison, Benjamin D. Gordon, Tong Ihn Lee, Nicola J. Rinaldi, Kenzie D. Macisaac, Timothy W. Danford, Nancy M. Hannett, Jean-Bosco Tagne, David B. Reynolds, Jane Yoo, Ezra G. Jennings, Julia Zeitlinger, Dmitry K. Pokholok, Manolis Kellis, Alex P. Rolfe, Ken T. Takusagawa, Eric S. Lander, David K. Gifford, Ernest Fraenkel et Richard A. Young : Transcriptional regulatory code of a eukaryotic genome. *Nature*, 731(7004):99–104, September 2004.

[19] Stacey L. Harmer, John B. Hogenesch, Marty Straume, Hur-Song Chang, Bin Han, Tong Zhu, Xun Wang, Joel A. Kreps et Steve A. Kay : Orchestrated transcription of key pathways in arabidopsis by the circadian clock. *Science*, 290(5499):2110–2113, December 2000.

[20] Alan G. Hinnebusch : Mechanisms of gene regulation in the general control of amino acid biosynthesis in saccharomyces cerevisiae. *Microbiological Reviews*, 52(2):248–273, June 1988.

[21] Wolfgang Huber, Vincent J. Carey, Li Long, Seth Falcon et Robert Gentleman : Graphs in molecular biology. *BMC Bioinformatics*, 8(S6): S8, September 2007.

[22] Dirk Husmeier et Adriano V. Werhli : Bayesian integration of biological prior knowledge into the reconstruction of gene regulatory networks with bayesian networks. *In* San Diego University of California, éditeur : *Proc Life Science Society Comput Syst Bioinform Conf.*, volume 6, pages 85–95. Biomathematics and Statistics Scotland, Edinburgh, United Kingdom, Life Science Society, August 2007.

[23] M. Kanehisa : The KEGG database. *Novartis Foundation symposium*, 247:91–101,101–103,119–128,244–252, 2002.

[24] Chang Sik Kim : Bayesian Orthogonal Least Squares (BOLS) algorithm for reverse engineering of gene regulatory networks. *BMC Bioinformatics*, 8:251, July 2007.

[25] Jong-Min Kim, Yoon-Sum Jung, Engin A Sungur, Kap-Hoon Han, Changyi Park et Insuk Sohn : A copula method for modeling directional dependence of genes. *BMC Bioinformatics*, 9:225, May 2008.

[26] Gunter B. Kohlhaw : Leucine biosynthesis in fungi : entering metabolism through the back door. *Microbiology and Molecular Biology Reviews*, 67(1): 1–15, March 2003.

[27] Michael Lässig : From biophysics to evolutionary genetics : statistical aspects of gene regulation. *BMC Bioinformatics*, 8(S6):S7, September 2007.

[28] Eugene Novikov et Emmanuel Barillot : Regulatory network reconstruction using an integral additive model with flexible kernel functions. *BMC Systems Biology*, 2:8, January 2008.

[29] M. Schena, D. Shalon, R. W. Davis et P. O. Brown : Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, October 1995.

[30] Gavin Sherlock, Tina Hernandez-Boussard, Andrew Kasarskis, Gail Binkley, MateseJohn C., Selina S. Dwight, Miroslava Kaloper, Shuai Weng, Heng Jin, Catherine A. Ball, Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, David Botstein et J. Michael Cherry : The Stanford Microarray Database. *Nucleic Acids Res.*, 29(1):152–155, January 2001.

[31] Eduardo Sontag, Anatoly Kiyatkin et Boris N. Kholodenko : Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinformatics*, 20(12):1877–1886, August 2004.

[32] Diethard Tautz : Evolution of transcriptional regulation. *Current Opinion in Genetics and Development*, 10(5):575–579, October 2000.

[33] Xian Wang, Ao Li, Zhaohui Jiang et Huanqing Feng : Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC Bioinformatics*, 7:32, January 2006.

[34] Gregory A. Wray, Matthew W. Hahn, Ehab Abouheif, James P. Balhoff, Margaret Pizer, Matthew V. Rockman et Laura A. Romano : The evolution of transcriptional regulation in eukaryotes. *Molecular Biology and Evolution*, 20(9):1377–1419, September 2003.

[35] Wei-Sheng Wu, Wen-Hsiung Li et Bor-Sen Chen : Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. *BMC Bioinformatics*, 7:427, September 2006.

[36] Wei-Sheng Wu, Wen-Hsiung Li et Bor-Sen Chen : Identifying regulatory targets of cell cycle transcription factors using gene expression and chip-chip data. *BMC Bioinformatics*, 8:188, June 2007.

[37] Marcelo J. Yanovsky et Steve A. Kay : Signaling networks in the plant circadian system. *Current Opinion in Plant Biology*, 4(5):429–435, October 2001.

[38] Min Zou et 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, November 2005.