Reconstruction de réseaux de gènes à partir de données d'expression par déconvolution centrée autour des hubs

Adel Ait-Hamlat

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

THÈSE DE DOCTORAT DE

Sorbonne Université

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Adel Ait-hamlat

Pour obtenir le grade de

DOCTEUR de Sorbonne Université

Sujet de la thèse :

Reconstruction de réseaux de gènes à partir de données d'expression par déconvolution centrée autour des hubs

soutenue le 25 février 2019

devant le jury composé de :

Mme. Alessandra Carbone	co-directeur de thèse
M. Thierry Jaffredo	co-directeur de thèse
M. Pierre Charbord	encadrant de thèse
M. Charles Durand	encadrant de thèse
M. Alberto Policriti	Rapporteur
Mme. Elisabeth Remy	Rapporteur
M. Jean-daniel Zucker	Examinateur

Table des matières

Chapitre 1

Avant-propos

Ce rapport vient conclure une thèse financée par l'Institut de Biologie Paris-Seine de Sorbone Université Paris VI dans le but de promouvoir les projets de rapprochement d'équipes de recherche de spécialités différentes. Mon travail de thèse, entamé en octobre 2014, s'est déroulé sous la double tutelle de Mme Alessandra Carbone, directrice de l'équipe de recherche Génomique Analytique du Laboratoire de Biologie Computationnelle et Quantitative, et de Mr Thierry Jaffredo, chef de l'équipe de recherche Migration et Différentiation des Cellules Souches Hématopoïétiques au sein du laboratoire de Biologie du Développement de l'institut de Biologie Paris-Seine. Mme Carbone dirige une équipe composée de mathématiciens, informaticiens, et bio-informaticiens qui travaillent à la modélisation mathématique de problématiques biologiques et à l'implémentation algorithmique de ces modèles. L'équipe dirigée par Mr Jaffredo est quant à elle composée de biologistes dont le travail se base sur des expérimentations en paillasse visant à tester des hypothèse biologiques.

L'objectif fixé initialement de cette thèse consistait à analyser la communication moléculaire entre cellules stromales et cellules souches hématopoïétiques, les premiers constituant la niche *in vivo* des seconds (une caractérisation des cellules souches hématopoïétiques sera présentée dans le dernier chapitre de ce document.) La stratégie scientifique établie alors se structurait en deux temps, un premier dédié

à la production de données par des expérimentations de biologie en paillasse, et un second à l'analyse de ces données par des outils algorithmiques. Le premier temps consistait à réaliser des expériences de cultures de chaque type cellulaire isolé et d'une co-culture des deux types cellulaires mis en contact. Le deuxième temps serait alors consacré à l'analyse des données prélevées de ces différentes cultures en comparant le système composé des deux types cellulaires en contact aux systèmes composées des types cellulaires isolés. Chaque système cellulaire devait être caractérisé par le réseau des interactions des gènes actifs dans ce système permettant ainsi de comparer les systèmes en comparant les réseaux caractéristiques de chacun. Le but de la deuxième partie était donc de développer une méthode de production de réseaux géniques et d'une méthode de comparaison de ces réseaux.

N'ayant au début de cette thèse aucune expérience en manipulations biologiques de paillasse, j'ai passé les dix-huit premiers mois sous le bienveillant encadrement de Mr Pierre Charbord et Mr Charles Durand à me former à cet exercice. Malheureusement, après plusieurs tentatives, les expériences de co-culture se sont révélées impossibles à réaliser avec le matériel biologique dont nous disposions, l'interaction de contact recherchée entre les cellules stromales et hématopoïétiques ne se réalisant que trop rarement. Il a donc était décidé de concert avec l'ensemble de mes encadrants de rediriger mon travail vers un développement purement méthodologique qui a aboutit aux résultats décrits dans ce qui suit. Néanmoins, ces mois passés à me former puis à réaliser des expérimentations biologiques en paillasse m'ont permis, en plus du savoir faire pratique acquis, de prendre conscience du travail fastidieux qui doit se faire en amont de l'analyse de données. En effet, ayant jusque là commencé mes différentes activités de recherche à partir des données, je me souciais peu voire pas du tout du processus qui permit la génération de ces données. Aujourd'hui, fort de cette expérience, j'ai une vision plus globale du processus de recherche en biologie et en bio-informatique.

Chapitre 2

Introduction

2.1 Motivations

Dans un organisme multicellulaire, les cellules sont différenciées par les gènes qu'elles expriment et par conséquent par les protéines traduites à partir de ces gènes exprimés. Une cellule ne produit en effet que les protéines nécessaires à sa fonction qui lui confèrent un phénotype propre. Par ailleurs, les gènes codant pour les protéines ne représentent qu'une petite fraction de l'ADN. Le séquençage du génome humain au début du siècle a par exemple révélé que les gènes ne représentaient pas plus de 3% de l'ADN humain [1]. On estime aujourd'hui ce taux à environ 1.5% [2, 3]. De plus, le nombre de gènes diffère fortement entre organismes, le riz possède par exemple plus de deux fois plus de gènes que l'homme, révélant que la complexité d'un organisme ne provient pas du nombre de gènes mis en jeu, mais de la dynamique des interactions entre ces gènes [4].

L'étude des systèmes biologiques complexes ne peut désormais plus suffire à identifier les gènes en jeu et leurs rôles individuels. Elle doit décrire les interactions qui existent entre ces gènes pour révéler les propriétés d'ensemble du système [5]. En régulant l'expression des gènes d'une cellule, ces interactions contrôlent la production des protéines nécessaires à la fonction spécifique de la cellule et lui permettent

de s'adapter à son environnement. La régulation de l'expression des gènes peut se produire à plusieurs niveaux. Au niveau de la transcription, la régulation se fait par l'intermédiaire de facteurs de transcription (FT), des protéines qui se lient aux régions régulatrices de gènes cibles pour réprimer ou induire leur transcription. Chez les eucaryotes, la régulation peut également impliquer des facteurs épigénétiques tels que des enzymes pour la méthylation de l'ADN et des remodeleurs de la chromatine [6].

Ces interactions entre gènes peuvent être modélisées en un réseau permettant d'analyser les systèmes cellulaires dans leur ensemble grâce aux outils développés dans la théorie des graphes. Les réseaux de régulation de gènes (RRGs) ainsi obtenus permettent de représenter par des arêtes dirigées les relations causales directes entre gènes régulateurs et leurs cibles. Beaucoup de données peuvent être utilisées pour modéliser les RRGs. Les données les plus explicites permettent d'observer les interactions entre FT et gènes régulés. Ces données sont toutefois rares. Beaucoup plus fréquentes, les données d'expression des gènes se sont accumulées dans les bases de données grâce au développement rapide des technologies permettant de mesurer les niveaux d'expression de plusieurs milliers de gènes simultanément et à cout relativement réduit. Ainsi, les quinze dernières années ont vu l'émergence d'un grand nombre de méthodes visant à inférer des RRGs à partir des données d'expressions des gènes représentés. Néanmoins, malgré la profusion des méthodes de reconstruction de RRGs développées à ce jour, reposant sur des bases théoriques variées, ce problème est loin d'être résolu.

Le but de cette thèse est de décrire une nouvelle méthode d'inférence de RRGs à partir de données d'expression de gènes. Cette méthode, que nous appelons HubNeD (pour "Hub-centered Network Deconvolution"), présente des performances nettement supérieures aux méthodes existantes. Nous organisons cette thèse comme suit : une première partie sera consacrée à la description du contexte biologique et à une in-

troduction à la théorie des graphes. Ces notions nous permettront ensuite d'analyser les RRGs des génomes complets de *Saccharomyces cerevisiae* et d'*Escherichia coli*, deux modèles de régulation de gènes bien établis, considérés dans la communauté comme des réseaux de référence. Cette analyse permettra de mettre en exergue les caractéristiques globales de RRGs et d'en dégager trois qui serviront d'hypothèses à la méthode présentée. Nous reviendrons ensuite sur l'état de l'art des méthodes existantes, avant de définir la nouvelle méthode développée ici et d'en comparer les performances avec trois autres méthodes représentatives des différentes familles méthodologiques.

2.2 Contexte biologique

La très grande diversité du vivant repose sur des mécanismes biologiques communs. Toute l'information nécessaire au développement puis au maintient d'un être vivant, information héritée et transmise à la génération suivante, est identiquement stockée dans chacune des cellules qui le composent. Cette information est inscrite dans les molécules doubles brins d'ADN qui forment les chromosomes. Elle est codée en une longue séquence à partir d'un alphabet de quatre lettres, A, C, G, T, pour respectivement Adénine, Cytosine, Guanine, et Thymine, les nucléotides qui composent l'ADN. Le mécanisme de décodage de cette information est lui aussi commun à tous les êtres vivants. Des bouts de la séquence d'ADN, appelés gènes, sont d'abords transcrits en molécules simple brin d'ARN messagers (ou ARNm) qui sont, pareillement à l'ADN, des sequences de nucléotides avec simplement les nucléotides T remplacés par des nucléotides U (Uracile). La synthèse des protéines se fait par traduction des séquences d'ARNm en une séquence composée à partir de vingt acide-aminés. Ce processus, appelé dogme central de la biologie moléculaire par Francis Crick en 1956[7], est illustré dans la Figure 2.1.

La synthèse d'ARNm est réalisée par une protéine appelée ARN polymérase. La Figure 2.2 schématise l'action de l'ARN polymérase dont l'activité est contrôlée par

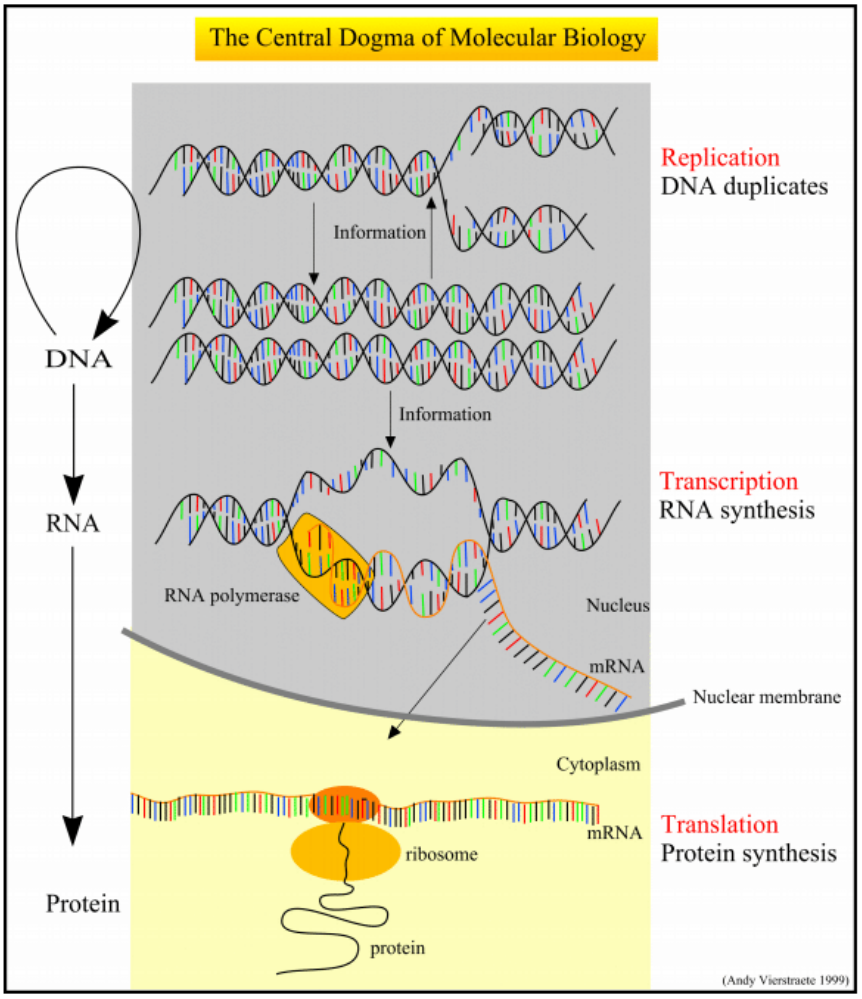


FIGURE 2.1 – Le dogme central de la biologie représentant les deux étapes de l'expression des gènes (transcription et traduction), ainsi que la réplication qui permet la duplication de l'ADN. Cette figure a été récupérée de "http ://users.ugent.be/ avierstr/principles/centraldogma.html".

différents facteurs de transcription. Un facteur de transcription (FT) se fixe à l'ADN et contrôle ainsi l'expression du gène en aval. Ce contrôle peut être activateur ou répresseur. Un FT se lie à des régions spécifiques d'amplification en *trans* (loin dans la séquence d'ADN transcrite) ou en *cis* sur le promoteur en amont du gène. Certains FTs pour être actifs se lient à d'autres protéines appelées *co-facteurs*.

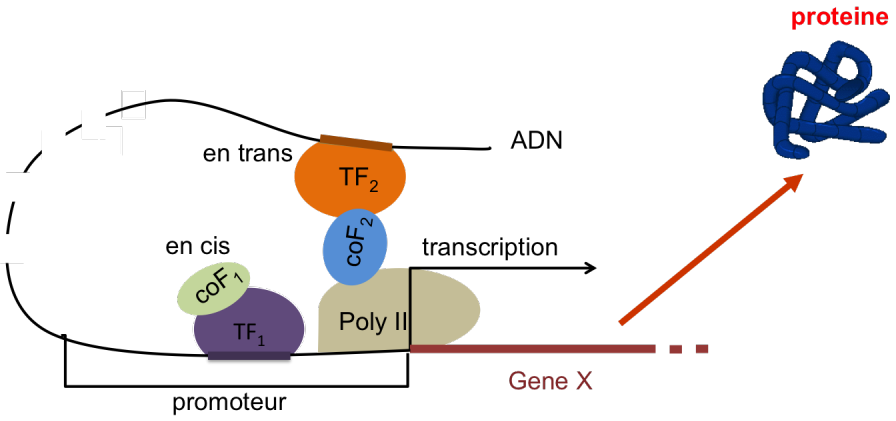


FIGURE 2.2 – Régulation de la transcription d'un gène. FT et co-facteurs s'associent pour se fixer sur des séquences spécifiques de l'ADN afin de réguler l'activité de l'ARN polymérase qui transcrit un gène en aval des sites de fixation.

2.3 Les graphes appliqués à la biologie

Les graphes, ou *réseaux*, sont des modèles utiles pour représenter des systèmes composés d'entités qui interagissent. Ces modèles se présentent en un schéma qui relie les entités, molécules ou gènes dans le cas des réseaux cellulaires, par des liens (traits ou flèches) qui représentent des relations entre les entités. Les premiers réseaux cellulaires modélisaient des voies métaboliques [8, 9]. Puis on s'est intéressé à d'autres formes d'interactions, comme celles qui permettent de représenter les associations de protéines en complexes ou celles décrivant les régulations de l'expression des gènes.

Un réseau métabolique représente l'ensemble des réactions chimiques d'un organisme qui transforment substrats (molécules d'entrée) en produits (molécules de sortie). En général, des molécules interviennent comme catalyseurs pour aider ces réactions à se réaliser. Ces catalyseurs sont des enzymes, protéines produites par l'organisme. Les réactions chimiques d'un métabolisme peuvent s'enchainer quand les produits en sortie d'une réaction sont utilisés comme substrats pour d'autres réactions. Les réseaux métaboliques permettent alors de représenter l'ensemble de

ces réactions par des noeuds pour représenter les réactions chimiques et les arêtes pour représenter les enzymes.

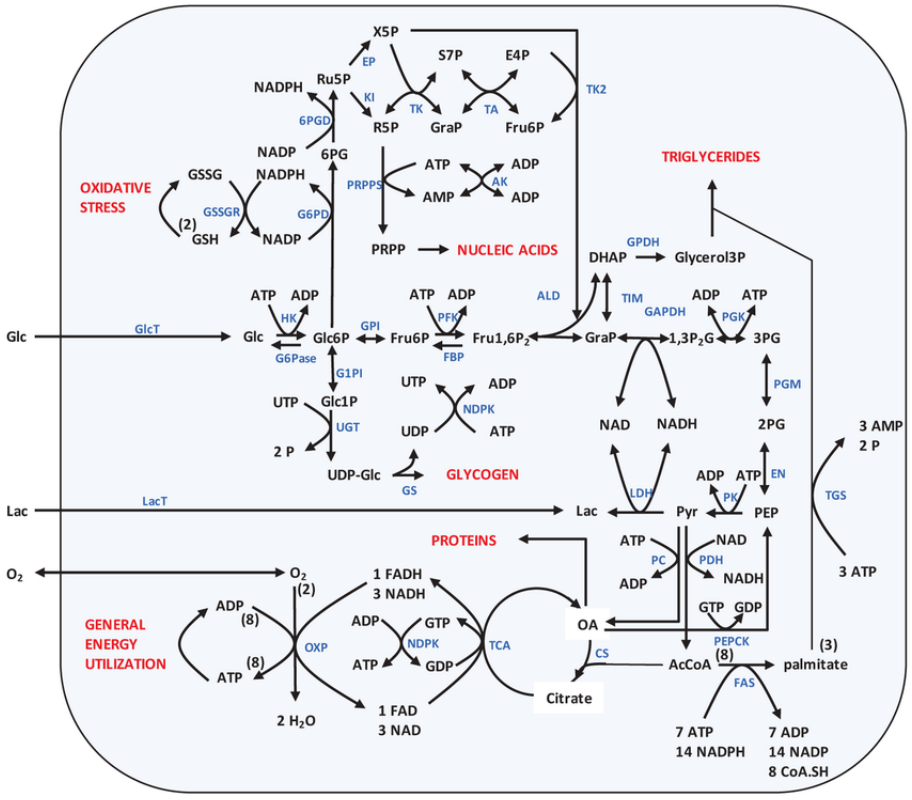


FIGURE 2.3 – Exemple d'un réseau métabolique représentant le métabolisme du carbone pris de [10].

Les réseaux protéine-protéine rendent compte des associations physique entre protéines. Ces associations peuvent servir plusieurs fonctions. Une protéine peut par exemple se lier à une autre pour aider à la transporter d'un compartiment cellulaire vers un autre. D'autres protéines doivent se lier formant un complexe protéique pour pouvoir être actives (l'hémoglobine est par exemple un complexe formé par plusieurs protéines). Une protéine peut également se lier à une autre pour pour la modifier (par phosphorilation, par exemple). Les noeuds d'un réseau protéine-protéine représentent les protéines et une arête entre deux protéines signifie une

association physique entre elles.

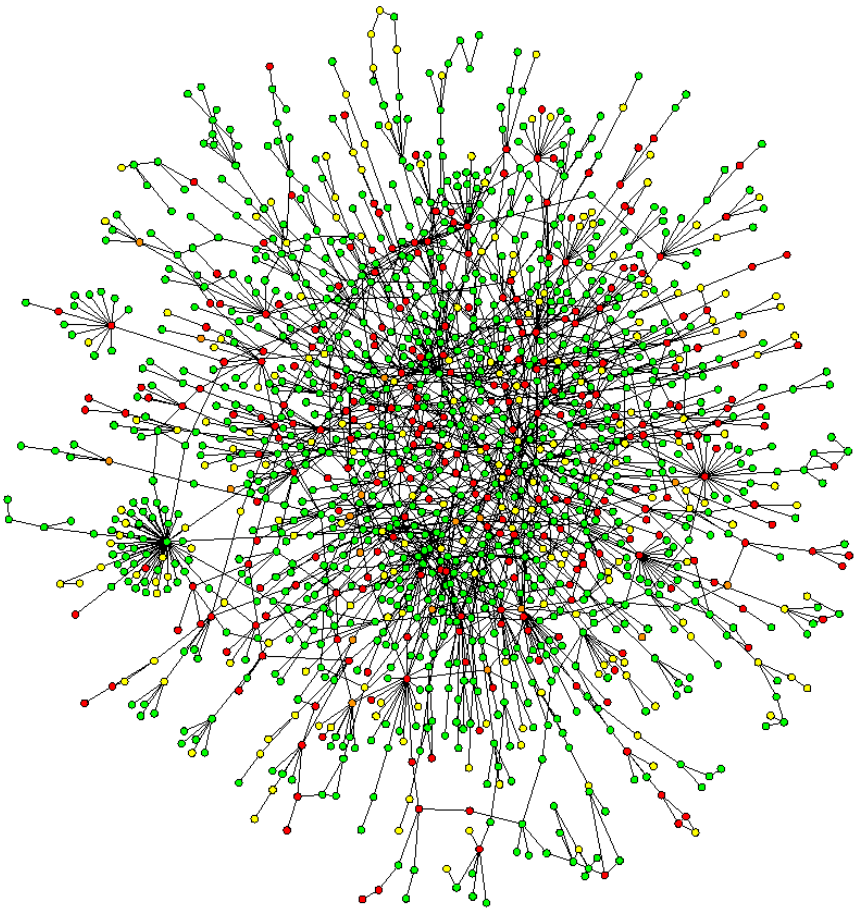


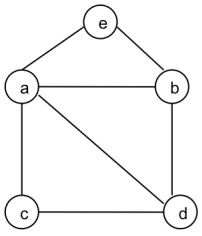
FIGURE 2.4 – Réseau des interactions protéine-protéine chez *Saccharomyces cerevisiae* [11].

Les réseaux de régulation de gènes (RRGs) sont les modèles qui nous intéresse dans ce travail de thèse. Un RRG représente les relations entre gènes régulateurs et leurs cibles, les premiers contrôlant l'expression des deuxièmes. La Figure 2.2 schématise le contrôle de l'expression des gènes par les FT. La régulation de l'expression des gènes peut cependant se réaliser par d'autres vecteurs. Les remodeleurs de la chromatines sont par exemple d'autres protéines de régulation de la transcription. Ces protéines peuvent changer la conformation de la chromatine (structure formée par l'ADN compacté) pour en relâcher certaines régions qui deviennent alors accessibles à la machinerie de transcription et aux FT. La régulation de l'expression d'un gène peut également se dérouler entre la transcription et la traduction, par l'intermédiaire de petites séquence d'ARN (microARN) qui se lient spécifiquement à des ARNm transcrits pour en empêcher la traduction en protéine. Les RRGs considéré dans ce présent document ne tiennent compte que des régulation par FTs. Leurs arêtes relient gènes régulateurs et gènes régulés, les protéines issues de la traduction des premiers agissant comme FT contrôlant la transcription des derniers.

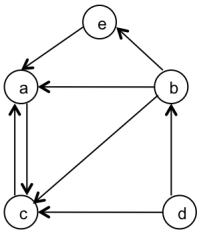
2.4 Introduction à la théorie des graphes

Un **graphe** $G = (N, A)$ est défini par un ensemble de noeuds N et par des arêtes $A \subseteq N^2$. Le graphe G est dit orienté si les arêtes dans A ont une orientation. Une arrête $(i, j) \in A$ est alors orientée du noeud i vers le noeud j et est différente de l'arrête (j, i) qui est d'orientation opposée. Deux exemples de graphes G_1 et G_2 sont représentés ci-dessous. G_1 est non orienté et $G2$ est orienté ; ils sont définis par :

$G_1 = (N_1, A_1)$ et $G_2 = (N_2, A_2)$ avec
$N_1 = N_2 = \{a, b, c, d, e\}$
$A_1 = \{(a, b), (a, c), (a, d), (a, e), (b, d), (b, e), (c, d)\}$
$A_2 = \{(a, c), (b, a), (b, c), (b, e), (c, a), (d, c), (d, b), (e, a)\}$



G_1



G_2

Nous présentons dans ce qui suit des notions issues de la théorie des graphes. Ces notions nous seront utiles pour l'étude des différents graphes abordés dans ce travail. Nous considérerons pour la suite de cette section un graph $G = (N, A)$ de n noeuds ($|N| = n$).

La **matrice d'adjacence** d'un graphe est une matrice booléenne de dimension $n \times n$. Si Adj est la matrice de G, alors :

$$Adj(i,j) = \begin{cases} 1 & si \quad (i,j) \in A \\ 0 & sinon \end{cases}$$

Les matrices d'ajacence des graphes G_1 et G_2 ci-dessus sont respectivement :

$$Adj_{G_1} = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array} \quad et \quad Adj_{G_2} = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

On peut remarquer que Adj_{G_1} est symétrique ($Adj_{G_1}[i,j] = Adj_{G_1}[j,i]$), parce que G_1 n'est pas orienté, alors que Adj_{G_2} n'est pas symétrique, parce que G_2 est orienté.

Deux noeuds i et j sont dits **adjacents** s'il sont reliés par une arrête dans le

graph, *i.e.* $(i, j) \in A$

Un **graphe complet** est un graphe où deux noeuds quelconques sont adjacents. Un graphe complet à n noeuds contient $\frac{n(n-1)}{2}$ arrêtes. On peut facilement arriver à ce résultat par comptage des arrêtes. En commençant par les arrêtes d'un noeud quelconque, adjacents aux $n-1$ autres noeuds, puis en comptant les arrêtes d'un deuxième noeud en évitant de recompter l'arrête le reliant au noeud déjà parcouru, donc $n-2$ arrêtes, et ainsi de suite, pour aboutir à un nombre total d'arrêtes égal à $(n-1)+(n-2)+...+1$, somme égale à $\frac{n(n-1)}{2}$. Notons que ces formules sont vraies quand G est sans boucle, c'est à dire sans arête qui lie un noeud à lui même. Les graphes considérées dans ce travail seront tous considérés sans boucles.

La densité d'un graphe est le rapport du nombre d'arrêtes qu'il contient et du nombre d'arrêtes du graphe complet de même taille (même nombre de noeuds). La densité d'un graphe $G = (N, A)$ est donc $\frac{2|A|}{n(n-1)}$ où n est le nombre de noeuds du graphe.

Un graphe $G' = (N', A')$ est un sous-graphe de G si et seulement si $N' \subseteq N$ et $A' \subseteq A$. Le sous-graphe est dit **maximal** si il n'existe pas d'arrête dont une extrémité est dans G' et l'autre n'est pas dans G'.

Une clique est un sous-graphe complet. $G'_1 = (\{a, b, e\}, \{(a, e), (a, b), (b, e)\})$ est une clique de taille 3 de G_1.

Le voisinage d'un noeud i est l'ensemble des noeuds adjacents à i. Formellement :

$$\forall i \in N, \quad V(i) = \{j \in N \mid (i, j) \in A\}$$

A titre d'exemple, dans le graphe G_1, le voisinage de b est $V(b) = \{a, d, e\}$.

Dans le cas d'un graphe orienté, la notion de voisinage s'étend naturellement aux voisinages entrant et sortant, respectivement définis par :

$$\forall i \in N, \quad V^+(i) = \{j \in N \mid (i,j) \in A\}$$
$$\forall i \in N, \quad V^-(i) = \{j \in N \mid (j,i) \in A\}$$

Ainsi, dans G_2, $V^+(a) = \{c\}$ et $V^-(a) = \{b, c, e\}$.

Le degré d'un noeud est le nombre d'éléments dans son voisinage :

$$\forall i \in N, \quad Deg(i) = |V(i)| = |\{j \in N \mid (i,j) \in A\}$$

Le noeud a du graphe G_1 a pour degré 4, les noeuds b et c sont de degré 3 et les noeuds d et e sont de degré 2.

Nous parlerons de degré sortant et degré entrant d'un noeud lorsqu'il s'agira d'un graphe orienté, pour représenter le nombre de noeuds compris respectivement dans le voisinage entrant et le voisinage sortant de ce noeud.

$$\forall i \in N, \quad Deg^+(i) = |V^+(i)| = |\{j \in N \mid (i,j) \in A\}$$
$$\forall n \in N, \quad Deg^+(i) = |V^+(i)| = |\{j \in N \mid (j,i) \in A\}$$

A titre d'illustration, dans G_2, $Deg^+(a) = 1$ et $Deg^-(a) = 3$.

Un chemin est défini par une suite ordonnée de noeuds adjacents d'un graphe. Si le graphe est orienté, le chemin doit évidemment respecter l'orientation des arrêtes, c'est à dire que chaque noeud du chemin doit être parmi le voisinage sortant du précédent. La longueur d'un chemin est le nombre d'arrêtes traversées par le chemin. Si le premier noeud d'un chemin est i et le dernier est j, nous dirons que le chemin relie i à j. Un chemin qui relie un noeud à lui même est un cycle.

Si G n'est pas orienté, il est dit connexe si chaque paire de ses noeuds est reliée par un chemin. Un graphe orienté est connexe si le graphe non orienté obtenu en ignorant l'orientation de ses arrêtes est connexe.

Une composante connexe de G est un sous-graphe connexe maximal de G. Par définition, il n'existe pas de chemin entre deux noeuds appartenant à deux composantes connexes différentes.

La distance géodésique $d(i,j)$ est la longueur minimum des chemins reliant i à j dans G. Notons que dans un graphe orienté $d(i,j)$ ne coincide pas nécessairement avec $d(j,i)$. En effet, dans G_2, $d(d,e) = 2$ alors que $d(e,d) = \infty$ puisqu'il n'y a pas de chemin reliant e à d dans G_2. Dans un graphe non orienté une distance géodésique est infinie si et seulement si les noeuds appartiennent à deux composantes connexes différentes. La notion de distance géodésique nous permettra d'étudier globalement un graphe en observant si sa structure permet des distances géodésiques plus ou moins grandes entre ses noeuds. Nous nous servirons alors de la moyenne des distances géodésiques $\frac{2}{n(n-1)} \sum_{i \neq j} d(i,j)$.

Le diamètre d'un graphe est la plus grande distance géodésique entre les noeuds de ce graphe.

Le coefficient de clustering mesure pour un noeud d'un graphe la proportion de ses voisins connectés entre eux. Soit $A_i = \{(j,k) \in A \mid j \in V(i) \quad et \quad k \in V(i)\}$ l'ensemble des arrêtes de G qui connectent les voisins de i. Le coefficient de clustering de i est alors défini par : $C_i = \frac{2|A_i|}{deg(i)(deg(i)-1)}$. Cette dernière notion est importante pour étudier la modularité d'un graphe, en observant si ses arrêtes ont tendance à se regrouper dans des régions particulières du graphe, alors appelées modules, ou s'ils sont plutôt réparties de manière homogène. Nous nous servirons alors de la moyenne des coefficients de clustering des noeuds d'un graphe.

Chapitre 3

Réseaux de régulation de gènes

Nous consacrons ce chapitre à l'étude des RRGs, en commençant par décrire les différents type de données utilisées pour modéliser les RRGs puis en analysant les caractéristiques topologiques de deux modèles bien établis de RRG. La dernière partie de ce chapitre sera consacrée à une revue générale des différentes familles de méthodes de reconstruction des RRGs à partir des données d'expression des gènes.

3.1 Modélisation des RRGs

Différents types de données expérimentales sont utilisées pour inférer les relations causales de régulation entre gènes. Les données les plus informatives sont fournies par des expériences d'immunoprécipitation de la chromatine (ChIP-chip pour Chromatin ImmunoPrecipitation on chip) révélant les liaisons physiques entre FTs et régions promotrices d'un gène. Ces données peuvent ensuite être combinées à une analyse des séquences pour identifier de courtes séquences conservées dans les régions promotrices, révélant des sites potentiels de liaison de FTs. Les données d'expression des gènes sont une autre source pour l'inférence des RRGs. Elles se présentent sous la forme de matrices donnant les mesures d'expressions d'un grand nombre de gènes pour un nombre beaucoup plus petit d'observations. Dans le cadre le plus simple, les observations correspondent à différents réplicas biologiques col-

lectés indépendamment, on parle alors de données d'expression de gènes à l'équilibre. Des expériences plus sophistiquées permettent de produire des séries temporelles ou d'appliquer des perturbations contrôlées au système avant de mesurer l'expression des gènes. Les données Knock Out / Knock Down (KO / KD) sont des exemples de données de perturbations dans lesquelles les observations sont obtenues après avoir baissé ou augmenté l'expression de gènes choisis. Ces dernières sont les données d'expression les plus faciles à interpréter puisque les gènes différentiellement exprimés, avant et après une perturbation d'un gène particulier, sont susceptibles d'être régulés par ce gène (de manière directe ou indirecte).

En combinant données ChIP-chip, données de perturbations, et données de conservation des séquences, les réseaux de régulation des génomes entiers de *Saccharomyces cerevisiae* [12, 13] et de *Escherichia coli* [14] ont été établis et sont à ce jour les modèles de RRG les plus fiables dont nous disposons. En raison du cout élevé des expériences ChiP-chip à large échelle, il est toutefois rare de disposer de l'ensemble de ces données quand on recherche à reconstruire le RRG d'un système. En revanche, le développement rapide des technologies telles que "microarrays" ou "RNA-seq", que nous décrirons au chapitre 3.3.1, permettant de mesurer l'expression d'un grand nombre de gènes simultanément et à cout relativement réduit, a permis d'accumuler de grandes bases de données d'expression. Il s'en est suivi une forte demande pour des méthodes computationnelles permettant de traiter ces données pour en inférer les RRGs sous-jacents. Ainsi, les quinze dernières années ont vu l'émergence de grand nombre de méthodes visant à répondre à cette demande.

3.2 Caractéristiques des RRGs

Dans ce qui suit, nous allons mettre en évidence les caractéristiques générales des RRGs en analysant les deux RRGs mentionnées plus haut. Comme indiqué dans la section précédente, ils ont été obtenus en combinant données Chip-chip, données de conservation pour l'identification des séquences de fixation des TFs et données d'ex-

pression KO/KD. Nous détaillerons la méthodologie de modélisation de ces RRGs en dernière partie de ce chapitre, quand nous introduirons les bases de données DREAM dont ils ont été tirés. Nous considérons simplement pour l'instant ces RRGs comme les modèles les plus proches de la réalité des interactions de régulation *in vivo* dont nous disposons. Nous allons donc les étudier de plus près pour en faire sortir des caractéristiques générales. Certaines de ces caractéristiques sont souvent utilisées par les méthodes pour restreindre l'espace de recherche aux réseaux dont les topologies satisfont ces caractéristiques.

La Figure 3.1 donne une représentation du RRG d' *E. coli* et permet de se faire une idée de la grande complexité de ce réseau. Pour en faire une analyse plus détaillée et en extraire les propriétés topologiques générales, nous caractérisons ce réseau ansi que celui de *S. cerevisiae* par les mesures numériques d'analyse des graphes introduites dans le chapitre 2.4.

Une première approche d'analyse d'un réseau est d'en observer la **distribution des degrés**. Elle permet une caractérisation importante de la structure globale d'un réseau en représentant, pour un degré particulier, la fraction des noeuds du réseau qui ont ce degré. Soit $G = (N, A)$ un réseau avec $|N| = n$ et soit n_d le nombre de noeuds de degrés d dans G. La proportion des noeuds de degré d vaut alors $P(d) = \frac{n_d}{n}$ (nous notons $P(d)$ la probabilité qu'un noeud ait pour degré d). La distribution des degrés du réseau G est alors l'ensemble des valeurs $P(d)$ en fonction des degrés du réseaux d. Beaucoup d'études ont montré que les réseaux empiriques (réseaux qui modélisent des interactions observées entre des entités du monde réel) affichent des distributions de degrés proches des lois de puissance ("power law") [16, 17, 18]. La fréquence des noeuds de degré d s'écrit alors :

$$P(d) = \lambda d^{-\gamma} \tag{3.1}$$

Où λ et γ sont des réels positifs. Les réseaux dont la distribution des degrés suit une loi de puissance sont dit invariants d'échelle ou sans-échelle (*scale-free*) avec

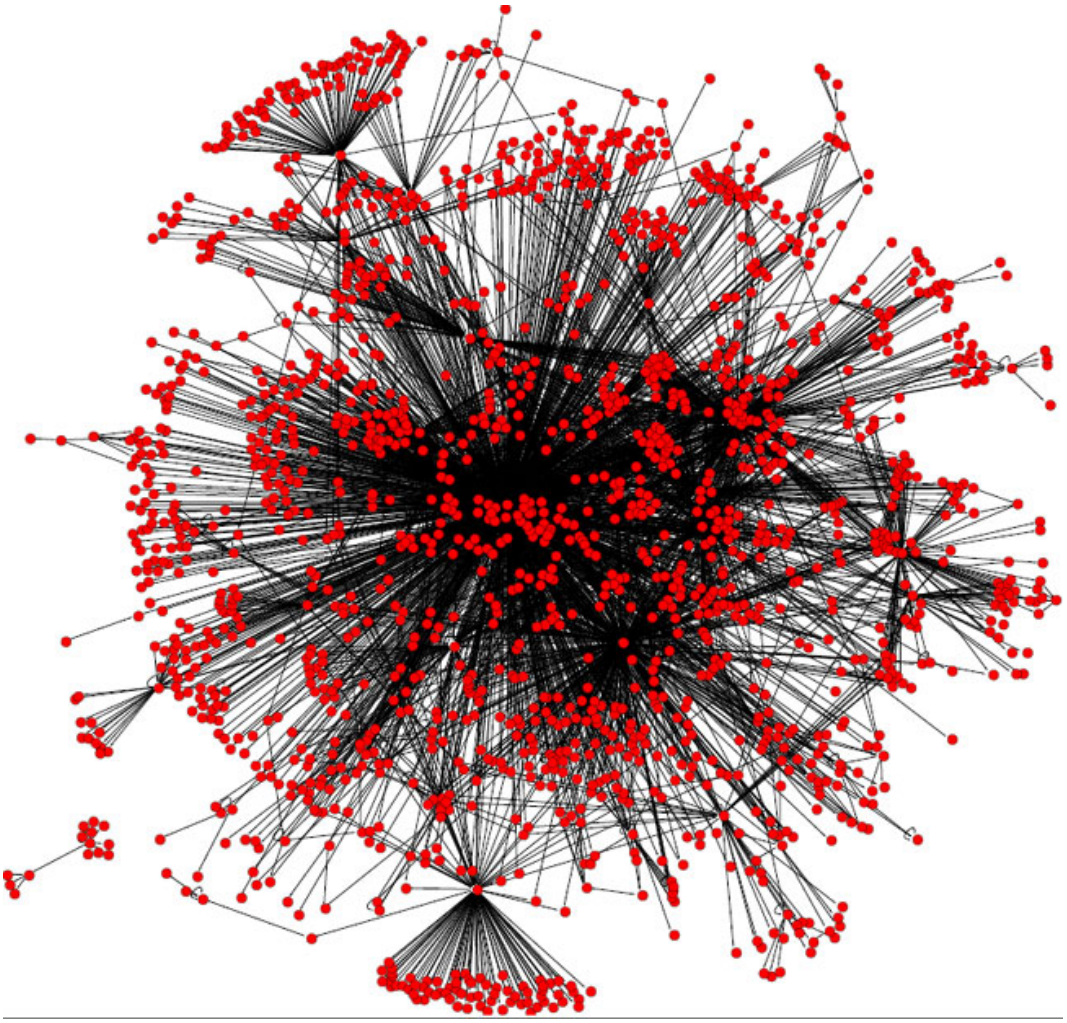


FIGURE 3.1 – RRG d'*E. coli* [15] révélant la grande complexité des réseaux de régulation.

γ pour paramètre d'échelle [16, 17]. En prenant les logarithmes des deux cotés de l'équation 3.1, on peut vérifier qu'on obtient une relation linéaire entre $log(p(d))$ et $log(d)$:

$$log(P(d)) = -\gamma log(d) + log(\lambda) \tag{3.2}$$

On peut ainsi mesurer, pour un réseau particulier, le caractère sans échelle de sa

topologie, en mesurant l'intensité de la relation affine entre $log(P(d)$ et $log(d)$:

$$SE = cor(log(P(d)), log(d))^2 \tag{3.3}$$

La fonction *cor* utilisée dans l'équation (3.3) est une fonction de corrélation linéaire qui peut se calculer à travers une regression linéaire ou directement par des scores statistiques comme la corrélation de Pearson. Ces deux notions seront définies plus loin dans ce rapport quand on présentera les différentes méthodes de reconstruction des RRGs. Pour l'instant on se contente de voir *cor* comme une fonction qui prend deux séries numériques et renvoie un nombre entre -1 et 1. $cor(x, y) = 1$ (resp. $cor(x, y) = -1$) quand y peut s'écrire comme une fonction affine de x à pente positive (resp. négative). $cor(x, y) = 0$ quand les deux séries ne présentent aucune relation linéaire. Le score SE est donc entre 0 et 1. On dira que les réseaux dont les scores SE sont proches de 1 ont des topologies proches de celles des réseaux sans échelle.

La Figure 3.2 montrent les distributions des degrés des deux réseaux de référence de *S. cerevisiae* et *E. coli* (en haut) et les transformations en logarithmes de ces distributions, $log(P(d))$ en fonction de $log(d)$ (en bas). Dans les graphiques en bas, les droites en orange montrent les meilleures associations linéaire entre $log(P(d))$ et $log(d)$. Pour un réseau donné, plus la courbe $log(P(d))$ en fonction de $log(d)$ est proche de la droite verte, plus sa distribution des degrés est proche d'une loi de puissance, et plus la topologie du réseau considéré ressemble à celle d'un réseau sans échelle. On voit que les deux distributions ne se rapprochent que partiellement de la loi de puissance, les scores de corrélation SE valant 0.74 et 0.75 pour respectivement le réseau d'*E.coli* et celui de *S. cerevisiae*. La Figure 3.2 montrent que les distributions des degrés de ces deux réseaux sont mieux modélisées par des lois de puissance tronquées exponentiellement (*exponentially truncated power law*) [19] représentée par les courbes vertes. De manière générale, cette loi est définie par :

$$P(d) = \lambda d^{-\gamma} e^{-\alpha r} \tag{3.4}$$

Ce qui devient après transformation en logarithme :

$$log(P(d)) = -\gamma log(d) - \alpha d + log(\lambda) \tag{3.5}$$

On peut ici, comme pour le cas de la loi de puissance, voir à quel point la distribution des degrés d'un réseau se rapproche d'une loi puissance exponentiellement tronquée. Pour ce faire, on recherche la combinaison linéaire de d et $log(d)$ qui se rapproche le plus de $log(P(d))$, puis on mesure à quel point cette transformation linéaire est effectivement proches de $log(P(d))$, c'est le principe de la régression linéaire multiple. On obtient là encore des scores de concordance à la loi entre 0 et 1. Ces scores sont égaux à 0.88 et à 0.9 pour respectivement *E. coli* et *S. cerevisiae* ce qui prouve que cette loi est effectivement un meilleur modèle pour la distribution des RRGs.

L'observation de la distribution des degrés permet de dégager une première caractéristique des réseaux RRGs, qu'ils ont en commun avec beaucoup de réseaux empiriques (réseaux sociaux par exemple) et plus généralement avec tous les réseaux similaires aux réseaux sans échelle. La grande majorité des noeuds de ces réseaux a un degré faible alors qu'un petit nombre de noeuds ont des degrés relativement très élevés. Les noeuds de degrés forts sont appelés hubs. On peut par exemple prendre le cas du RRG d'*E.coli* où 1077 des 2055 arêtes totales (54%) sont incidentes aux 1% des noeuds les plus connectés. Les hubs de tels réseaux ont un importance capitale, autant d'un point de vue structurel, que d'un point de vue fonctionnel.

D'un point de vue structurel, ces noeuds confèrent au réseau une remarquable résilience aux erreurs aléatoires[17, 20]. En effet, une erreur aléatoire affectera très probablement un des noeuds très majoritaires faiblement connectés, ce qui aura pas ou peu d'incidence sur la connectivité globale du réseau. Cette propriété a été

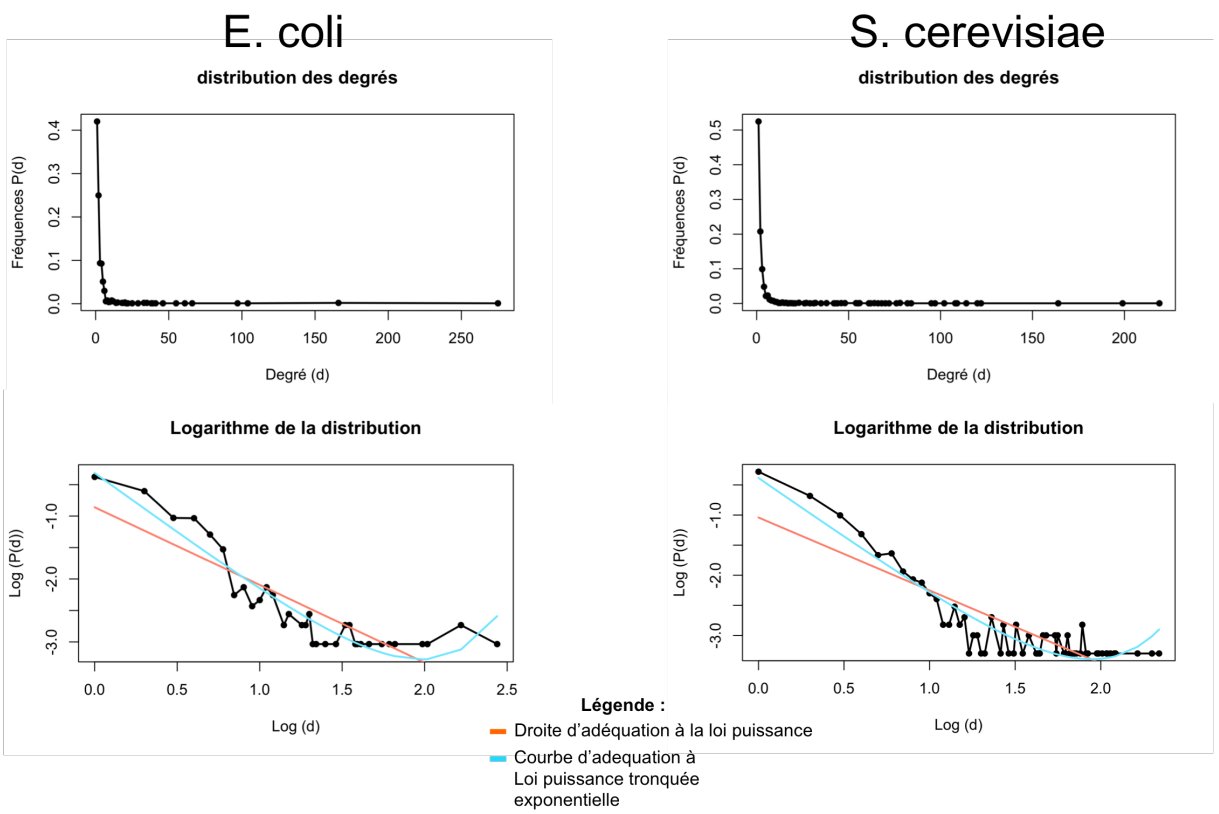


FIGURE 3.2 – Les distribution des degrés des RRGs de référence d'*E.coli* et de *S. cerevisiae*. Les droites en orange correspondent aux droites de regression linéaire entre $log(P(d))$ et $log(d)$. Elles correspondent aux lois de puissances caractéristiques des topologies sans échelle. Les distributions des degrés sont plus proches des courbes en vert qui correspondent à des distributions en loi de puissance tronquée par une exponentielle.

analysée dans [20] en utilisant les modèles issus de la théorie de la percolation [21], théorie qui étudie la fragmentation d'un réseau engendrée par une suppression d'un sous-ensemble aléatoire de ses noeuds. Il a ainsi été prouvé que pour un large éventail de réseaux sans échelle, le seuil critique de percolation est nul. Cela signifie que pour fragmenter un réseau scale-free, c'est à dire casser une de ses composantes connexes en deux ou plusieurs composantes connexes plus petites, il faut virtuellement enlever

tous les noeuds qui la constituent.

Si les hubs constituent une force contre les erreurs aléatoires, ils sont une faiblesse contre des attaques ciblées. En effet, si on supprime ses hubs, un réseau devient morcelé en un grand nombre de petites composantes connexes, ce qui le rend défaillant. On voit là l'intérêt thérapeutique évident que représentent les hubs d'un RRG, car ce sont des cibles intéressantes pour des médicaments qui visent à endommager un système biologique particulier (une bactérie nocive par exemple).

D'un point de vue fonctionnel, identifier les hubs d'un réseau permet de mieux le comprendre. Ça permet également d'avoir des leviers d'action pour le contrôler. Revenons pour illustrer ce point au contexte biologique qui nous intéresse. La Figure 3.3 montre que les arêtes des hubs sont essentiellement sortantes aussi bien dans le RRG de yeast que dans celui d'E.coli. Les hubs des RRGs sont donc des TFs qui régulent beaucoup d'autres gènes. Cette caractéristique des RRGs fait sens d'un point de vue biologique. D'une part, il n'y a pas grand intérêt d'avoir un grand nombre de régulateurs pour un même gène cible. La combinatoire de régulation qui émerge des structures avec seulement deux régulateurs qui contrôlent l'expression d'un même gène cible, permet déjà des programmes de régulation subtils du gène cible (voir la dynamique des motifs des réseaux de régulation[22]). D'autre part, il est plus parcimonieux d'assigner la régulation d'un nouveau gène, apparu par duplication d'un gène déjà présent dans le système, aux TFs qui régulent le gène à l'origine de la duplication. Ainsi, plus le nombre de gènes régulés par un TF est grand, plus ce gène a de chance de gagner de nouvelles régulations (arêtes sortantes dans le RRG) au cours de l'évolution du RRG. C'est le principe du riche qui devient plus riche ("rich gets richer"). Ce principe est au coeur d'un modèle génératif de réseaux sans échelle proposé dans [16] et que nous décrirons plus loin.

Lorsqu'un chercheur s'intéresse à un ensemble de gènes impliqués dans une fonction biologique particulière, s'il identifie les hubs du RRG qui contrôle l'expression

de ces gènes, il saura quels régulateurs clés sont impliqués dans la fonction biologique qui l'intéresse. Il pourra également, en contrôlant l'expression des hubs (en petit nombre par rapport à l'ensemble des gènes), contrôler l'ensemble du système. Yamakana a par exemple montré que par le contrôle de l'expression de seulement quatre TFs, il pouvait reprogrammer une cellule somatique en cellule souche[23]. La reprogrammation cellulaire est un secteur clé de recherche en biologie moléculaire et pourrait permettre de grandes avancés en médecine. Reprogrammer une cellule d'un individu permettrait par exemple de fournir une resource personnalisée de tissus pour reconstituer les cellules perdues dans le cas d'une maladie dégénérative. La reprogrammation cellulaire permettrait également de générer du sang synthétique en laboratoire.

On comprend donc le grand intérêt qu'il y a à identifier les hubs d'un RRG. Pour identifier les hubs, il faut d'abord reconstruire le RRG, c'est en tout cas l'approche classique : Un chercheur commence par sélectionner les gènes impliqués dans une fonction biologique qui l'intéresse, il se sert alors d'une méthode disponible pour reconstruire le RRG de ces gènes (dans la majorité des cas une méthode computationnelle appliquée à des données d'expression), il se concentre enfin sur les hubs du RRG reconstruit. C'est ce protocole que met en place WGCNA[24, 25], la méthode de reconstruction de RRGs à partir de données d'expression la plus utilisée dans les récentes publications en biologie des systèmes. Nous décrirons cette méthode dans le chapitre dédié à l'état de l'art des méthodes d'inférence des RRGs à partir des données d'expression. La nouvelle méthode développée au cours de ce travail de thèse sera alors détaillée dans le chapitre suivant. L'idée centrale derrière notre méthode, que nous présentons ici par anticipation, est de considérer le problème dans l'autre sens : identifier les hubs, c'est reconstruire le réseau. Nous commençons par identifier les hubs directement à partir des données d'expression, puis nous reconstruisons le RRG sur la base de ces hubs. C'est ce changement de perspective qui rend notre méthode fondamentalement différente de toutes celles déjà développées et qui permet

E.coli

ID gène	Degré sortant	Degré entrant	Degré total
G55	275	1	276 (#1)
G268	166	1	167 (#2)
G333	166	1	167 (#3)
G92	104	0	105 (#4)
G23	95	3	98 (#5)
G769	0	8	8
G876	0	8	8
G3221	0	8	8
G4125	0	8	8
G110	27	6	33

Yeast

ID gène	Degré sortant	Degré entrant	Degré total
G41	217	2	219 (#1)
G138	199	0	199 (#2)
G76	164	0	164 (#3)
G1	120	3	123 (#4)
G224	120	0	120 (#5)
G1251	14	0	14
G1773	14	0	14
G3339	14	0	14
G3588	14	0	14
G4176	14	0	14

FIGURE 3.3 – **Tableau des gènes les plus connectés.** Les parties supérieures des tableaux indiquent les degrés entrants et sortants des cinq noeuds de plus fort degré sortant, classé par ordre décroissant de degré sortant. Les parties inférieures des tableaux indiquent les degrés entrants et sortants des cinq noeuds de plus fort degré entrant, classé par ordre décroissant de degré entrant. La dernière colonne indique le degré total. On voit que les noeuds avec les plus d'arêtes sortantes sont les hubs du réseau (le classement par ordre décroissant de degré total des cinq hub des réseaux est indiqué entre parenthèse dans la dernière colonne)

un saut de performances.

Poursuivons à présent la caractérisation topologique des RRGs de *S. cerevisiae* et d'*E.coli*. Le tableau de la Figure 3.4 présente les mesures numériques des deux RRGs comparativement à deux réseaux empiriques et deux réseaux générés par deux modèles que nous allons décrire. Les réseaux empiriques sont le réseau du métro parisien[26] et le réseau social des membres actif de Facebook en 2011[27]. Les

deux modèles de génération sont celui d'Erdös-Rényi [28] pour les réseaux aléatoires (colonne E-R dans la Figure 3.4) et celui de Barabási-Albert [16] pour les réseaux sans échelle (colonne B-A dans la Figure 3.4).

Le premier modèle est paramétré par le nombre de noeuds n et le nombre d'arêtes a et génère un réseau uniformément au hazard parmi tous les graphs possibles de n noeuds et a arêtes. La génération commence par un réseau vide à n noeuds puis connecte a paires de noeuds aléatoirement, chaque paire ayant la même probabilité $p = \frac{2a}{n(n-1)}$ de former une arête.

Le deuxième modèle s'appuie sur le principe du "rich get richer" mentionné plus haut rebaptisé attachement préférentiel dans [16]. Ce modèle est itératif et est paramétrisable par le nombre de noeuds n et le nombre d'arêtes a_n à ajouter à chaque itération. La construction commence par un réseau à a_n noeuds sans arrêtes. On ajoute alors $n - a_n$ noeuds itérativement, et à chaque nouveau noeud ajouté, on l'attache avec a_n arêtes aux noeuds existants, en privilégiant les noeuds dont les degrés sont les plus élevés. Pour une description plus précise de ce modèle, plaçons nous à une itération au cours de la construction, et supposons que le réseau déjà construit a n' noeuds, et chaque noeud i du réseau a un degré d_i. On ajoute alors un nouveau noeud et a_n arêtes. Chaque arête ajoutée relie le nouveau noeud à un noeud déjà crée i avec la probabilité $p(i) = \frac{d_i+1}{\sum_{k=1}^{n'} d_k+1}$. Le réseau final aura donc $a_n(n - a_n)$ arêtes.

Les paramètres des deux modèles définis ici sont fixés de sorte à ce que les réseaux crées soient les plus proches possible du réseau d'*E. coli* en terme de nombre de noeuds et de nombre d'arêtes. Les paramètres du modèle aléatoire d'Erdös-Rényi sont donc $n = 1081$ et $a = 2055$ et ceux du modèle avec attachement préférentiel de Barabási-Albert sont $n = 1081$ et $a_i = 2$, pour obtenir un réseau final avec 2158 arêtes. Les valeurs des colonnes B-A et E-R du tableau de la Figure 3.4 ont été obtenues en moyennant les mesures correspondantes sur 100 réseaux générés avec

les même paramètres.

Une première observation à faire à partir de la Figure 3.4 est que les deux RRGs sont peu denses. *E. coli* contient par exemple seulement 2055 régulations (arêtes) pour 1081 gènes (noeuds). Il faut toutefois noter que les RRGs d' *E. coli* et de *S. cerevisiae*, considérés comme vrais (ou réels) pour les analyses de topologie et les évaluations des méthodes, manquent probablement de certaines régulations en raison de la méthodologie conservatrice de leur modélisation dont le but été de produire les réseaux de régulation les plus sûrs possibles. On reviendra plus en détail sur cette méthodologie quand on décrira les bases de données dont ils ont été extraits.

On peut remarquer ensuite que les rapports entre degrés maximaux et degrés moyens sont nettement plus grands pour les deux RRGs le réseau social Facebook et le réseau sans échelle de Barabási-Albert, que pour le réseau aléatoire d'Erdös-Rényi et le réseau du métro parisien (notons que le réseau social Facebook empêche des noeuds de degrés supérieurs à 5000). On retrouve dans les trois premiers réseaux la caractéristique des topologies dominées par des hubs. En comparant ces rapports entre le RRG d'*E. coli* et le réseau sans-échelle de Barabási-Albert (qui est de même taille que le RRG d'*E. coli*), on peut également observer que le degré maximal est encore plus distant du degré moyen dans le RRG. Cette différence peut également être observée dans les distributions des degrés de la Figure 3.2, où les queues des courbes $log(P(d))$ (les valeurs de $log(P(d))$ pour d élevé) sont notablement au dessus des droites oranges qui symbolisent les réseaux sans échelle. Ceci tend à montrer que les RRGs sont plus fortement dominés par leurs hubs que les réseaux sans échelle.

Une autre caractéristique importante des réseaux sans échelle et leur petite distance géodésique moyenne. Pour rappel, la distance géodésique moyenne et le nombre moyen d'arêtes à parcourir au minimum pour aller d'un noeud du réseau à un autre. Cette caractéristique permet d'estimer la qualité de la connectivité globale d'un réseau. On voit dans la Figure 3.4 que la distance géodésique moyenne du RRG d'*E.*

coli est comparable à la distance géodésique moyenne du réseau sans échelle de même taille (généré par le modèle de Barabási-Albert) alors que la distance géodésique moyenne du réseau aléatoire de même taille (généré par le modèle d'Erdös-Rényi) est environ 60% plus grande. Le réseau social Facebook présente la plus petite distance géodésique (4.7) relativement à sa très grande taille (720 millions de noeuds) et sa très petite densité ($2.6e^{-10}$). Cette valeur a baissé à 3.5 dans le réseau de 2016 [29]. Les réseaux avec une distance géodésique moyenne ont une structure décrite de petit monde ("small world"). Il est intéressant de noter que bien qu'il soit le plus dense en arêtes, le réseau du métro parisien est le moins bien connecté, avec une distance géodésique moyenne supérieure à 11 stations (le trajet de longueur minimal qu'il faut, en moyenne, pour passer d'une station à une autre, passe par 10 autres stations.)

La dernière caractéristique que nous abordons ici est celle du coefficient de clustering des réseaux. Dans le RRG d'*E.coli*, ce coefficient est égal à 2%, ce qui signifie que seulement 2% des structures de chaîne x-y-z, où le noeud x est adjacents à deux autres noeuds y et z, sont des triangles avec y adjacent à z. Dans le RRG de *S. cerevisiae* ce coefficient est encore plus petit et vaut 1.3%. Il faut cependant noter que les structures triangulaires sont plus fréquentes dans le RRG d'*E.coli* que dans le réseau sans échelle (modèle Barabási-Albert) ou le réseau aléatoire (modèle d'Erdös-Rényi) de même taille, surement à cause de l'intérêt biologique que présentent les motifs de régulation à 3 gènes adjacents (par exemple les boucles rétroactives [22]). Les méthodes d'inférence des RRGs à partir des données d'expression vont toutefois préférer fortement pénaliser les structures triangulaires dans les réseaux reconstruits, voire les interdire. Nous expliquerons pourquoi quand nous présenterons une revue de ces méthodes.

	Ecoli	Yeast	Paris	FB	B-A	E-R
# noeuds	1081	1994	295	721M	1081	1081
# arrêtes	2055	3935	383	68.7M	2158	2055
densité	3.5e-3	2e-3	8.8e-3	2.6e-10	3.7e-3	3.4e-3
Degré maximal	275	219	9	5000	167	11.7
Degré minimal	1	1	1	1	1.9	0
Degré moyen	3.8	3.9	2.3	99	4.1	3.8
Degré max / degré moyen	72.36	56.15	3.91	50.5	40.73	3.07
# composantes connexes	26	4	1	132	1	22.1
Taille de la plus grande composante connexe	985	1984	295	720.3M	1081	1055
Coefficient de clustering	2%	1.3%	1.6%	8.51%	0.7%	0.5%
Distance géodésique moyenne	4.2	3.9	11.77	4.7	4.34	6.8
diamètre	13	9	33	_	7	11

FIGURE 3.4 – **Tableau des caractéristiques topologiques.** Les colonnes Yeast et E.coli correspondent aux deux RRGs analysés dans ce chapitre. La colonne Paris correspond au réseau du métro parisien. Les deux autres colonnes sont générés, B-A par le modèle de génération des réseaux sans échelle, et E-R par le modèle de génération des réseaux aléatoire.

3.3 Reconstruction des RRG à partir des données d'expression

Nous présentons dans cette section les différentes méthodes de reconstruction des RRGs à partir des données d'expression. Nous commençons par décrire les technologies permettant de mesurer l'expression des gènes puis nous présenterons une revue des différentes approches développées pour traiter ces données dans le but de

reconstruire des RRGs.

3.3.1 Mesure des niveaux d'expression des gènes

La mesure d'expression par puces à ADN a longtemps été la biotechnologie la plus couramment utilisée pour mesurer les niveaux d'expression transcriptionnelle des gènes. Sur une puce (en verre, silicium ou plastique) sont rangées des sondes contenant des fragments d'ADN synthétiques représentatifs des gènes dont on veut mesurer l'expression. Chacun de ces fragments est une molécule d'ADN simple brin qui se lie spécifiquement à une séquence complémentaire par hybridation. En effet, l'ADN dénaturé (simple brin) retrouve spontanément sa forme naturelle en double hélice lorsqu'il est en présence d'un brin complémentaire. Les deux brins apparient leurs bases complémentaires A=T et G=C par des liaisons hydrogènes. Pour mesurer l'expression des gènes d'un échantillon, on en extrait les ARNm transcrits de ces gènes, avant de les convertir en ADNc grâce à une enzyme transcriptase inverse. Ensuite, l'ADNc est marqué avec un colorant fluorescent avant d'être appliqué sur la puce lui permettant de s'hybrider avec les sondes. Après l'élimination de l'ADNc non hybridé, l'intensité de fluorescence de chaque sonde est mesurée révélant la concentration d'ARN transcrits et donc le niveau d'expression du gène correspondant. Une dernière étape consiste à normaliser les niveaux d'intensité de fluorescence[30]. Une seule puce peut contenir des milliers de sondes et par conséquent mesurer le niveau d'expression de tous les gènes d'une espèce. Cette technologie est relativement peu couteuse mais est affectée par beaucoup de bruit provenant de la préparation des échantillons, de la dynamique d'hybridation ou de la saturation des molécules fluorescentes. Il a été suggéré que le bruit dans les expériences de puces à ADN suit une distribution log-normale[31].

RNA-seq (pour RNA sequencing) est une technologie plus récente de mesure de l'expression des gènes reposant sur le séquençage complet des ARNm prélevés sur échantillons[32]. Dans les expériences RNA-seq, les ARN isolés sont d'abord hydro-

lysés en courts fragments appelés reads de longueur variant en général entre 100 et 400 nucléotides. Ces reads sont alors convertis en ADNc par l'action d'une transcriptase inverse avant d'être amplifiés puis séquencés. Le séquençage se fait par des technique de séquençage à haut débit permettant le séquençage de millions de reads en une traite. Les reads séquencés sont ensuite alignés sur un génome de référence permettant ainsi de déterminer l'abondance des différents transcrits[33]. Cette technologie a trois avantages majeures par rapport aux puces à ADN. Premièrement, la technologie des puces à ADN se limite à mesurer l'expression des gènes dont la séquence est déjà connue alors que RNA-seq permet de mesurer l'expression de gènes qui n'ont pas encore été identifiés. Deuxièmement, les données d'expression obtenues par RNA-seq sont moins bruitées puisqu'elles sont débarrassées du bruit qui résulte des erreurs d'hybridation. Troisièmement, la technologie RNA-seq permet d'obtenir les mesures d'expression absolues des gènes alors qu'avec les micro puces les mesures sont relatives (une condition par rapport à une autre par exemple). Cependant, les données RNA-seq sont plus couteuses à produire parce qu'elle nécessitent davantage de traitement et plus de puissance de calcul, les données d'expression générés par puces à ADN demeurent les plus répandues dans les bases de données.

Les deux technologies présentées plus haut permettent d'obtenir les valeurs d'expression des gènes exprimés dans un échantillon d'une culture cellulaire particulière. En reproduisant la procédure pour plusieurs échantillons nous obtenons plusieurs observations du système étudié, chacune caractérisée par les valeurs d'expression de l'ensemble des gènes exprimés dans ce système. Nous pouvons ainsi construire une matrice d'expression avec les gènes dans une dimension et les observations dans l'autre. Si une seule expérience de mesure permet d'obtenir les valeurs d'expression de tous les gènes du système, jusqu'à plusieurs dizaines de milliers, le nombre d'expériences différentes qu'il est possible de réaliser est limité par le cout et le temps nécessaires à la réalisation de ces expériences. Les matrices d'expression produites sont donc déséquilibrées avec beaucoup de gènes et relativement très peu d'observa-

tions. Cette caractéristique particulière aux données d'expression des gènes, connue sous le nom de la malédiction de la dimensionnalité ("dimensionality curse"), est l'une des principales raisons qui font de la reconstruction des réseaux à partir de ces données un problème si compliqué.

On distingue trois types d'observations selon la nature des expériences de préparation des échantillons à mesurer.

1. Les données temporelles sont produites en contrôlant le temps qui sépare les mesures d'expression. Ils permettent d'avoir une vue dynamique du système.

2. Les données de perturbations correspondent à des mesures d'expression après avoir perturbé le système, en empêchant par exemple l'expression d'un ou plusieurs gènes.

3. Les données d'expression du sytème à l'état d'équilibre sont les données les plus répandues. Elles correspondent à des mesures effectuées de manière indépendante. Les matrices produites sont alors traitées par les méthodes computationnelles afin d'inférer le réseau de régulation sous-jacent. Nous présentons dans la prochaine section les principales stratégies développées à cet effet.

3.3.2 Méthodes de reconstruction des RRGs à partir des données d'expressions à l'équilibre

Dans la suite de ce document nous parlerons de **réseau réel** pour représenter l'ensemble des regulations qui ont effectivement cours dans un système biologique donné. Nous supposerons que ces régulations sont bien déterminées, elles sont causales et directes. Par contraste, nous utiliserons **réseaux reconstruits** pour parler des RRGs inférés par les méthodes à partir des données. Nous cherchons à ce que le réseau reconstruit soit le plus proche possible du réseau réel. Cette section sera dédiée à la présentation détaillée des différentes familles de méthodes qui visent à traiter les données d'équilibre. Nous commencerons par une description formelle et générique des données d'entrée et de sortie des méthodes. Ces dernières seront en-

suite décrites en les groupant en trois grandes familles : réseaux de co-expression, réseaux reconstruits par sélection de régulateurs et réseaux bayesiens. Pour une description plus exhaustive, nous orientons le lecteur vers les nombreuses publications qui font la revue de ces méthodes[34, 35, 36, 37, 38].

Données d'expression à l'équilibre

Soit X une matrice d'expression de dimension $m \times n$ avec n gènes en colonnes, m observations en lignes et $m \ll n$. Chaque colonne X_i correspond au profil d'expression du gène i, c'est à dire les mesures d'expression de ce gène à travers l'ensemble des observations. D'une manière générale, les méthodes d'inférence de RRGs, reçoivent en entrée une telle matrice et renvoient une matrice d'adjacence de dimension $n \times n$ qui attribue un score à chaque paire de gène. Plus ce score est élevée plus on accorde de confiance à l'arrête liant la paire. Un réseau est reconstruit à partir de la matrice d'adjacence délivrée par une méthode en fixant un seuil et en connectant par une arrête chaque paire de gènes dont le score est supérieur au seuil.

Nous supposerons que les observations correspondent à différentes mesures à l'état d'équilibre du système, *i.e.* les lignes de X sont indépendantes. Les différentes méthodes d'inférence de RRGs à partir de ce type de données peuvent être regroupées en trois grandes famille : les réseaux de co-expression, les réseaux par sélection de régulateurs et les réseaux Bayesiens.

Réseaux de co-expression

Ces méthodes reconstruisent les RRGs en mesurant les similarités des profils d'expression des gènes. Deux gènes dont les profils d'expression présentent une grande similarité sont dits co-exprimés, ils sont conjointement faiblement ou fortement exprimés aux même observations. Deux tels gènes sont alors statistiquement dépendants laissant supposer qu'ils sont associés dans un même complexe de régulation, leur relation pouvant être directe ou indirecte. Les deux principaux scores communément utilisés pour mesurer la similarité de deux profils d'expression sont

la correlation de Pearson [39] et l'information mutuelle basée sur la théorie de l'information développée par Shannon [40].

La corrélation de Pearson se mesure directement à partir des profils d'expression de deux gènes i et j comme suit :

$$cor(i,j) = \frac{\sum_k (X_{ki} - \overline{X_{.i}})(X_{kj} - \overline{X_{.j}})}{\sqrt{\sum_k (X_{ki} - \overline{X_{.i}})^2)}\sqrt{\sum_i (X_{kj} - \overline{X_{.j}})^2}} \tag{3.6}$$

Où $\overline{X_{.l}}$ représente la moyenne des valeurs d'expression d'un gène l, (*i.e.* $\overline{X_{.l}} = \frac{\sum_k X_{kl}}{m}$).

L'information mutuelle mesure un score de similarité de deux variables discrètes (à valeurs entières). Les valeurs d'expression des gènes étant continues (à valeurs réelles), il est donc nécessaire de commencer par une étape de discrétisation des données. Supposons qu'on cherche à discrétiser une série de m valeurs réelles $z[1], z[2], ..., z[m]$ d'une variable z. Supposons également que z_{min} est la valeur minimale de z et z_{max} sa valeur maximale, *i.e.* $\forall i \in [1,m] \mid z[i] \in [z_{min}, z_{max}]$. Pour discrétiser z il faut d'abord partitionner l'intervalle $[z_{min}, z_{max}]$ en k sous-intervalles disjoints $[z_{min}, z_1[, [z_1, z_2[, ..., [z_{k-1}, z_{max}]$. On transforme alors z en z' en posant $z'[i] = 1 \quad si \quad z[i] \in [z_{min}, z_1[$, $z'[i] = k \quad si \quad z[i] \in [z_{k-1}, z_m ax[$ et $z'[i] = l \in \{2, 3, ..., k-1\} \quad si \quad z[l] \in [z_{l-1}, z_l[$. Il y a autant de façon de discrétiser z que de manières de partitionner l'intervalle initial et les solutions diffèrent en fonction du nombre de sous-intervalles et des points de coupe $z_1, z_2, ..., z_{k-1}$. On peut par exemple fixer le nombre de sous-intervalles en paramètre et tous les sous-intervalles auront la même taille. On peut également décider que tous les sous-intervalles aient le même nombre de valeurs. Le choix de la technique de discrétisation et de ses paramètres a une influence sur les réseaux reconstruits et souvent les chercheurs qui se servent des méthodes basées sur l'information mutuelle doivent jouer avec plusieurs techniques de discrétisation et leurs paramètres pour analyser les différences entre les réseaux reconstruits[41]. Supposons maintenant que la matrice d'expression X soit discrétisée en χ. Nous

notons χ_i l'ensemble des valeurs entières que peut prendre l'expression discrétisée d'un gène i, ainsi $\forall k \in [1,m], \chi_{ki} \in \chi_i$. La probabilité marginale que le gène i ait une valeur particulière χ_i de χ_i est alors égale à :

$$\forall \chi_i \in \chi_i, \quad P(\chi_i) = \frac{|\{k \in [1,m] \quad | \quad \chi_{ki} = \chi_i\}|}{m}$$

Par extension, la probabilité jointe que le gène i ait une valeur particulière χ_i de χ_i et que le gène j ait une valeur particulière χ_j de χ_j est égale à :

$$\forall \chi_i \in \chi_i \quad et \quad \forall \chi_j \in \chi_j, \quad P(\chi_i, \chi_j) = \frac{|\{k \in [1,m] \quad | \quad \chi_{ki} = \chi_i \quad et \quad \chi_{kj} = \chi_j\}|}{m}$$

Nous pouvons maintenant définir le score d'information mutuelle $IM(i,j)$ pour deux gènes i et j :

$$IM(i,j) = \sum_{\chi_i \in \chi_i, \chi_j \in \chi_j} P(\chi_i, \chi_j) log(\frac{P(\chi_i, \chi_j)}{P(\chi_i)P(\chi_j)}) \tag{3.7}$$

Il existe d'autres scores permettant de mesurer la similarité de séries numériques (par exemple les corrélation de Kendall et de Spearman) mais nous n'en parlerons pas dans cette thèse, les deux définis plus haut étant les plus largement utilisés. Ces scores sont utilisés pour identifier les gènes co-exprimés révélant des liens de régulation existant entre eux. D'une part, une corrélation ou une information mutuelle nulle impliquent une indépendance statistique, ce qui laisse supposer que les gènes correspondant ne sont liés par aucun lien de régulation. D'autre part, il est biologiquement raisonnable de penser que le signal capturé par les scores de similarité (les scores de similarité élevés) révèlent les gènes appartenant à un même complexe de régulation. Ce signal entre gènes co-exprimés peut cependant être direct ou indirect. Un signal direct est une relation causale entre un gène régulateur et sa cible, les deux gènes sont adjacents dans le RRG. Un signal indirect émerge entre deux gènes dont les profils d'expressions ont un score élevé bien qu'ils ne soient pas adjacents dans le réseau. La Figure 3.5 illustre les deux cas où se manifestent des

signaux indirects dans le cadre le plus simple de trois noeuds liés par deux arrêtes (deux régulations causales). Dans le cas d'une chaine de régulation, où un gène i régule un gène k qui à son tour régule un gène j, les gènes i et j vont évidemment afficher des comportements d'expression similaires bien qu'ils ne soient pas adjacents dans le réseau. L'autre cas correspond à une co-régulation de deux gènes i et j non adjacents mais tous deux régulés par un troisième gène k. Les profils d'expression des deux cibles co-régulées ont en effet tendance à montrer des similitudes, car l'expression de ces gènes sont sous contrôle du même régulateur k. On voit ainsi que le signal indirect est transitif émergeant par convolution du signal direct. Pour inférer les RRGs, les méthodes qui se basent sur les scores de similarité des profils d'expression doivent donc éliminer les relations indirectes. Cette étape s'appelle la **déconvolution** et chaque méthode propose sa propre solution pour répondre à ce problème. Nous présentons dans la suite quelques méthodes phares en expliquant les solutions qu'ils proposent pour la déconvolution du signal de co-expression.

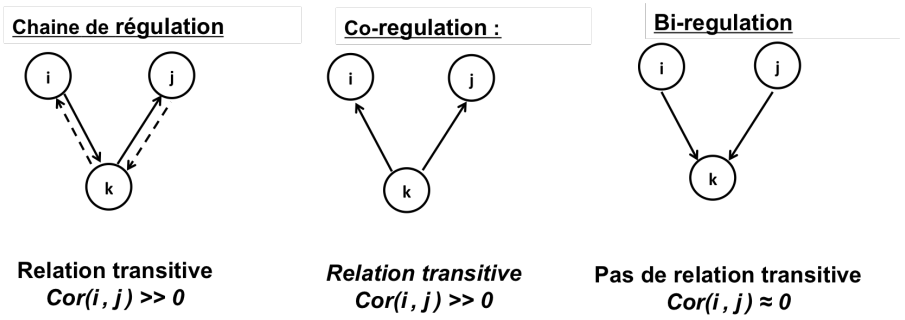


FIGURE 3.5 – Relations causales et transitives

WGCNA[24, 25] (pour *Weighted Gene Co-expression Network Analysis*) est la méthodologie la plus fréquemment utilisée dans les publications de biologie des systèmes. La première étape dans protocole WGCNA est d'élever les scores de corrélations de Pearson (ou d'autres scores de corrélations) à une puissance $\beta > 0$. Comme les valeurs absolues des scores de corrélations sont entre 0 et 1 (voir l'equation 2.1), cette opération baisse les scores de corrélation. De plus, l'écart entre les

scores de corrélation, une fois élevés à une puissance β positive, augmente exponentiellement avec la puissance β. Cette opération peut donc être considérée comme une déconvolution globale en supprimant un partie du signal qui aurait pu être retenu sans l'élévation à la puissance β. Seul le signal le plus fort est retenu. La puissance β est fixé de manière à garantir une distribution des degrés qui suit une loi de puissance (voir la section *caractéristiques des RRGs* du chapitre *Réseaux de régulation de gènes*). Il est tout de même important de noter que WGCNA ne prétend pas reconstruire des réseaux causaux mais simplement des réseaux de co-expression. La reconstruction du réseau peut même se faire après une opération de convolution du signal une fois l'opération d'élévation à la puissance β réalisée (à travers la matrice TOM, pour Topological Overlap Matrix[42]). Nous signalons pour finir que WGCNA permet également de grouper les gènes du système en modules de co-expression par un clustering hiérarchique. Nous reviendrons plus en détails sur cette étape quand on comparera la méthode de clustering de gènes développée dans cette thèse au clustering proposé par WGCNA.

D'autres méthodes basées sur la corrélation de Pearson appliquent une déconvolution locale en se servant des scores de **corrélation partielle**[43, 44]. Pour définir ce qu'est une corrélation partielle nous devons d'abord définir la notion de régression linaire. Cette notion permet d'analyser la relation linéaire entre deux variables. Considérons à titre d'illustration deux séries $x[1], ..., x[m]$ et $z[1], ..., z[m]$. Pour chercher à trouver la transformation linéaire de z qui se rapproche le plus près de x on cherche à trouver $\alpha = argmin \sum_{k=1}^{m} (x[k] - \alpha z[k])^2$. La relation linéaire s'écrit alors $x = \alpha z + r_{xz}$, où r_{xz} est le vecteur des résidus de la régression linéaire. Plus petite est la norme du vecteur r_{xz}, plus forte est la relation linéaire entre x et z. En considérant maintenant une troisième variable y, le score de corrélation partielle du couple (x, y) conditionnellement à z est simplement le coefficient de corrélation de Pearson des résidus r_{xz} et r_{yz} On obtient alors la relation :

$$Cor(x, y|z) = cor(r_{xz}, r_{yz}) = \frac{cor(x, y) - cor(x, z)cor(y, z)}{\sqrt{1 - cor(x, z)^2}\sqrt{(1 - cor(y, z)^2}} \tag{3.8}$$

$Cor(x, y|z)$ est faible quand $cor(x, y) \approx cor(x, z)cor(y, z)$. Ceci peut se produire quand $Cor(x, y) \approx 0$, $Cor(x, z) \approx 0$ $Cor(y, z) \approx 0$. Ce cas n'est pas intéressant. Si par contre $Cor(x, y)$ est élevé, une valeur basse de $Cor(x, y|z)$ peut alors signifier qu'une part significative de la corrélation $cor(x, y)$ est indirecte et s'explique par les corrélations $Cor(x, z)$ et $Cor(y, z)$. La corrélation partielle peut donc être utilisée pour retirer les effets extérieurs des scores de corrélations entre gènes et, en théorie, aboutir à des scores indiquant des dépendances directes. En pratique cette approche se montre toutefois peu efficace pour traiter les matrices de données où le nombre d'observations est inférieur au nombre de variables, ce qui est toujours le cas dans le cas des matrices d'expression de gènes [38, 34]. La notion de corrélation partielle sera

ARACNE[45] (pour *Algorithm for the Reconstruction of Accurate Cellular NEtworks*) est l'une des première méthode d'inférence des RRGs basée sur l'information mutuelle. Elle applique une déconvolution locale en considérant pour chaque triplet de gènes que c'est la paire, dont le score d'information mutuelle est le plus petit, qui est en relation indirecte (en s'appuyant sur *data processing inequality*[46]).

CLR[47] (pour *Context Likelihood or Relatedness network*) est une autre méthode d'inférence des RRGs basée sur l'information mutuelle. Cet algorithme corrige le score d'information mutuelle d'une paire de gène par apport à la distribution empirique des scores d'informations mutuelle de chaque gène de la paire. En pratique CLR attribue un score d'adjacence à chaque paire de gène i, j comme suit :

$CLR(i, j) = \sqrt{(CLR(i)^2 + CLR(j)^2)}$ avec

$CLR(i) = max(0, \frac{IM(i,j) - \overline{IM(i,.)}}{<IM(i,.)>})$ et $CLR(j) = max(0, \frac{IM(i,j) - \overline{IM(j,.)}}{<IM(j,.)>})$

Pour tout gène k, $\overline{IM(k,.)}$ et $< IM(k,.) >$ sont respectivement la moyenne et l'écart type des scores d'information mutuelle du gène k aux autres gènes, définis comme suit :

$$\overline{IM(k,.)} = \frac{sum_{l=1}^{N} IM(k,l)}{n} \quad \text{et} \quad < IM(k,.) >= \sqrt{\frac{\sum_{l=1}^{N} (IM(k,l) - \overline{IM(i,.)})^2}{n-1}}$$

C3NET[48] (pour *Conservative Causal Core Network*) est une autre méthode basée sur l'information mutuelle. Elle applique une déconvolution conservatrice en limitant à un le nombre d'arrêtes par gène. Chaque gène k est connecté au gène $l = argmax \quad IM(k,.)$ (chaque gène est lié au gène avec qui son score d'information mutuelle est maximal.) Le réseau reconstruit a au maximum n arrêtes et essaye de retrouver une structure de base du réseau réel. BC3net[49] (pour *bagging CENET*) applique l'algorithme C3NET sur plusieurs matrices de données, chacune générée par sous-échantillonnage de la matrice initiale (bootstrap[50].) La matrice d'adjacence finale est obtenue par consensus des matrices obtenues à chaque application de C3NET.

Réseaux par selection de régulateurs

L'inférence des RRGs peut également être vue comme une sélection des régulateurs à assigner à chaque gène ("feature selection problem"). Nous avons mentionné la faible densité en arrêtes des RRGs quand nous avons analysé leurs caractéristiques topologiques, ces méthodes intègrent *a priori* cette propriété en pénalisant les structures avec beaucoup de régulateurs pour un gène.

Tigress[51] (pour *Trustful Inference of Gene REgulation with Stability Selection*) sélectionne les régulateurs pour chaque gène en appliquant l'algorithme Lasso[52]. Cet algorithme permet de sélectionner les variables les plus prédictives d'une variable réponse en résolvant un problème de regression linéaire multiple avec une pénalisation de la norme l_1 des coefficients de regression. Nous avons défini plus haut l'equation qui permet de calculer le coefficient de regression linaire lorsqu'il

n'y a qu'une variable explicative (voir le paragraphe dédié à la corrélation partielle de la section précédente.) Cette équation s'étend naturellement quand il y a k variables explicatives $z_1, ..., z_k$. Il s'agit alors de trouver la fonction linéaire de ces variables qui se rapproche le plus de la variable réponse x, ce qui revient à résoudre le problème : $\alpha_1, ..., \alpha_k = argmin\{\sum_{l=1}^{m}(x[l] - \sum_{j=1}^{k}\alpha_j z_j[l])^2\}$. La transformation linéaire s'écrit alors $x = \alpha_1 z_1 + \alpha_2 z_2 + ... + \alpha_k z_k + r_{xz}$ où r_{xz} est le vecteur des résidus de la régression linéaire multiple. Là encore, plus petite la norme de r_{xz} plus forte est la relation linaire entre variables explicatives et variable réponse. L'algorithme Lasso ajoute à ce problème d'optimisation (trouver les coefficients de régression) une pénalité sur le nombre de coefficients non nuls. Le problème s'écrit alors : $\alpha_1, ..., \alpha_k = argmin\{\sum_{l=1}^{m}(x[l] - \sum_{j=1}^{k}\alpha_j z_j[l])^2 + \lambda \sum_{j=1}^{k}|\alpha_j|\}$

La contrainte sur la norme des coefficients va faire que certains de ces coefficients seront nuls (ou presque nuls). Le paramètre λ permet de contrôler l'effet de la pénalité. En l'augmentant, on réduit le nombre de coefficient non nuls. Plus un coefficient est élevé, plus la variable descriptive associée est fortement impliqué dans l'expression de la variable réponse. Dans le cas de l'inférence des régulateurs pour un gène, ce dernier est considéré comme une variable réponse et tous les autres gènes comme variables descriptives. Pour stabiliser les résultats, la méthode Tigress applique plusieurs fois l'algorithme Lasso pour un gène donné, et à chaque fois retient les cinq gènes de plus forts coefficients. Le score final attribué à chaque pair régulateur - gène cible est alors proportionnel au nombre de fois où le régulateur a fait partie des gènes dont le coefficient est parmi les cinq plus élevés.

MRNET [53] (pour *Minimum Redundancy NETworks*) s'appuie sur les scores d'information mutuelle et la technique de sélection de variables connue sous le nom de *Minimum Redundancy Maximum Relevance* (MRMR)[54]. C'est une méthode agglomérative qui pour chaque gène cible x, commence par sélectionner parmi les autres gènes celui qui maximise le score d'information mutuelle avec x. La sélection des pro-

chains régulateurs se fait alors de manière itérative, à chaque itération le nouveau régulateur est sélectionné parmi ceux qui n'ont pas encore été choisi comme celui qui maximise l'information mutuelle avec le gène cible x (pertinence maximale *maximum relevance*), et qui minimise l'information mutuelle avec les gènes déjà sélectionnés comme régulateurs (redondance minimum *minimum redundancy.*) Concrètement, si on nomme $z_1, ..., z_{k-1}$ les gènes déjà sélectionnés, le k-ième régulateur est choisi comme suit : $z_k = argmax\{IM(z_k, x) - \frac{\sum_{l=1}^{k-1} IM(z_k, z_l)}{k-1}\}$

GENIE3 [55] (pour *GEne Network Inference with Ensemble of trees*) se sert d'une forêt d'arbres de régressions pour procéder à la sélection des régulateurs d'un gène. Comme la regression linéaire, un arbre de régression permet d'estimer une variable réponse par un ensemble de variables descriptives. Pour décrire le fonctionnement de cette méthode, nous posons x un gène cible (variable réponse) et $z_1, ..., z_k$ l'ensemble des autre gènes (variables descriptives). On démarre avec l'ensemble O des m observations disponibles et on calcule la dispersion de x sur O par $D_O(x) = \sum_{l \in O} (x[l] - \overline{x})^2$ où $\overline{x}$ est la moyenne empirique (*i.e.* $\overline{x} = \frac{\sum_{l \in O} x[l]}{m}$). On commence la construction de l'arbre de régression en cherchant la partition de O en deux sous ensembles d'observations disjoints O_1 et O_2 de telle sorte à maximiser la baisse de la dispersion de x définie par $\Delta D_O(x) = D_O(x) - D_{O_1}(x) - D_{O_2}(x)$. O_1 et O_2 sont recherchés par rapport à tous les gènes. Pour un gène z_j, on parcourt toutes les valeurs γ permettant une nouvelle partition de l'ensemble des observations O en $O_1 = \{s \in O \mid z_j[s] < \gamma\}$ et $O_2 = \{s' \in O \mid z_j[s'] > \gamma\}$. Cette procédure est alors réitérée à chacun des deux noeuds crées et ainsi de suite jusqu'arriver à des feuilles avec une seule observation. Chaque noeud de l'arbre est donc associé à un sous ensemble d'observations qui a été obtenu par partition d'un ensemble parent par rapport à une variable explicative particuliére. Une fois l'arbre reconstruit, l'intensité d'une relation entre un régulateur z_j et la variable x est calculée en sommant les réductions de dispersions aux noeuds où z_j a servi pour faire la partition. Les arbres de regression ont beaucoup d'avantages mais souffrent d'instabilité et de sur

apprentissage. Pour contourner ces limites, GENIE3 applique plusieurs fois l'algorithme décrit plus haut à des sous-échantillons de la matrice d'expression initiale (bootstrap), et les scores finaux d'une paire régulateur - gène cible s'obtient alors par agrégation des scores obtenus pour chaque arbre créé (d'où le nom de forêt d'arbres de régression).

Toutes les méthodes décrites jusqu'ici infèrent à partir des données d'expression une matrice attribuant un score à chaque paire de gènes. Une étape importante de ces méthodes est de distinguer entre interaction directes et interactions indirectes. Un réseau est alors reconstruit en fixant un seuil puis en liant d'une arrête toutes les paires dont le score est supérieur au seuil.

Réseaux Bayesiens

Contrairement aux méthodes vues jusqu'à présent, dans le cadre méthodologique des réseaux Bayesiens, la structure du réseau causale est directement inférée à partir des données. Les réseaux Bayesiens sont des modèles aussi bien graphiques que probabilistes. Nous commençons donc cette section par présenter quelques notions de probabilité qui nous serons utiles pour expliciter la théorie des réseaux Bayesiens.

La théorie des probabilités est le language mathématique permettant de quantifier l'incertitude. Pour définir les concepts de base de la théorie des probabilités, nous nous servirons de l'exemple suivant : une urne opaque contenant trois boules noires numérotées de 1 à 3 (notées respectivement n_1, n_2, n_3), et une boule rouge qui porte le numéro 0, notée r_0. Soit l'expérience aléatoire qui consiste à tirer simultanément deux boules de l'urne.

L'univers Ω est l'ensemble des résultats possibles de l'expérience aléatoire, $\Omega = \{n_1n_2, n_1n_3, n_1r_0, n_2n_3, n_2r_0, n_3r_0\}$.

Une réalisation est un élément de l'univers et **un évènement** est un ensemble de réalisations. Par exemple l'évènement "la boule rouge fait partie des deux boules

tirées" est $R = \{n_1 r_0, n_2 r_0, n_3 r_0\}$.

La probabilité d'un évènement est le rapport entre le nombre de réalisations de l'évènement et le nombre d'évènement de l'univers. La probabilité de l'évènement R est $P(R) = \frac{|R|}{|\Omega|} = 3/6 = 1/2$. Comme pour tout évènement E, $E \subseteq \Omega$, $0 \leq |E| \leq |\Omega|$ et donc $0 \leq P(E) \leq 1$. Aussi, $P(\Omega) = 1$.

La probabilité jointe de deux évènements E_1 et E_2 est la probabilité de l'intersection de E_1 et E_2 :

$$P(E_1, E_2) = P(E_1 \bigcap E_2) = \frac{|\{\omega \in \Omega \ : \ \omega \in E_1 \ \ et \ \ \omega \in E_2\}|}{|\Omega|}$$

La probabilité conditionnelle de E_1 sachant E_2 est définie par :

$$P(E_1|E_2) = \frac{P(E_1, E_2)}{P(E_2)} \tag{3.9}$$

Deux évènement E_1, E_2 sont dits indépendants si et seulement si :

$$P(E_1, E_2) = P(E_1)P(E_2) \tag{3.10}$$

Si E_1 et E_2 sont indépendants, l'équation 3.9 devient alors :

$$P(E_1|E_2) = \frac{P(E_1, E_2)}{P(E_2)} = P(E1) \tag{3.11}$$

L'indépendance est une notion importante pour le formalisme des réseaux Bayésiens, et plus généralement en probabilité et en statistiques. On note $E_1 \perp E_2$ pour dire que les évènements E_1 et E_2 sont indépendants. Dans le cas contraire nous notons $E_1 \propto E_2$. Intuitivement, l'équation 3.11 peut s'interpréter comme : deux évènements E_1 et E_2 sont indépendants si savoir que E_2 s'est réalisé ne change pas la probabilité

que E_1 se réalise. Autrement dit, une observation de E_2 n'apporte pas d'informations sur E_1. En pratique, pour montrer que deux évènement sont indépendants, soit on le suppose *a priori*, soit on prouve l'équation 3.10. Revenons à l'exemple de l'urne cité plus haut, et l'évènement R : "la boule rouge fait partie des deux boules tirées". Considérons maintenant un nouvel évènement Imp : "la somme des numéros des boules est impaire". On vu que $R = \{n_1r_0, n_2r_0, n_3r_0\}$, et on peut facilement voir que $Imp = \{n_1n_2, n_2n_3, n_1r_0, n_3r_0\}$. On peut alors facilement montrer que ces deux évènements sont indépendants : $P(R, Imp) = P(R \bigcap Imp) = P(\{n_1r_0, n_3r_0\}) = 2/6 = P(R)P(Imp) = 3/6 \times 4/6$.

En pratique, on préfère définir quantitativement les évènements pour faciliter leur manipulation numérique. On se sert alors de variables aléatoires. **Une variable aléatoire** attribue une valeur numérique à chaque réalisation. Une variable aléatoire X est donc une fonction $X : \Omega \to \mathbb{R}$. Nous suivrons ici les notations classiques en théorie des probabilités, en symbolisant les variables aléatoires par des lettres en majuscule, attention à ne pas confondre X la variable aléatoire avec X matrice d'expression utilisée plus haut. On peut voir que les deux évènements R et Imp peuvent être définis par les variables aléatoires X : "somme des numéros des boules tirées" et Y : "nombre de boules rouges tirées". R devient alors $Y = 1$, et Imp devient X est impair. Toutes les équations vues plus haut s'étendent naturellement aux variables aléatoires.

Causalité

Bien qu'elle soit très intuitive, la notion de causalité n'a pas de définition mathématique. Cette idée fait débat depuis longtemps entre philosophes. Hume, philosophe anglais du XVIIIème siècle, l'un des principaux fondateurs de l'empirisme anglais, va jusqu'à nier son existence, arguant qu'elle n'est qu'une fausse idée qu'on se fait de la relation qui lie deux évènements lorsqu'on voit se répéter la conjonction de ces évènements.

Pearl, qui fonde les réseaux Bayésiens en 1988 [56], définie pourtant ces réseaux comme causaux. Le but est de lier deux variables par une arête, symbolisant une relation de causalité directe, lorsqu'on ne peut pas contester une relation de causalité entre ces deux variables. La causalité implique une dépendance, mais la réciproque n'est pas vraie. Ainsi, par contraposé, si deux variables sont indépendantes, il n'y a pas de relations causales entre elles. D'autres situations peuvent permettre de réfuter une relation de causalité entre deux variables. Pour les identifier, Pearl se sert de la notion d'indépendance conditionnelle.

On dit que que X et Y sont indépendantes conditionnellement à Z (noté $X \perp Y|Z$), si $P(X|Y,Z) = P(X|Z)$. Autrement dit, une fois Z connu, une observation sur Y n'apporte plus aucune information sur X. Pour illustrer cette notion centrale d'indépendance conditionnelle nous reprenons le cas simple de 3 variables et deux liens causaux. Les trois structures possibles sont représentées dans la Figure 3.13.

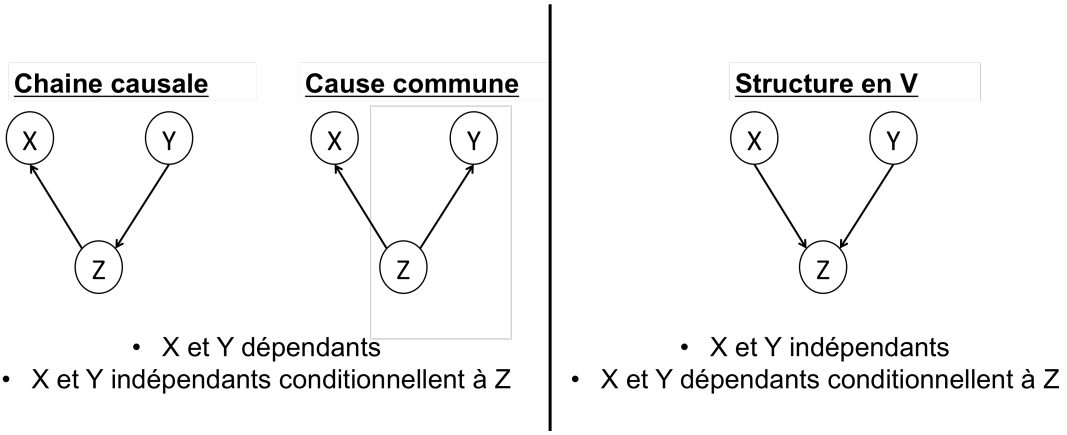


FIGURE 3.6 – Indépendance conditionnelles et réseaux Bayesiens

Si on suppose que la variable aléatoire X représente le prix des tomates au marché, Y le taux d'ensoleillement et Z la production de tomates. On obtient une relation de chaine causale entre ces trois variables $Y \to Z \to X$, le taux d'ensoleillent influence la production de tomates qui à son tour impacte leur prix. Si on sait que le taux

d'ensoleillement a été élevée, sans savoir quelle a été la production de tomates, on peut la supposer élevée, et par conséquent supposer que le prix des tomates sera relativement bas, X et Y sont dépendants. Si maintenant on apprend que la production de tomates à été très basse cette année, à cause d'une épidémie affectant les récoltes par exemples, on en déduira que le prix des tomates sera élevé, indépendamment du taux d'ensoleillement, $X \perp Y|Z$.

Supposons à présent que X (resp Y) est une variable booléenne indiquant si ma pelouse (resp. la pelouse de mon voisin) est mouillée, et que Z est une variable booléenne qui indique s'il a plu durant la nuit. On a bien une relation de cause commune $X \leftarrow Z \rightarrow Y$. Si en sortant le matin j'observe que ma pelouse est mouillée, je supposerai qu'il a plu durant la nuit et que par conséquent la pelouse de mon voisin est probablement également mouillée, X et Y sont dépendants. Supposons maintenant que j'ai appris avant de sortir qu'il a plu durant la nuit, ayant vu l'information à la télévision par exemple, je peux directement en supposer que la pelouse de mon voisin est mouillée. Apprendre qua ma pelouse l'est ne m'apportera aucune information supplémentaire, $X \perp Y|Z$.

Le dernier cas est opposé aux deux précédents. Supposons pour l'illustrer deux gènes X et Y desquels les expressions sont indépendantes. On se met dans le cas simple ou un gène est soit exprimé soit non exprimé ($X = 1$ si le gène correspondant est exprimé, $X = 0$ sinon). Supposons maintenant un troisième gène Z qui n'est exprimé que si $X = 1$ ou $Y = 1$. Le réseau de régulation est donc représenté par une structure en V, $i \rightarrow k \leftarrow j$. Comme X et Y sont indépendants, si je n'ai aucune information sur l'état de Z, et si j'observe $X = 1$, je ne peux rien inférer de l'état de Y. Si maintenant je sais que X est exprimé, observer $X = 0$ permet d'en déduire que $Y = 1$, X et Y sont donc dépendants conditionnellement à Z. Les structures en V seront exploitées pour détecter des hubs indépendants par la nouvelle méthode présentée au prochain chapitre.

Nous allons maintenant caractériser chacune des trois structures causales décrites au dessus du point de vue de la probabilité jointe des trois variables X, Y, et Z. En passant à trois variables, l'équation 3.9 devient $P(X,Y,Z) = P(X|Y,Z)P(Y,Z)$. On étend tout aussi naturellement l'équation 3.11 à trois variables, qui devient $P(X,Y,Z) = P(X|Y,Z)P(Y,Z)$. Il suffit à chaque fois de considérer que Y,Z composent une seule et même variable (ce qui ne pose aucun problème conceptuel), pour aboutir à ces résultats.

Les deux premières structures, chaine causale et cause commune, impliquent les même relations de dépendance *a priori* entre X et Y et d'indépendance conditionnelle $Y \perp X|Z$. On peut donc écrire la probabilité jointe des trois variables comme suit :

$$P(X,Y,Z) = P(X|Y,Z)P(Y,Z) = P(X|Z)P(Z|Y)P(Y) \tag{3.12}$$

La structure en V implique une dépendance *a priori* entre X et Y et une dépendance conditionnelle $X \perp Y|Z$. On peut donc écrire la probabilité jointe des trois variables comme suit :

$$P(X,Y,Z) = P(Z|X,Y)P(X,Y) = P(X|Z)P(X)P(Y) \tag{3.13}$$

On peut voir que chacune de ces structures à trois noeuds se traduit par une probabilité jointe des variables décomposée en produit des probabilités de chaque variable conditionnellement à son voisinage entrant. Nous pouvons à présent définir un réseau Bayésien en généralisant les relations vues pour les structures simples à trois noeuds. Un réseau Bayesien est un graph $G = (\{X_1, ..., X_n\}, A)$ où les noeuds représentent les variables aléatoires X_i et les arêtes les relations causales directes entre ces variables. La probabilité jointe des variables du réseau s'écrit alors :

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|V^-(X_i)) \tag{3.14}$$

Où $V^-(X_i)$ représente le voisinage entrant de X_i, c'est à dire ses causes directes (parfois appelés les parents de X_i dans le réseau). Il est important de noter que le réseau doit être sans cycles pour que la décomposition 3.14 soit possible. Dans le cadre des réseaux Bayesiens, le but est donc de reconstruire un réseau qui soit conforme aux indépendances conditionnelles estimées à partir des données. Comme nous l'avons déjà mentionné, une relation de causalité implique une dépendance, mais la réciproque n'est par vraie. Ainsi, plusieurs réseaux différents peuvent être conformes à une même probabilité jointe. Ces réseaux forment une classe d'équivalence appelée classe markovienne.

Inférence des structures des réseaux Bayesiens

Il existe deux familles de méthodes pour l'inférence de la structure d'un réseau Bayesien. La première commence par calculer des scores d'indépendances entre les variables pour reconstruire un réseau conforme. La deuxième parcours l'espace des réseaux possibles et attribue un score de vraisemblance à chaque réseau considéré. Nous brièvement décrire les principes de ces différentes familles d'approches.

Les méthodes basées sur les tests d'indépendance commence par un réseau complet. Des tests statistiques d'indépendance sont alors appliqués à chaque paire de variables et les arêtes entre variables déclarés indépendantes sont retirées. Le graph est alors itérativement modifié en fonction des résultats de tests d'indépendance entre chaque paire de variables conditionnellement à différents ensemble de variables tierces. Explicitons cette modification en considérons le cas d'un paire de variables X_i et X_j dont on teste l'indépendance conditionnelle par rapport à une seule variable tierce X_k. Deux cas peuvent alors se présenter :

-Si X_i et X_j n'étaient pas reliées par une arête, et qu'ils sont déclarés dépendant conditionnellement à X_k, on se trouve dans le cas d'une structure en V, on oriente

les arêtes de X_i vers X_k et de X_j vers X_k.

-Si X_i et X_j étaient reliées par une arête, et qu'ils sont déclarés indépendants conditionnellement X_k, ce qui correspond à une chaine causale passant par X_k, ou à une structure avec X_k comme cause commune à X_i et X_j. On supprime alors l'arête entre X_i et X_j et les arêtes entre X_i et X_k d'une part et X_j et X_kd'autre part reste sans orientation.

Au fur et à mesure des tests exécutés, les arêtes retirés et orientés doivent respecter les contraintes imposées par les tests précédents. On ne peut pas par exemple créer une structure en V là où une chaine causale ou une structure commune avaient été déclarées. On voit là la première limite de ce genre d'approches, l'ordre d'exécution des tests a un effet potentiellement important sur les réseaux reconstruits. A la fin, on obtient un réseau bayesien qui représente la classe markovienne des réseaux supportant la loi de probabilité jointe des variables composée à partir des indépendances conditionnelles. Certaines arêtes peuvent alors être orientées par propagation de l'orientation des arêtes déjà orientées, sans bien sûr créer des structures en V là où elle n'existaient pas. L'exemple le plus connu de cette famille est l'algorithme PC [57].

La deuxième famille de méthodes est opposée à la première dans son principe de base. Les méthodes de cette deuxième catégorie parcours l'espace des réseaux possibles et évalue chaque réseau par rapport aux données. Un réseau particulier considéré implique une probabilité jointe des des données. Un score de vraisemblance est alors attribué à ce réseau en calculant la probabilité *a posteriori* d'obtenir les données si la probabilité jointe des variables était celle induite par le réseau (cette probabilité *a posteriori* est appelée la vraisemblance du modèle). D'autre scores (dits de parcimonie) combinent score de vraisemblance et pénalité sur la complexité du réseau, pour privilégier les réseaux les moins complexes pour un même score de vraisemblance (les scores de parcimonie les plus connus sont le score AIC [58] et le score BIC [59]).

Depuis leur invention, les réseaux Bayesiens ont suscité beaucoup d'intérêt chez les scientifiques de différents domaines qui cherchent à modéliser des réseaux causaux sous-jacents à des systèmes de variables aléatoires. En revanche, en ce qui concerne la reconstruction de RRGs à partir de données d'expression, ces méthodes se montrent peu efficaces [34], principalement en raison du faible nombre d'observations comparativement aux variables, ce qui affaibli nettement la puissance des tests d'indépendance et des calculs de vraisemblance. On peut citer par exemple [60, 61, 62], des méthodes d'inférence de RRGs basées sur la méthodologie des réseaux Bayesiens développées au début des années 2000. Très peu de méthodes de ce genre ont été proposées plus récemment.

3.3.3 Les bases de données DREAM

Les projets DREAM (pour "Dialogue on Reverse-Engineering Assessment and Methods") [64] ont été développés pour permettre une comparaison des différentes méthodes d'inférence de réseaux biologiques. Ces projets s'organisent en différents défis adressant différents problèmes biologiques. Pour chaque défi, des données biologiques sont proposées à la communauté, puis des mesures standard de performance permettent d'évaluer comparativement la capacité des méthodes à retrouver des réseaux de référence à partir de ces données. Nous nous intéressons ici à DREAM4 [55, 38, 65] et DREAM5 [34], deux éditions qui ont trait à la reconstruction de RRGs à partir de données d'expressions.

DREAM4 met au défi les méthodes d'inférence des RRGs sur des bases de données *in silico*. Pour produire ces bases, les données d'expression des gènes ont été obtenues par simulation numérique à partir d'un certain nombre de sous-réseaux extraits du RRG de référence de *S. cerevisiae*. L'extraction des sous-réseaux et les procédures de simulation ont été réalisées avec GeneNetWeaver dont on peut trouver une description détaillée dans [66]. Brièvement, la procédure d'extraction recherche, pour un nombre de noeuds fixé, des sous-réseaux du RRG de référence de manière

à ce que les caractéristiques topologiques du sous-réseau soient similaires à celle du réseau entier (même densité d'arêtes, même coefficient de clustering, etc...). La simulation associe à chaque sous-réseau extrait un modèle dynamique défini par un système non linéaire d'équations différentielles, augmentées par des termes de bruit suivant des lois gaussiennes. Autrement dit, le taux de transcription de chaque gène est exprimé en fonction du niveau d'expression de ses régulateurs, plus un terme de bruit pour plus de réalisme. Les données simulées à partir de chaque sous-réseau sont de trois types : séries temporelles, KO/KD et multifactorielles. Nous nous intéresserons ici aux données multifactorielles qui correspondent aux données d'expression des gènes à l'état d'équilibre, ce cas étant en pratique le plus compliqué et le plus courant. Nous nous concentrerons donc sur cinq réseaux de la base de donnée DREAM4, chacun comprenant cent gènes et associé à une matrice de données d'expression multifactorielles de 100 observations indépendantes (matrices de dimensions 100×100).

DREAM5 met les méthodes évaluées face à une situation plus réaliste puisque celle-ci se voient proposer de reconstruire les RRGs complet d'*E. coli* et *S. cerevisiae* à partir de données *in vitro*, obtenues par des expériences de mesure d'expression des gènes par puces à ADN. Une matrice de données est ainsi associée à chaque RRG de référence. Chaque matrice a été produite en assemblant des données d'expression générées par différentes expériences transcriptomiques sur puces mais réalisées sur la même plateforme Affymetrix (un des leaders mondiaux dans la production des puces à ADN). Plus de détails concernant la récupération, l'assemblage et la normalisation des données d'expression peuvent être trouvés dans [34]. Chaque matrice d'expression est composée de données de nature différente : temporelle, KO/KD et multifactorielle. Les informations sur chaque observation sont fournies dans un tableau de métadonnées. Pour nous mettre dans la situation qui nous intéresse dans cette thèse, toutes les observations seront considérées comme multifactorielles, en ignorant simplement les informations de métadonnées. La méthode de reconstruc-

tion des RRGs de référence a également été décrite dans [34]. Ces réseaux ont été obtenus en combinant données Chip-chip, données de conservation pour l'identification des séquences de fixation des TFs, et données d'expression KO/KD. Seules les arêtes présentant le plus fort support expérimental on été maintenues. Ainsi, les arêtes qui composent ces RRGs de référence représentent très vraisemblablement de véritables interactions entre des FTs et leurs gènes cibles. Par contre, il est probable que certaines vraies régulations soient absentes des RRGs de référence. Pour ce travail de thèse, nous avons écarté les noeuds de degré 0 (sans arêtes) des RRGs de référence de la base DREAM5. Nous obtenons ainsi un réseau avec 1081 noeuds et 2055 arêtes pour *E. coli* et un réseau de 1994 noeuds et 3935 arêtes pour *S. cerevisiae*. A chacun de ces réseaux est associée une matrice d'expression avec 805 et 536 observations pour respectivement *E. coli* et *S. cerevisiae*.

Chapitre 4

HubNeD

Nous décrivons dans ce chapitre HubNeD ("Hub-centered Network Deconvolution"), une nouvelle méthode d'inférence de RRGs à partir de données d'expression à l'état d'équilibre. Comme les méthodes présentées dans le chapitre 3.3.2, HubNeD considère que les RRGs sont peu denses et que leurs coefficients de clustering sont bas. Le réseau est donc reconstruit après une étape de déconvolution qui cherche à distinguer entre signal direct et signal indirect. En plus des hypothèses classiques, HubNeD en introduit une supplémentaire : la topologie des RRGs est centrée sur des hubs très dominants. HubNeD est en effet la seule approche qui commence par détecter les hubs du système directement à partir des données d'expression. Pour nous, identifier les hubs c'est reconstruire le réseau. Cette démarche va à contre-sens des approches classiques qui reconstruisent le réseau pour identifier les hubs. Afin d'identifier efficacement les hubs à partir de données d'expression bruitées et de grande dimension, HubNeD réduit l'espace de recherche en écartant des gènes considérés comme exclusivement régulés (qui ne sont pas des FTs). Pour ce faire, nous capturons les corrélations les plus élevées de chaque gène pour construire un graphe où les communautés densément interconnectées forment des clusters homogènes de co-régulation (CHC). Les profils d'expression des gènes présents dans le même CHC présentent des similarités mutuelles élevées, indiquant que ces gènes sont suscep-

tibles d'être des cibles d'un programme de régulation spécifique contrôlé par un ou plusieurs régulateurs cachés (cachés dans le sens où les régulateurs sont, dans la plupart des cas, absents du CHC).

Cette stratégie d'inférence des hubs en amont de la reconstruction du réseau, permet à HubNeD de réduire considérablement le taux d'erreurs lors de la reconstruction, conduisant à des performances significativement supérieures à celles des autres méthodes considérées, WGCNA[24, 25], CLR [63], MrNet[53], GENIE3[55].

4.1 Données d'expression et topologie des RRGs

L'analyse des RRGs de référence décrite en première partie, nous a permis de mettre en évidence des caractéristiques importantes que nous intégrons à la méthode comme hypothèses :

1. Les RRGs sont organisés autour d'un petit nombre de hubs très dominants en terme de connectivité.

2. Les RRGs sont peu denses en arêtes et contiennent une faible proportion de triangles.

Une dernière hypothèse a été établie en analysant les données d'expression associées à ces réseaux. Nous allons pour cela comparer les corrélations calculées à partir des données d'expression d'*E. coli* récupérées de la base DREAM5 et les corrélations calculées à partir de données simulées. Dans chaque cas, nous étudions la relation entre les plus fortes corrélations et la topologie du RRG de référence d'*E. coli*. Les données simulées ont été obtenues en appliquant GeneNetWeaver au RRG de référence d'*E. coli*, imitant ainsi la procédure de simulation *in silico* utilisée dans DREAM4.

Comme cela se fait classiquement, nous utilisons la corrélation de Pearson (3.6) pour mesurer les similarités entre les paires de profils d'expression des gènes. Concrètement, nous comparons corrélations directes et indirectes. Les corrélations directes

correspondent aux régulations directes, c'est à dire aux arêtes orientées du réseau de référence. Les corrélations indirectes correspondent à des co-régulations. Elles émergent entre deux gènes cibles qui ne sont pas directement connectées dans le réseau réel mais partagent un ou plusieurs régulateurs communs. Les profils d'expression de deux gènes cibles co-régulés affichent en effet des similitudes, comme on peut plus généralement l'observer entre deux effets d'une cause commune. L'analyse des corrélations calculées à partir de données simulées, comparativement à la topologie du réseau, révèle que les corrélations directes (boîte à moustaches orange à gauche de la Figure 4.1, gauche) sont supérieures aux corrélations indirectes des paires de gènes co-régulés par un à huit régulateurs communs (boîtes à moustaches bleues à gauche de la Figure 4.1, gauche). Contrairement au comportement observé sur les données simulées, dans le contexte des données réelles d'*E. coli*, les corrélations directes ont des valeurs plus faibles que les corrélations indirectes (boîtes à moustaches orange et bleues de la Figure 4.1, centre). Ici, les corrélations les plus élevées correspondent à des paires de gènes co-régulées par cinq régulateurs communs ou plus.

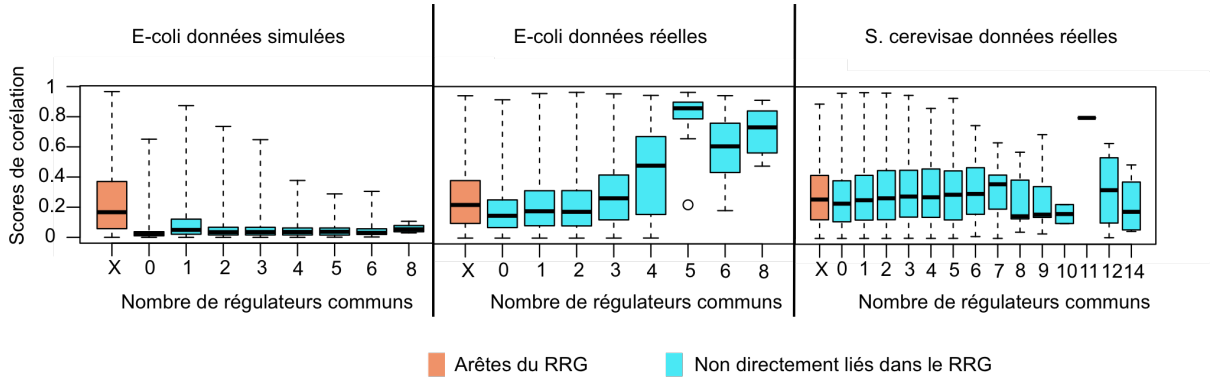


FIGURE 4.1 – Correspondance entre fortes corrélations et topologies du réseau. Dans le cas des données simulées, les plus fortes corrélations correspondent aux relations directes dans le réseau. Dans le cas des données simulées, elles correspondent aux relations de co-régulation indirectes.

Cette différence entre les données simulées et les données réelles montre que la procédure de simulation ne rend pas compte de toute la complexité du problème de reconstruction de RRGs à partir de données transcriptomiques réelles. Comme l'expression simulée d'un gène est calculée en fonction uniquement des expressions de ses régulateurs, nous pouvons considérer que les données d'expression simulées donnent une image complète du système. Ceci n'est cependant pas vrai dans le contexte de données réelles, où l'expression d'un gène n'est pas seulement une fonction des mesures d'expression de ses régulateurs, mais peut également impliquer d'autres paramètres des régulateurs, tels que leur conformation, leur méthylation ou leur association avec certains signaux chimiques pouvant déclencher ou inhiber leur activité de régulation. Par conséquent, les données d'expression réelles donnent une image incomplète du réseau ; *in vivo*, les mesures d'expression ne sont qu'une des dimensions des relations multidimensionnelles entre régulateurs et cibles. On comprend donc pourquoi les méthodes de reconstruction des RRGs, qui généralement partent du principe que les corrélations les plus élevés correspondent à des relations directes dans le réseau, génèrent un niveau élevé d'erreurs de prédiction, expliquant la chute de leurs performances lorsqu'elles sont appliquées à des données réelles par rapport à des données simulées. Par exemple, Aracne [45], une des premières méthodes de reconstruction de RRGs, fait la distinction entre relations directes et indirectes en considérant les gènes trois à trois, et en interdisant de lier la paire dont le score de corrélation est le plus faible. Au regard de ce que nous avons montré plus haut, cette approche est adaptée aux données simulées, mais face à des données réelles, génère de nombreuses erreurs de prédiction. Pour éviter ces erreurs, lorsqu'on traite des données réelles, ce qui est évidemment le but de la méthode développée, **nous considérons les corrélations les plus élevées comme des signaux de co-régulation uniquement**.

Dans le contexte des données réelles de *S. cerevisiae*, nous n'observons aucune correspondance entre topologie du RRG et corrélations significatives, qui ne corres-

pondent dans ce cas ni à des régulations directes ni à des co-régulations (boîtes à moustaches à droite de la Figure 4.1). Ceci peut soit être le résultat d'une sélection trop stricte lors de la construction du RRG de *S. cerevisiae*, qui manquerait alors d'une grande partie des régulations réelles, soit par le fait que les régulations post-transcriptionnelles et post-traductionnelles ont un rôle si important chez les eucaryotes qu'elles annulent complètement la relation entre corrélations des données d'expression et topologie du RRG.

4.2 La méthode HubNeD

La Figure 4.2 montre pas à pas le déroulement des quatre étapes de la méthode HubNeD dans le cas simple du RRG de référence présenté en haut à gauche de la figure. Ces étapes seront détaillées dans les prochaines sections de ce chapitre. En entrée HubNeD reçoit une matrice d'expression associée au réseau de référence, le but étant de reconstruire un réseau le plus proche possible du réseau de référence. Les quatre étapes sont :

1. Calcul des scores de corrélation de Pearson pour chaque paire de gènes.
2. Construction du graphe de co-régulation à partir des scores de corrélation les plus élevés.
3. Inférer les hubs du réseau à partir des scores de corrélation et du graphe de co-régulation.
4. Déconvolution centrée sur les hubs.

Les flèches bleues indiquent les trois résultats de HubNeD :

— **R1 :** Des communautés fortement connectées sont extraites du graphe de co-régulation. Elles constituent des groupes de gènes partageant le même programme de régulation et susceptibles d'être impliqués dans les mêmes processus fonctionnels.

— **R2 :** Un classement des gènes par ordre décroissant d'un score qui est plus élevé pour des gènes que nous inférons comme hubs.

— **R3 :** Une matrice d'adjacence calculée par déconvolution des scores de corrélation centrée sur les hubs. Un réseau est construit en connectant par une arête chaque paire de gènes dont le score d'adjacence est supérieur au seuil fixé. Un réseau reconstruit est affiché en bas à gauche de la Figure 4.2. Il est ensuite comparé au véritable réseau (réseau au milieu) pour évaluer la distance entre réseau reconstruit et réseau réel.

Dans ce qui suit, nous noterons X la matrice d'expression de dimension $m \times n$, où m est le nombre d'observations et n le nombre de gènes. La matrice de corrélation C se calcule donc en appliquant l'équation 3.6 à chaque paire de profils d'expression (*i.e.* aux paires de colonnes de X) pour ainsi obtenir une matrice symétrique de dimension n gènes $\times$ n gènes. Nous noterons ainsi dans ce qui suit $C_{i,j}$ le score de corrélation des profils d'expression des gènes i et j.

4.3 Clustering en groupe de co-regulation

Pour minimiser les erreurs de construction, nous nous limitons à l'exploration des signaux de corrélation les plus forts. Nous cherchons ici à représenter ces signaux de confiance élevée en un graphe. Nous montrons comment s'effectue la construction de ce graphe que nous appelons graphe de co-régulation à partir de la matrice de corrélation. Cette étape est illustrée dans la Figure 4.3 où à gauche on a la matrice de corrélation calculée à partir de la matrice d'expression. Les cases vertes sur la matrice de corrélation indiquent pour chaque ligne (gène) les points extrêmes de son profil de corrélation, c'est-à-dire les gènes (colonnes) les plus fortement corrélés avec le gène correspondant à la ligne. En pratique, pour un gène i, on note MP_i l'ensemble des "Meilleurs Partenaires" du gène i, c'est à dire l'ensemble des gènes le plus fortement corrélés avec i :

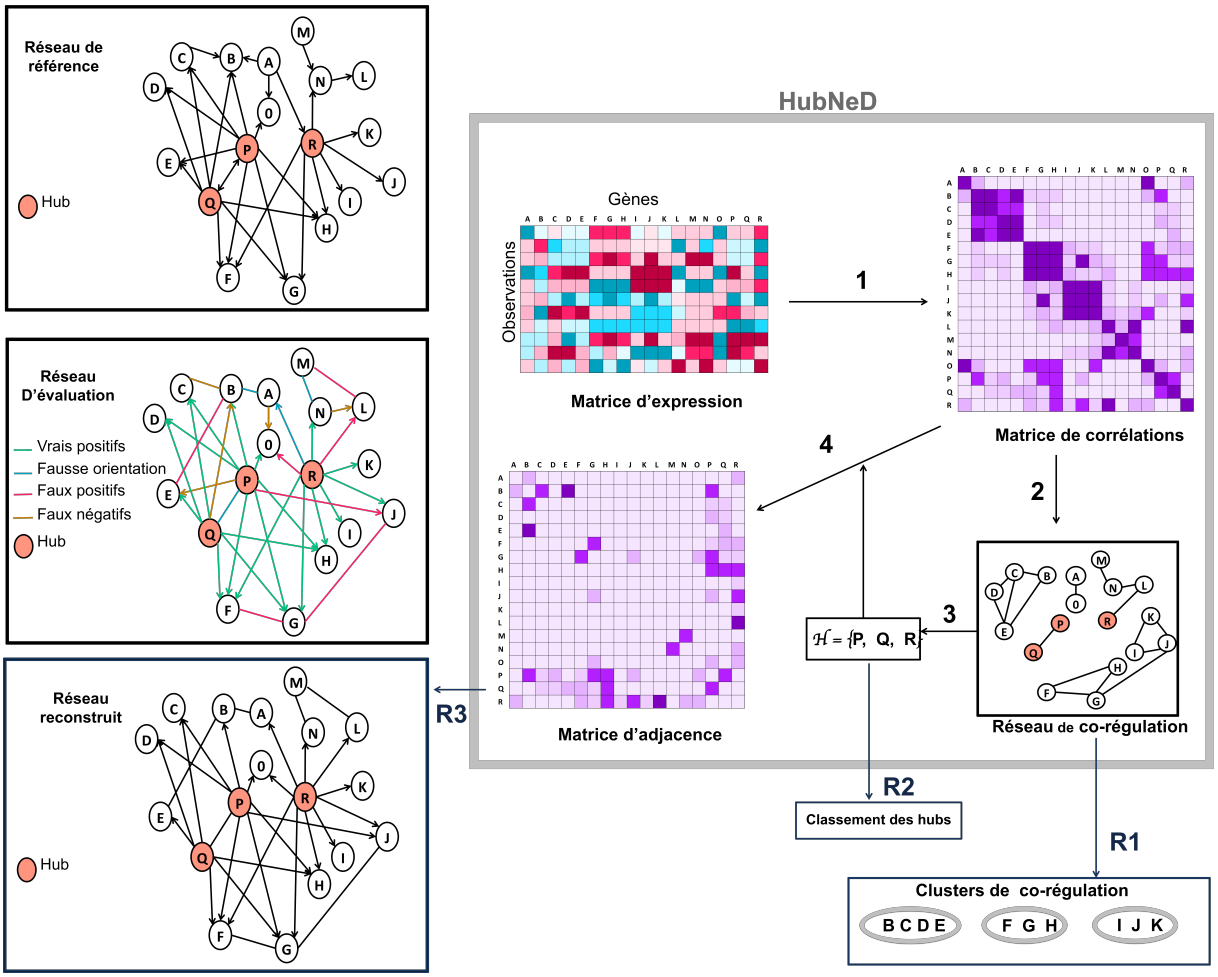


FIGURE 4.2 – Étapes du déroulement de la reconstruction d'un RRG par HUbNeD sur un exemple simple. Les 4 étapes sont représentés par les flèches numérotées de 1 à 4. 1 : Calcul des scores de corrélation de Pearson

$$MP_i = \{j \neq i \mid C_{i,j} > \bar{C_i}, + \alpha < C_i, >\} \tag{4.1}$$

où :

$$\bar{C_i}, = \frac{\sum_{k \neq i} C_{i,j}}{n} \quad et \quad < C_i, >= \sqrt{\frac{\sum_{k \neq i} (C_{i,j} - \bar{C_i},)^2}{n-1}} \tag{4.2}$$

Le nombre de cases vertes peut varier d'un gène à l'autre en fonction de la distribution des scores de corrélation de chaque gène aux autres gènes. Notons aussi qu'on peut avoir une case verte en case (i, j) mais pas en case (j, i), *i.e* on peut avoir $j \in MP_i$ et $i \notin MP_j$. Le graphe de co-régulation est ensuite construit en 2 étapes :

1. On construit un réseau orienté reliant i à j s'il y a une case verte à la ligne i et colonne j, *i.e.* si j fait partie des gènes les plus fortement corrélés à i ($j \in MP_i$).
2. Le graphe de co-expression final est construit en connectant uniquement les gènes qui sont mutuellement connectés dans le réseau précédent, *i.e.* i et j sont connectés si j fait partie des gènes les plus fortement corrélés à i, et i fait partie des gènes les plus fortement corrélés à j ($j \in MP_i$ et $i \in MP_j$). Ce faisant, nous établissons un lien entre deux gènes s'il existe un signal fort relatif et mutuel. On considère alors les arêtes de ce réseau comme révélant les relations les plus fortes du système. A partir de ce réseau, des communautés fortement connectées sont extraites en tant que clusters homogènes de co-régulation (CHC). Dans la Figure 4.3, les gènes de ces communautés sont en vert. L'identification des communautés fortement connectées à partir du graphe de co-régulation se fait en deux étapes. Toues les cliques sont premièrement extraites du graphes. Deux cliques sont en suite fusionnées si elles contiennent au moins un gène en commun.

4.4 Calcul de la matrice d'adjacence.

4.4.1 Inférence des hubs

Les observations faites dans la section 4.1 nous ont amené à développer deux stratégies différentes pour manipuler les CHCs, selon qu'on fait face à des données réelles ou simulées. Bien entendu, HubNeD a été développé pour être appliqué à des données réelles. Nous précisons néanmoins la stratégie d'inférence des hubs spécifique aux données simulées pour, en comparant HubNeD aux autres méthodes dans ce

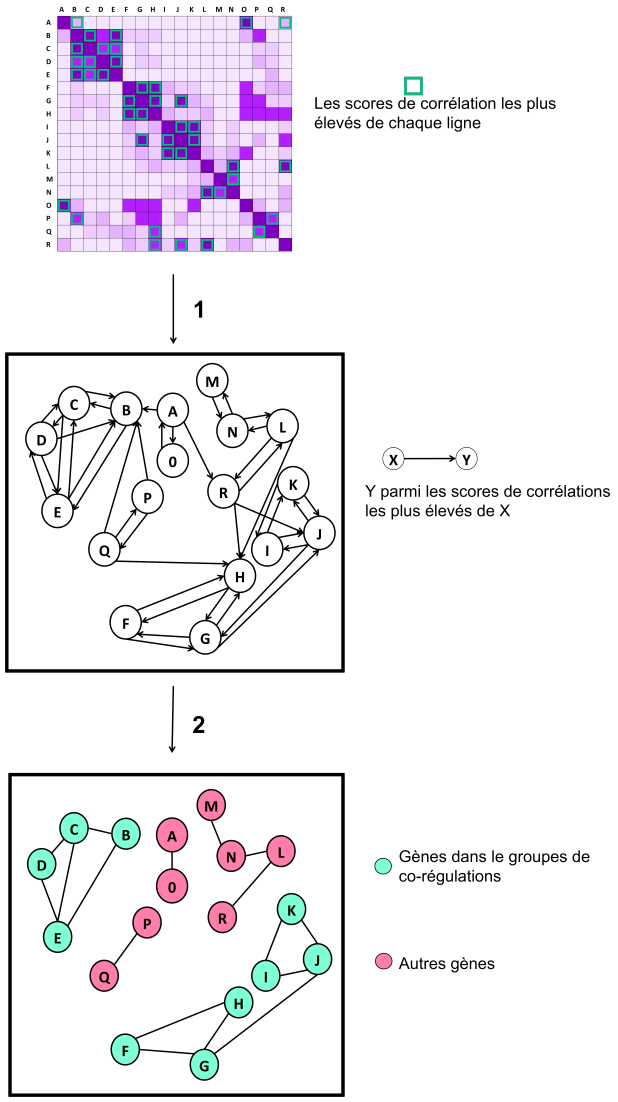


FIGURE 4.3 – Graphe de co-régulation à partir de la matrice de corrélation. L'étape 1 joint chaque gène aux gènes les plus fortement corrélés avec lui. L'étape 2 lie les gènes qui étaient mutuellement connectés à l'étape1.

contexte, montrer que l'idée de base d'HubNeD, trouver les hubs c'est trouver le réseau, lui permet de rivaliser avec les autres méthodes quand celles-ci affichent de bons niveaux de performances. Par ailleurs, quand on les teste sur les données réelles, les niveaux de performances de ces méthodes baissent drastiquement. Comme men-

tionné plus haut, nous supposons que cette nette baisse s'explique par le fait que les méthodes reconstruisent les réseaux à partir des signaux les plus forts. Si cette stratégie permet de capturer les relations directes de régulation dans le contexte des données simulées, dans celui des données réelles, elle résultera en un taux très élevé d'erreurs de prédiction, les signaux forts étant dans ce cas révélateurs de relations indirectes. Dans le contexte des données réelles, encore une fois celui qui nous intéresse, **HubNeD considère que les plus forts signaux de corrélation sont synonymes de co-régulation uniquement**. Il faut alors chercher les régulateurs parmi les gènes non capturés dans les CHCs (en rose dans la Figure 4.3), dont les corrélations sont cachées parmi les signaux de plus faible intensité.

Inférence des hubs à partir des données réelles

Nous utiliserons pour la suite de ce paragraphe CHC pour noter l'ensemble des clusters de co-régulation construits comme nous l'avons décrit dans la section précédente. CHC est donc un ensemble d'ensembles de gènes. Nous noterons $\overline{CHC}$ les gènes absents des CHCs. On peut voir cette décomposition comme une partition de l'ensemble des gènes en CHC et $\overline{CHC}$. Comme nous supposons que la co-expression (forte corrélation des profils d'expression) est un signal de co-régulation mais pas de régulation, les gènes contenus dans un ensemble de CHC (gènes en vert dans la Figure 4.3) sont considérés comme exclusivement régulés. Leurs régulateurs, donc également les hubs, sont à chercher dans $\overline{CHC}$. Cette partition permet donc de réduire l'espace de recherche des hubs à l'ensemble $\overline{CHC}$.

La méthode que nous proposons pour l'identification des hubs consiste à tirer profit des structures en V (ou bi-régulations) qui émergent entre deux hubs indépendants dans le réseau réel et leur nombreuses cibles communes. Par deux hubs indépendants dans le réseau réel, nous entendons qu'il n'y a aucune structure topologique impliquant les deux hubs qui puisse laisser penser une similarité des profils d'expression des ces hubs. C'est à dire qu'il n'y a pas de chemin liant un hub à l'autre, et, s'il y a un chemin qui lie un gène tiers à un des deux hub, il n'y a pas de chemin qui lie

ce troisième gène à l'autre hub. Nous allons donc rechercher les paires de gènes dans $\overline{CHC}$ qui sont faiblement corrélés entre eux mais dont les profils de corrélation aux CHCs présentent une forte similarité.

De manière cohérente avec les notations utilisées jusque là, nous utiliserons des lettres latines en minuscule pour symboliser les gènes. Nous ajoutons ici une nouvelle notation en lettres grecques en majuscule pour symboliser les clusters de CHC. Nous pouvons par exemple définir les CHCs de la Figure 4.3 par $CHC = \{\Gamma = \{b, c, d, e\}, \quad \Psi = \{f, g, h\}, \quad \Theta = \{i, j, k\}\}$.

On définit alors la corrélation entre un gène de $\overline{CHC}$ et un cluster de CHC comme suit : $\forall i \in \overline{CHC}, \quad \forall \Lambda \in CHC, \quad C_{i,\Lambda} = max_{k \in \Lambda} C_{i,k}$ où $C_{i,k}$ est le score de corrélation de Pearson de la paire (i, k).

Le score, qu'on note coH, de co-occurence de hub indépendants peut alors être défini comme suit : $\forall (i, j) \in \overline{CHC}^2, \quad coH_{i,j} = s(C_{i,j}) max\{\prod_{\Lambda \in CHC} C_{i,\Lambda} C_{j,\Lambda}\}$.

La fonction s permet de ne considérer que les paires de gènes dont les corrélations sont inférieures à un certain seuil, considérés alors comme indépendants. Cette fonction doit renvoyer des valeurs élevées (proches de 1) pour les corrélations proches de 0, et renvoyer 0 pour des corrélations dépassant un seuil fixé par en paramètre. On propose deux fonctions différentes, la sigmoïde permettant d'appliquer un seuillage souple $sigm(x) = \frac{\epsilon^\eta}{x^\eta + \epsilon^\eta}$, ou une fonction qui applique un seuillage drastique $\mathbb{1}(x) = \begin{cases} 1 \quad si \quad x < \epsilon \\ 0 \quad sinon \end{cases}$

Le paramètre ϵ est le seuil fixé en dessous duquel les corrélations impliquent indépendance. Le paramètre η permet de contrôler la pente de la fonction sigmoïde et ainsi de définir un seuillage plus ou moins souple. La Figure 4.4 montre les courbes de deux fonctions sigmoides et d'une fonction de seuillage drastique.

Nous pouvons à présent définir un score absolu H qui mesure la confiance qu'on attribue à un gène d'être parmi les hubs du réseaux à reconstruire.

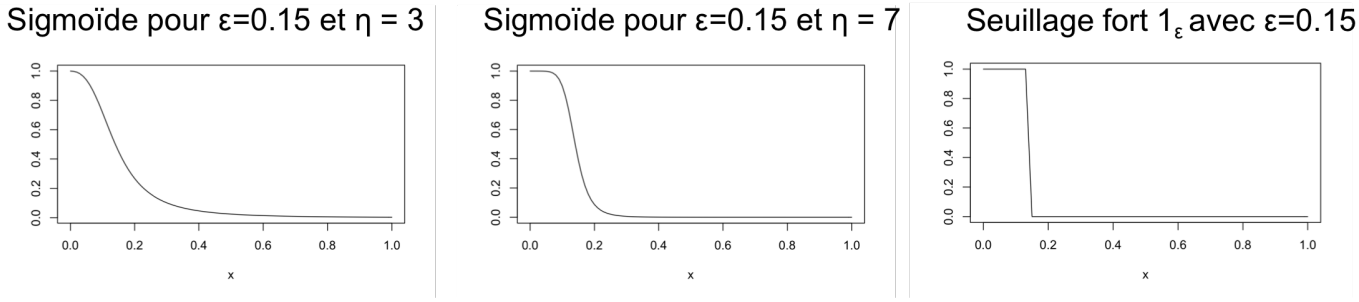


FIGURE 4.4 – courbes de fonctions de seuillage souple (les deux courbes à gauche) et une courbe pour le seuillage fort (à droite)

$$\forall i \in \overline{CHC}, \quad H_i = \frac{\sum_{k \neq i} coH_{i,k}}{\sum_{k \neq i} s_\theta(C_{i,k})} \tag{4.3}$$

Inférence des hubs à partir des données simulées

Nous avons vu dans la section 4.1 que dans le cas de données simulées, contrairement au cas des données réelles décrit au paragraphe précédent, les corrélations les plus fortes correspondent aussi bien à des relations directes qu'à des relations indirectes. Dans ce cas, les CHCs sont composés des hubs et certains de leurs cibles. Pour montrer comment ces structures communautaires apparaissent dans le cas des données simulées, considérons le premier des cinq réseaux de la base de données DREAM4 (Figure 4.5A). A partir des données simulées associées à ce réseau, nous avons construit un graphe de corrélation avec 100 arêtes correspondant aux 100 scores de corrélation les plus élevés (Figure 4.5B). Ce graphe contient des communautés densément interconnectées, chacune formée par un hub et certaines de ses cibles (une communauté est entourée par un cercle en pointillés dans la Figure 4.5B). On retrouve ici que les corrélations les plus fortes sont soit directes (en vert), soit indirectes (en rose). Comme mentionné plus haut, les corrélations directes correspondent à des régulations directes, représentées par des arêtes dirigées dans le réseau de référence. Ils connectent très probablement les hubs à leurs cibles simplement parce que la plupart des arêtes du réseau de référence sont incidentes aux

hubs. Nous voyons donc apparaître des structures de corrélation émergent autour d'un hub, avec des corrélations directes reliant le hub à certaines de ses cibles, et des corrélations indirectes reliant les cibles co-régulées par le hub, formant ainsi des communautés fortement intra-corrélées autour des hubs. Construit à partir des corrélations les plus fortes, le graphe de co-régulation va lui aussi voir se former en lui des structures similaires. Le graphe de co-régulation construit à partir des données d'expression associées à ce premier réseau de la base DREAM4 est affiché en Figure 4.5C. Nous retrouvons les CHCs qui incluent des hubs et certaines de leurs cibles, avec des liaisons hub à cibles correspondant à des régulations directes et des liens de cible à cible à des co-régulations indirectes. Nous observons dans ce graphe neuf communautés, dont six contiennent un hub. En outre, on peut observer que, comparativement aux communautés du graphe des cent plus fortes corrélations, les CHCs sont séparés les uns des autres, permettant une meilleure séparation des communautés contenant les hubs (Figure 4.5B,C).

Quand on fait face à des données simulées, l'inférence des hubs consistera donc à extraire un hub de chaque CHC. Nous avons déjà décrit les structures causales avec trois noeuds et deux arêtes (3.3.2), une première fois quand nous avons défini la corrélation partielle (équation 3.8) et une deuxième fois dans le cadre des réseaux Bayésiens. Nous allons nous servir de deux de ces structures pour identifier les hubs à partir des CHCs, la co-régulation et la bi-régulation (ou structure en V). D'une part, lorsqu'on le considère avec ses cibles, un hub est au centre des structures de co-régulations. D'autre part, deux hubs, s'ils ne sont pas directement connectés dans le réseau de référence, forment des structures de bi-régulation avec leurs cibles communes (Figure 4.6B).

Pour caractériser les structures de co-régulation et de bi-régulation, nous utilisons des scores de corrélation (équation 3.6) et de corrélation partielle (équation 3.8). Nous rappelons ici que la co-régulation $x \leftarrow z \rightarrow y$ est caractérisée par une forte

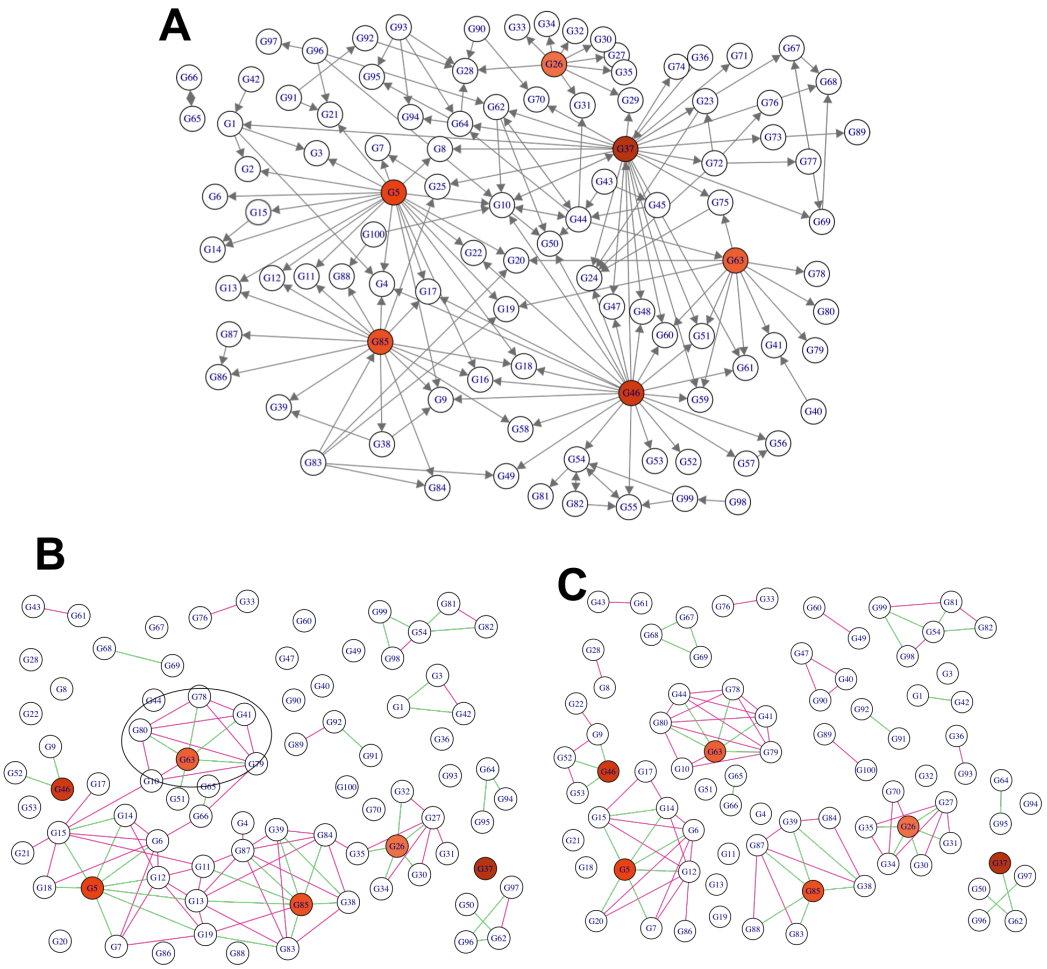


FIGURE 4.5 – **A.** Le premier réseau de la base de données DREAM4. Les hubs sont colorés en orange du plus foncé pour les hubs les plus connectés aux oranges plus claires pour les hubs d'ordre inférieur. **B.** Le réseau obtenu en connectant deux gènes si la valeur absolue de leur score de corrélation de Pearson se situe dans les 100 meilleurs scores. Les arêtes vertes sont des vrais positifs et les arêtes roses des faux positifs. **C.** Le graphe de co-regulation obtenu pour le même réseau, avec le même code couleur pour les arêtes. Les hubs sont plus clairement séparés dans le graphe de co-régulation que dans le réseau à sa gauche.

corrélation des deux cibles x et y et par une corrélation partielle faible (proche de 0) de x et y par rapport à z. En revanche, la bi-régulation $x \rightarrow z \leftarrow y$ est caractérisée par une faible corrélation (proche de 0) de x et y et par une corrélation partielle élevée de x et y par rapport à z.

Par conséquent, nous pouvons distinguer la co-régulation et la bi-régulation en définissant l'influence de z sur (x, y) comme la corrélation de x et y moins la corrélation partielle de x et y relativement à z :

$$I_{z:x,y} = C_{x,y} - C_{x,y|z} \tag{4.4}$$

Ainsi, l'influence est positive pour la co-régulation et négative pour la bi-régulation (voir Figure 4.6A). Nous définissons "l'influence endogène de z" comme la moyenne des influences pondérées du gène z sur l'ensemble des paires possibles formés par les autres gènes (x, y) :

$$I_z = \frac{\sum_{x \neq y \neq z} C_{z,x} C_{z,y} I_{z:xy}}{\sum_{x \neq y \neq z} C_{z,x} C_{z,y}} \tag{4.5}$$

Dans le même esprit, nous calculons "l'influence exogène" sur une paire de gènes (x, y) comme la moyenne des influences pondérées de tous les autres gènes z sur la paire :

$$I_{:xy} = \frac{\sum_{z \neq y,x} C_{z,x} C_{z,y} I_{z:xy}}{\sum_{z \neq x,y} C_{z,x} C_{z,y}} \tag{4.6}$$

En utilisant les scores d'influence endogènes et exogènes, HubNeD extrait un hub de chaque CHC en trois étapes (Figure 4.6C). La première étape est la construction du graphe de co-régulation identifiant les CHCs constituées d'un ensemble de noeuds fortement corrélés, où les corrélations peuvent être directes ou indirectes. Les corrélations directes représentent un échantillon des arêtes réelles du réseau, probablement les arêtes des hubs, celles-ci étant sur-représentées dans le réseau réel.

La deuxième étape présélectionne de chaque CHC les gènes ayant les influences endogènes les plus élevées. Notons que le nombre de CHC reste le même mais que le nombre de gènes dans un CHC est réduit. La dernière étape vise à prélever un hub de chaque CHC réduit, en maximisant un score de co-occurence de hub. Nous vérifierons de manière exhaustive toutes les combinaisons de hubs potentiels formées en choisissant un gène par CHC réduit. Nous évaluons de manière exhaustive chaque combinaison par la somme des influences exogènes sur toutes les paires et définissons l'ensemble final des hubs comme étant la combinaison avec le score minimal.

4.4.2 Déconvolution

HubNeD applique une déconvolution centrée sur les hubs de la matrice de corrélation C pour calculer une matrice d'adjacence A. Une fois que les hubs sont choisis, le réseau est reconstruit en donnant la priorité aux arêtes incidentes aux hubs et en pénalisant les arêtes entre les non-hubs. Plus précisément, pour chaque paire de gènes non-hubs x et y, nous retirons de leur score de corrélation de Pearson $C_{x,y}$, la racine carrée du produit de leurs scores de corrélation de Pearson respectifs $C_{h,x} \times C_{h,y}$ avec le hub h le plus proche de la paire (qui est le hub dont les scores de corrélation avec x et y ont le produit maximal $C_{h,x} \times C_{h,y}$). Intuitivement, h est le hub ayant l'effet le plus élevé sur la relation indirecte entre x et y. Formellement, si H est l'ensemble des hubs choisis et N l'ensemble des gènes (noeuds) du système :

$$\begin{aligned} \forall h_1, h_2 \in H, A_{h_1,h_2} = A_{h_2,h_1} = C_{H_1,H_2} \\ \forall h \in H, \forall x \in N \setminus H, A_{h,x} = C_{h,x}; A_{x,h} = 0 \\ \forall x, y \in N \setminus H, A_{x,x} = \min_{\forall h \in H} C_{x,y} - \sqrt{C_{x,h} \times C_{y,h}} \end{aligned} \tag{4.7}$$

Par conséquent, la réduction du score de corrélation de deux gènes non hubs est plus élevée lorsque leurs corrélations avec un même hub sont plus fortes. Cette étape de déconvolution produit une matrice d'adjacence servant à reconstruire le RRG (Figure 4.2). Sous l'hypothèse que les hubs dans les RRGs sont des FTs avec

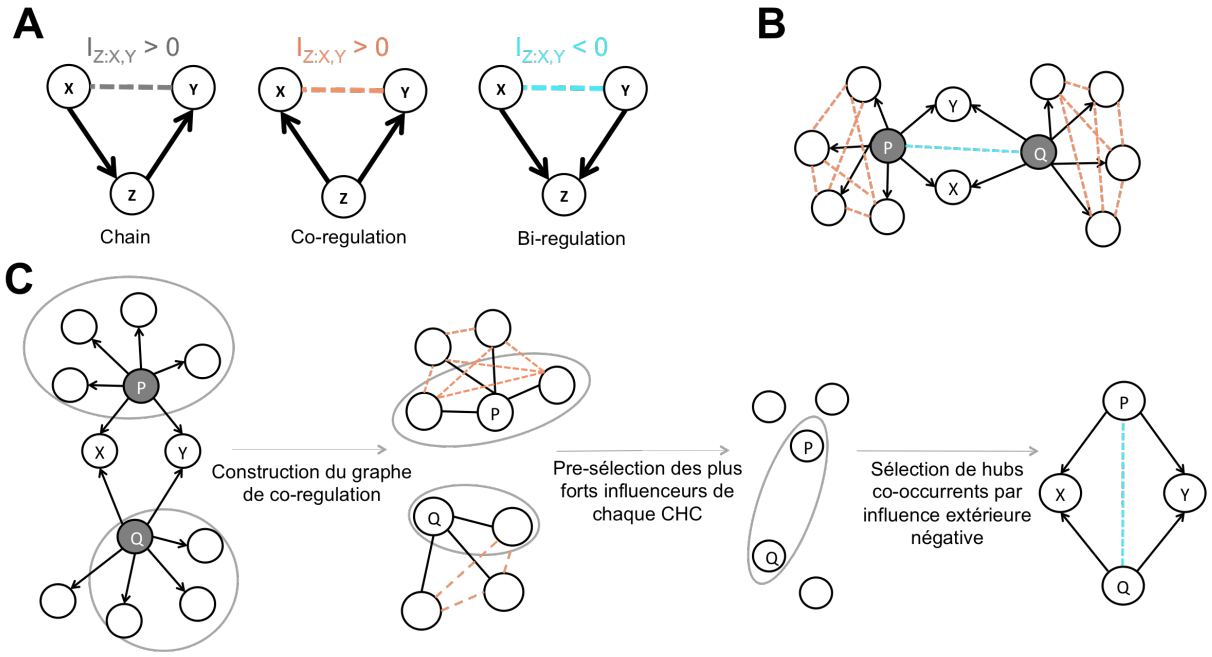


FIGURE 4.6 – **Usage des scores d'influence pour la détection des hubs dans le contexte des données simulées. A.** Les trois structures possibles avec trois noeuds x, y, z et deux arêtes $(x, z) et (y, z)$: chaîne (gauche), co-régulation (centre) et bi-régulation (droite). Pour chaque triplet, une ligne discontinue indique l'arête $I_{z:x,y}$, l'influence de z sur la relation entre x et y. Pour les chaînes et co-régulations, $I_{Z:X,Y}$ est significativement positif (lignes discontinues grise et orange). Pour la bi-régulation, $I_{Z:X,Y}$ est significativement négatif (ligne en pointillé bleu). **B.** Un réseau simple à 2 hubs déconnectés affiche des structures de co-régulation formées entre chaque hub et leurs cibles respectives (lignes pointillées orange) et des structures de bi-régulatriion formées entre les 2 hubs et leurs cibles communes x et y (ligne pointillée bleue) **C.** Procédure de sélection des hubs. Un simple réseau avec deux hubs (noeuds gris) est considéré (à gauche).

beaucoup de cibles, HubNeD oriente les arêtes incidentes aux hubs, des hubs vers leurs voisins, renvoyant ainsi des RRGs partiellement dirigés.

Chapitre 5

Evaluation des performances d'HubNeD

5.1 Données réelles

5.1.1 Clustering en groupe de co-regulation

Pour évaluer la qualité d'un CHC en tant que cluster de co-régulation, nous examinons la similarité des programmes de régulation des gènes qu'il contient. Le programme de régulation d'un gène étant défini par l'ensemble de ses régulateurs, la similarité des programmes de régulation de deux gènes est donnée par le chevauchement de leurs régulateurs respectifs. Nous définissons le score Hmg d'homogénéité d'un cluster comme étant le rapport entre le nombre de régulateurs communs et le nombre total de tous les régulateurs des gènes du cluster :

$$\forall \Gamma \in CHC, Hmg(\Gamma) = \frac{|\bigcap_{i\in\Gamma} |V^-(i)|}{|\bigcup_{i\in\Gamma} V^-(i)|} \tag{5.1}$$

Le meilleur cas, où tous les gènes d'un CHC ont exactement les mêmes régulateurs, correspond à un cluster de co-régulation parfaitement homogène dont le score d'homogénéité est égal à 1. Nous avons utilisé des scores d'homogénéité pour comparer

les CHCs à des clusters construits par la méthode de clustering hiérarchique partiel classiquement utilisée dans le protocole d'analyse de WGCNA [67]. Dans le contexte d'*E.coli*, nous montrons que les CHCs contiennent principalement des gènes ayant exactement le même ensemble de régulateurs, surpassant de manière significative les trois méthodes de clustering proposées par WGCNA (Figure 5.1, à gauche). Dans le contexte du RRG de *S. cerevisiae*, aucune méthode n'arrive à capturer des clusters homogènes, ce qui est cohérent avec l'absence de correspondance entre les corrélations élevées et la topologie du réseau observée dans le cas de *S. cerevisiae* (boîtes à moustaches à droite de la Figure 4.1).

5.1.2 Inférence des hubs

En caractérisant les topologies des RRGs, nous avons vu que ces réseaux étaient centrés sur un faible nombre de hubs très fortement connectés qui permettent de lier les autres noeuds moins connectés du réseau par de courts chemin (topologie du petit monde). Nous avons également vu que les RRGs, comme les réseaux sans échelle, affichent une tolérance remarquable face aux attaques aléatoires qui, affectant la grande majorité des noeuds faiblement connectés avec une plus grande probabilité, n'ont aucun impact sur la structure globale du système. Pour endommager le réseau, les attaques doivent cibler les hubs.

En biologie des systèmes, l'utilisation des méthodes de reconstruction de réseau à partir de données transcriptomiques ne vise pas seulement à décrire le système comme un réseau, mais surtout à identifier les hubs du système, en sélectionnant, en aval de la reconstruction du réseau, les gènes qui y sont plus connectés. WGCNA est l'approche la plus fréquemment utilisée dans les publications de biologie de systèmes (orientées réseaux). Comme mentionné dans la section précédente, WGCNA permet de trier les gènes du système en modules de co-expression. Les hubs sont ensuite choisis en recherchant les gènes présentant la plus forte connectivité dans chaque module et, si un trait biologique est disponible, qui affichent les corrélations les

plus élevées avec le trait. Toute autre méthode peut également être utilisée pour la sélection des hubs. En effet, ces méthodes produisent des matrices d'adjacence donnant un score de connectivité pour chaque paire de gènes. La connectivité d'un seul gène est alors simplement la somme de ses connectivités à tous les autres gènes.

Pour évaluer la capacité d'une méthode à trouver les hubs du système, nous supprimons du réseau réel un par un les noeuds par ordre décroissant de leurs connectivités prédites par la méthode, et examinons l'effet de leur suppression du réseau. Nous mesurons l'effet d'une suppression en mesurant la proportion de noeuds déconnectés (noeuds sans chemin les reliant dans le réseau) après la suppression. Nous considérons ici les composantes géantes des deux réseaux, ce qui signifie qu'avant toute suppression, la proportion initiale de noeuds déconnectés est zéro. En ce qui concerne *E.coli*, nous montrons qu'après la suppression des trois gènes de plus fortes connectivités selon HUbNeD, déjà 28% des noeuds sont déconnectés ; alors que pour les autres méthodes, il n'y a pas d'impact sur le réseau avant la 33ème suppression (Figure 5.2, à gauche). À la suppression des 18 noeuds les plus connectés selon HubNeD, la proportion de noeuds déconnectés atteint 48%, un niveau de fragmentation atteint les autres méthodes après la délétion de 142 noeuds pour CLR, 250 GENIE3 et 367 pour WGCNA. Cela montre que HubNeD est la méthode la plus efficace pour détecter les hubs du réseau. Ceci est encore plus évident quand on regarde l'impact des méthodes sur le RRG de *S. cerevisiae* où seul HubNeD est capable d'impacter le réseau, quand, après 100 suppressions, plus de 80% des noeuds sont déconnectés alors que les autres méthodes n'ont pratiquement aucun effet.

5.1.3 Reconstruction des réseaux

Nous évaluons les performances des différentes méthodes avec les scores classiques de précision et de sensibilité, mesurés en comparant les RRGs reconstruits avec le RRG de référence. La sensibilité est le rapport entre le nombre de vrais positifs (VP, arêtes présentes dans les deux réseaux) et le nombre d'arêtes dans le réseau

de référence, il indique la capacité de la méthode à capturer les arêtes du réseau de référence. La précision indique la robustesse du réseau reconstruit en mesurant le rapport entre le nombre de VP et le nombre d'arêtes du réseau reconstruit. Deux courbes permettent alors une évaluation comparative des performances. La courbe ROC ("Receiver Operating Characteristic") indique la sensibilité et la courbe PR ("Precision-Recall") indique la précision. Dans les deux cas, les aires sous les courbes (AUC PR et AUC ROC) délivrent des mesures de la performance de la méthode. Plus l'aire sous la courbe est élevée, plus la méthode est sensible (ROC) ou précise (PR). Dans les courbes ROC et PR de la Figure 5.3, nous pouvons voir que HubNeD est au dessus de toutes les autres courbes pour les deux systèmes d'*E.coli* et de *S. cerevisiae*. HubNeD affiche une amélioration remarquable à la fois de l'AUC PR et de l'AUC ROC sur les deux réseaux, démontrant en particulier une plus grande précision en raison de la stratégie efficace de HubNeD consistant à éviter les faux positifs dus à la convolution centrée sur les hubs.

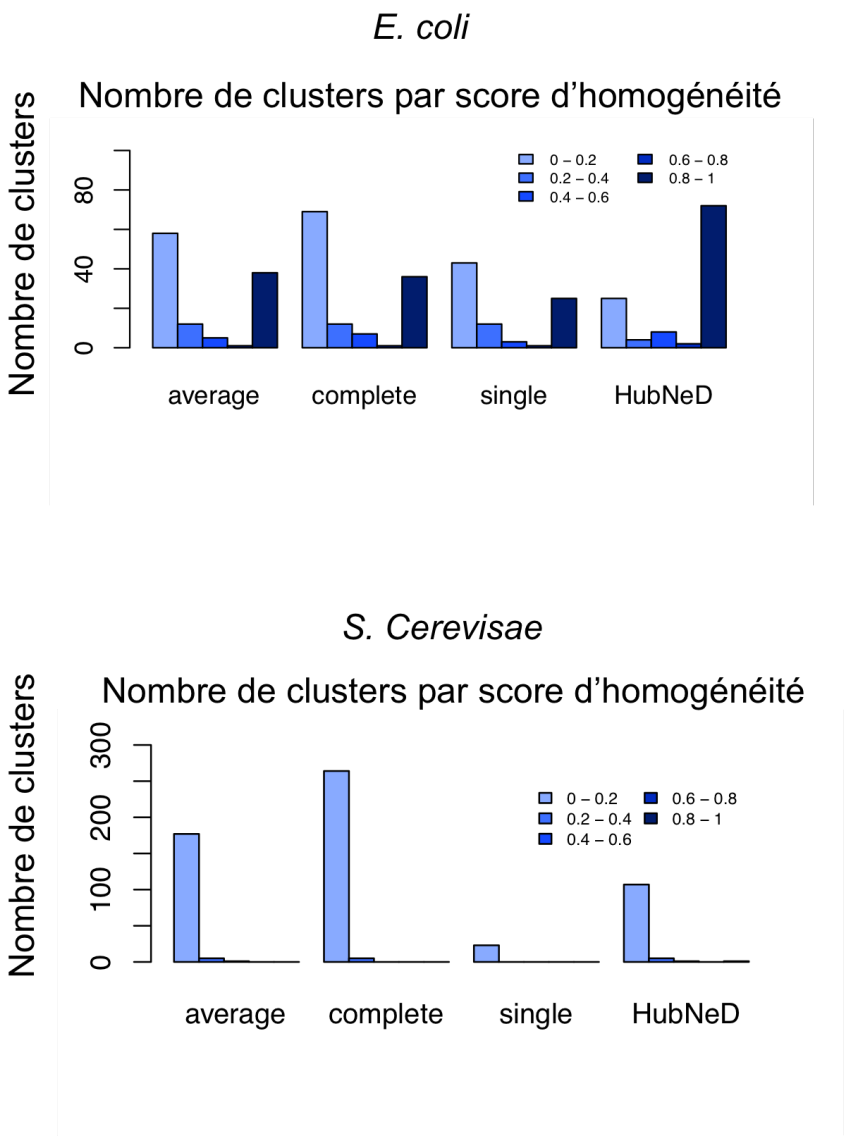


FIGURE 5.1 – **Barplot montrant les performances du clustering par co-régulation homogène d'HubNeD comparé à trois différentes méthodes de clustering partiel.** Les performances sont évaluées par l'homogénéité des clusters extraits. Pour un cluster spécifique, un score d'homogénéité de 0 signifie que les gènes dans le cluster ne partagent pas de régulateur commun et un score d'homogénéité de 1 signifie que tous les gènes ont exactement les mêmes régulateurs. HubNed est clairement la meilleure méthode pour capturer des grappes de co-régulation parfaitement homogènes pour *E.coli*. Toutes les méthodes fonctionnent mal pour *S. cerevisiae*.

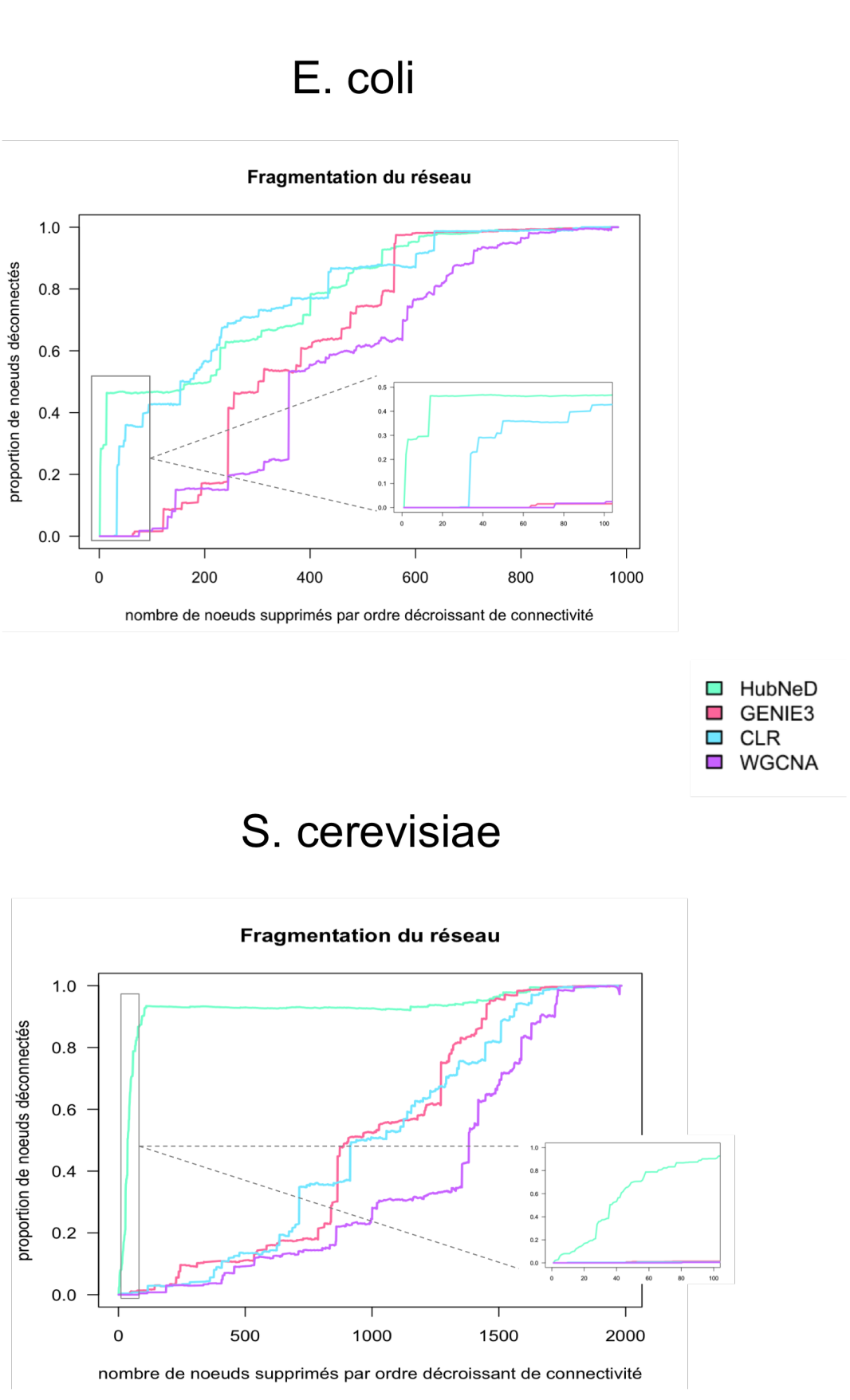


FIGURE 5.2 – Capacité des méthodes à fragmenter les RRGs d'*E.coli* et de *S. cerevisiae*

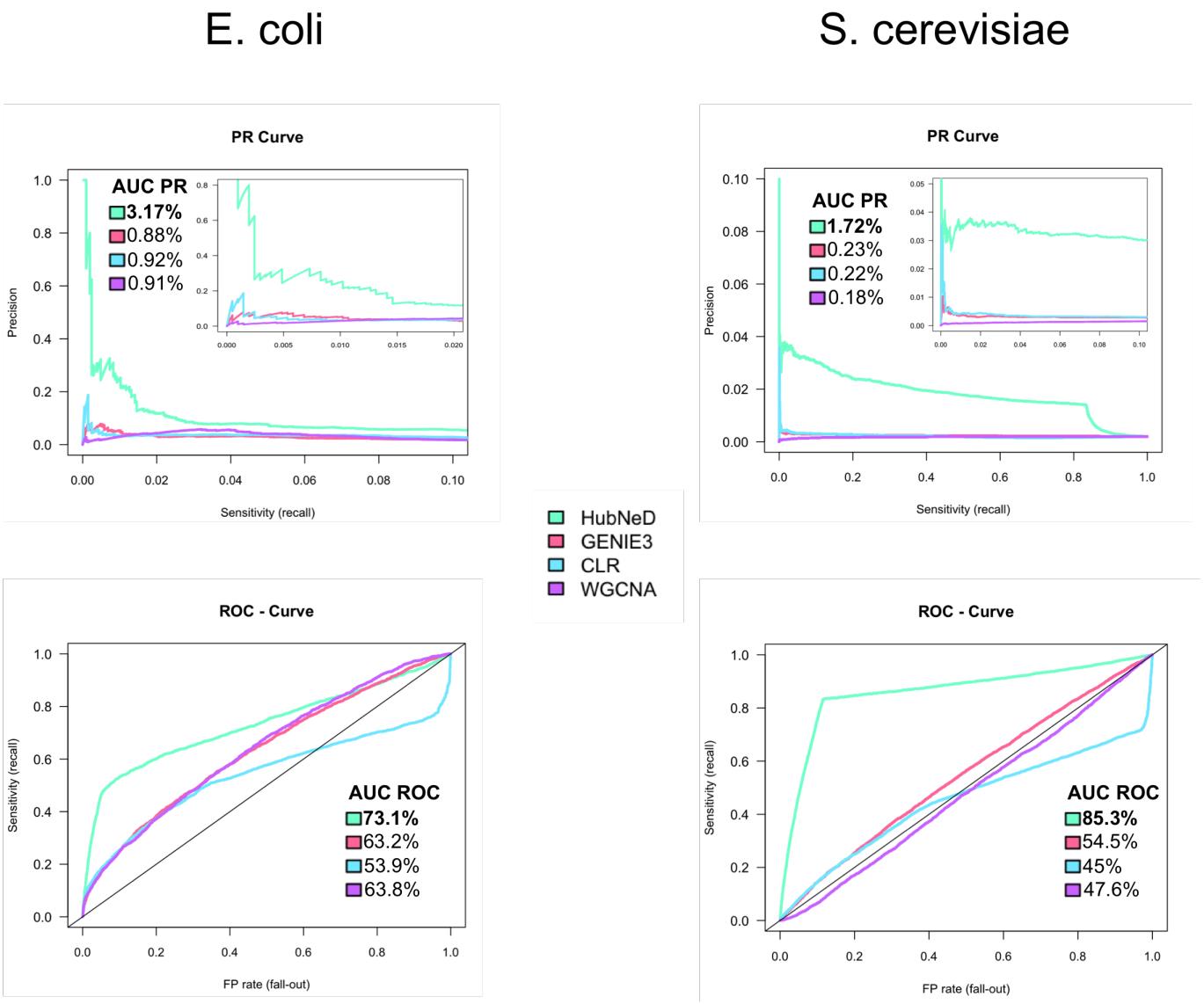


FIGURE 5.3 – **Courbes ROC et PR pour** *E.coli* **et** *S. cerevisiae*.

Chapitre 6

Application d'HUbNeD aux transcriptomes de cellules uniques

L'hématopoïèse, le processus de formation des cellules du sang (globules rouges, globules blancs et plaquettes), a lieu chez les mammifères adultes dans la moelle osseuse. L'ensemble des cellules des différents lignages est généré à partir d'un très petit nombre des cellules, les Cellules Souches Hématopoïétiques (CSH). Ces dernières ont la capacité de se reproduire à l'identique (auto-renouvellement) tout en donnant naissance aux cellules les plus immatures des différents lignages. L'équilibre auto-renouvellement/différenciation est rendu possible en état stationnaire par l'asymétrie de division de chaque CSH. Avant la naissance l'hématopoïèse a lieu dans d'autres sites que la moelle osseuse. Les premières CSHs apparaissent dans l'aorte dorsale chez l'embryon des vertébrés, puis migrent vers le foie foetal où elles s'amplifient avant de finalement coloniser la moelle osseuse en fin de gestation où elles seront maintenues tout au long de la vie.

L'émergence des premières CSH dans l'aorte dorsale embryonnaire résulte d'un processus de trans-différenciation des cellules endothéliales de la paroi ventrale de

l'aorte. Ce processus très court, de deux jours chez la souris et de quelques heures chez le poisson-zèbre, est déclenché par l'induction dans les cellules endothéliales (dites hémogéniques) de facteurs de transcription (FT) tel l'homéobox *Runx1*. Ces cellules vont alors acquérir d'autres marqueurs caractéristiques des cellules hématopoïétiques, alors qu'elles vont perdre les marqueurs caractéristiques des cellules endothéliales. On qualifie ce processus unique au cours du développement de transition endothélio-hématopoïétique (EHT).

Dans [68], Pereira et co comparent les transcriptomes de cellules uniques de deux populations cellulaires, les précurseurs hémogéniques (HP) d'une part, et les cellules souches et progéniteurs hématopoïétiques (HSPC) d'autre part. Comme indiqué plus haut la première population HP donne naissance à la seconde HSPC par trans-différenciation.

En décrivant la méthodologie d'application d'HubNeD aux données transcriptomiques générées par RNA-seq décrites dans cet article, nous montrons comment l'information d'un trait biologique peut être intégrée à la procédure HubNeD. Nous avons retiré de la matrice d'expression récupérée de [68] les gènes dont toutes les mesures d'expression étaient nulles, pour ainsi obtenir une matrice avec 16059 gènes et 144 observations indépendantes. Chaque observation correspond aux mesures d'expression des gènes d'une cellule unique. Le trait biologique est représenté par la nature des cellules étudiées, caractérisé ici par les phénotypes cellulaires HP et HSPC. L'analyse différentielle entre transcriptomes des cellules HP d'une part, et cellules HSPC d'autre part, permet ainsi de mettre en évidence des gènes essentiels à cette spécialisation. Dans ce qui suit nous commencerons par décrire la méthodologie suivie pour effectuer l'analyse différentielle des transcriptomes puis nous expliquerons comment les résultats de cette analyse ont été intégrés à la procédure HubNeD pour générer un réseau décrivant le système biologique HP/HSPC.

6.0.1 Expression différentielle basée sur l'analyse de variance.

Pour identifier les gènes différentiellement exprimés entre les deux conditions biologiques, nous les considérons un à un, et à chacun nous appliquons un test statistique classique d'analyse de variances anova (pour "analysis of variance"). Ce test développé par Ronald Fischer [69] permet de mesurer la différence statistique qui existe au sein des mesures d'expression d'un gène entre les deux conditions. Le but est donc de déterminer, pour un gène donné, si son expression dans chaque condition sont issues de deux distributions différentes (hypothèse positive : les conditions ont un effet sur l'expression du gène) ou s'ils représentent deux échantillons d'une même distribution (hypothèse nulle : les conditions n'ont pas d'effet sur l'expression du gène). Le résultat de ce test est une p-valeur, c'est à dire la probabilité d'obtenir, en supposant l'hypothèse nulle (en supposant que les expressions dans les deux conditions sont des échantillons d'une même distribution), la différence observée entre les mesures d'expressions dans les deux conditions. Ce test se base sur l'analyse comparative des variances inter et intra conditions. Il faut donc s'assurer, pour pouvoir appliquer ce test, que les expressions des différents gènes ont des variances comparables. Une caractéristique importante de la technologie RNA-seq est qu'elle produit des données d'expression déséquilibrées suivant la loi binomiale négative. Sans rentrer dans les détails de cette distribution, nous n'en mentionnerons ici qu'une caractéristique importante pour notre analyse : les mesures d'expressions d'un petit nombre de gènes varient très fortement comparativement à la grande majorité des gènes dont les expressions varient peu. Il est donc nécessaire de passer par une étape de normalisation avant d'appliquer des tests statistiques basés sur la comparaison des variances. Nous commençons donc par transformer la matrice d'expression des gènes par un logarithme en base deux. Cette opération est classiquement utilisée pour normaliser des données permettant de corriger le déséquilibre des variances et permettant ainsi d'appliquer les tests anova.

Nous pouvons maintenant décrire formellement la méthode d'analyse de variance en considérant l'exemple d'un seul gène X et deux conditions A et B. On écrira X^A

et X^B pour représenter les expressions du gène X sur respectivement les conditions A et B. On supposera également que la condition A contient n_A mesures et la condition B n_B mesures. Il y a donc au total $n = n_A + n_B$ mesures. La moyenne des expressions de X s'écrit donc : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$; et les moyennes des expressions de X au sein de chaque condition s'écrivent : $\overline{X^A} = \frac{1}{n_A}\sum_{i=1}^{n_A} X_i^A$ et $\overline{X^B} = \frac{1}{n_B}\sum_{i=1}^{n_B} X_i^B$.

Pour mesurer la variation totale des mesures d'expression du gène X, nous considérons la somme des carrés des écarts à la moyenne. On définit ainsi la Somme des Carrés Totale par : $SCT = \sum_{i=1}^{n} (X_i - \overline{X})^2$. L'idée de l'anova est de décomposer cette variation totale pour en mesurer la part qui est due aux variations au sein des mêmes conditions (variation intra-conditions) et la part qui est due à la variation entre les conditions (variation inter-conditions). En effet, la variation totale se décompose comme suit : $SCT = SC_{intra} + SC_{inter}$ avec

$SC_{intra} = \sum_{i=1}^{n^A} (X_i^A - \overline{X^A})^2 + \sum_{i=1}^{n^B} (X_i^B - \overline{X^B})^2$ et

$SC_{inter} = n_A(\overline{X^A} - \overline{X})^2 + n_B(\overline{X^B} - \overline{X})^2$.

Nous cherchons ici à tester l'hypothèse nulle H_0 = " les conditions n'ont pas d'effet sur les expressions du gènes X ". Nous mesurons alors la probabilité d'obtenir les données d'expression du gène X sachant H_0 et rejetons H_0, donc supposerons que les conditions ont un effet sur les expressions de X, si cette probabilité est inférieure à un seuil fixé. Pour ce faire nous définissons la statistique $F = \frac{\frac{SC_{inter}}{1}}{\frac{SC_{intra}}{n-2}}$ qui sous l'hypothèse H_0 suit une loi de Fischer de degrés de libertés 1 et $n-2$, respectivement le dénominateur du numérateur (définit comme le nombre de conditions moins 1), et le dénominateur du dénominateur (définit comme le nombre d'observations moins le nombre de conditions). Il est alors possible d'associer une p-valeur à cette statistique, c'est à dire $P(F|H_0)$, la probabilité d'obtenir pour F la valeur calculée à partir des données sachant l'hypothèse H_0. Si on fixe pour seuil 5%, nous déciderons que les conditions ont un effet sur l'expression du gène X si $P(F|H_0) < 0.05$, et par là prendrons un risque inférieur à 5% de se tromper. Quand nous appliquons ce test aux 16059 gènes de la matrice d'expression, nous obtenons un p-valeur associée à chaque gène. En sélectionnant tous les gènes dont la p-valeur est inférieure à 0.05,

nous risquons d'avoir jusqu'à 5% de fausses prédictions parmi les résultats positifs. Une dernière étape consiste donc à corriger les p-valeurs pour des tests multiples, et nous faisons cela en multipliant les p-valeurs par le nombre de tests réalisés (c'est à dire ici 16059 pour le nombre de gène). Cette correction, proposée par Bonferroni, n'est pas la seule. Bien qu'elle ne soit pas la plus sophistiquée, nous optant pour cette correction car c'est celle qui génère le moins de faux positifs. Nous sélectionnant pour finir tous les gènes sont les p-valeurs corrigées sont inférieures à 0.05. Ces gènes seront référencés dans la suite par DE pour gènes Différentiellement Exprimés.

6.0.2 Intégration du trait biologique à la méthode HUbNeD.

Nous allons maintenant décrire les deux étapes de l'analyse des données transcriptomiques par HubNeD. Pour rappel, la première consiste à extraire des groupes de co-régulation et la deuxième à identifier les hubs parmi les gènes absents de ces groupes.

Groupes de co-régulation associés au trait biologique.

Les groupes de co-régulation ont été construits en appliquant HubNeD à la matrice d'expression globale. Nous nous intéresserons ici aux groupes de co-régulation qui contiennent au moins un gène dans DE (Tableau $coReg_DE.csv$ en annexe). Cette approche à deux intérêts. Premièrement, elle permet d'organiser en les groupant un sous-ensemble de gènes dans DE. Deuxièmement, nous étendons les gènes dans DE à d'autres gènes qui n'ont pas été sélectionnés par l'analyse d'expression différentielle décrite plus haut. En effet, certains de ces groupes contiennent des gènes qui n'ont pas été sélectionnés par l'analyse d'expression différentielle (Tableau $coReg_DE.csv$ en annexe). Comme nous partons du principe que chaque groupe de co-régulation contient des gènes partageant le même programme de régulation, nous supposons qu'ils participent aux mêmes fonctions biologiques. Ainsi, un gène absent de DE présent dans un même groupe de co-régulation qu'un gène dans DE représente un intérêt pour le trait biologique.

Nous associons alors chacun de ces groupes de co-régulation à une ou l'autre des populations cellulaires HP ou HSPC en fonction des gènes DE qu'il contient. Par définition, les mesures d'expression d'un gène dans DE sont significativement supérieures dans les cellules HP que dans les cellules HSPC ou, inversement, supérieures dans HSPC par rapport à HP. Dans le premier cas on dira que le gène est associé à HP et dans le deuxième cas à HSPC. Un groupe de co-régulation est alors dit associé à HP (respectivement HSPC) si les gènes dans DE qu'il contient sont associés à HP (respectivement HSPC). Nous avons ainsi reconstruit 29 groupes de co-régulation associés au trait comportant de 3 à 36 gènes. Les gènes de chacun de ces groupes correspondent à une fonction biologique définie par une à trois catégories d'ontologie (Tableau $coReg_DE.csv$ en annexe). A titre d'exemples, pour les réseaux associés à HP (10 au total) on trouve :

-le groupe de co-régulation n°29 contenant 27 gènes impliqués dans l'angiogenèse, la formation de la membrane basale et la constitution des jonctions serrées. Parmi ces gènes il y a le FT *Epas1* qui a un rôle essentiel dans l'angiogenèse,

-le n° 120 contient 5 gènes "chaperone" impliqués dans la conformation correcte des protéines.

Pour les réseaux associés à HSPC (18 au total) on trouve :

-le n°7 contenant 36 gènes impliqués dans la réaction inflammatoire, la réponse immunitaire et le chimiotactisme,

-le n°116 contenant 23 gènes impliqués dans le cycle cellulaire,

-le n°12 contenant 24 gènes impliqués dans la traduction de protéines.

L'ensemble des réseaux associés à l'une des populations cellulaires est cohérent avec les fonctions reconnues de HP ou HSPC.

Hubs associés au trait.

Pour l'identification des hubs nous avons choisi de limiter la recherche aux gènes impliqués dans la transcription, c'est à dire les FTs, les cofacteurs formant avec

les FTs des complexes moléculaires permettant la trans-activation, et les facteurs épigénétiques qui rendent accessibles ou non l'accès au double brin d'ADN. Le choix de ces gènes est justifié par l'importance critique de ces facteurs pour la réalisation du processus d'EHT. Nous avons ainsi identifié 1765 gènes en utilisant la base de données DAVID [70] que nous nommerons globalement FTs dans la suite. Nous recherchons ainsi les hubs parmi les FTs n'appartenant pas aux groupes de co-régulation. HubNeD permet d'ordonner ces gènes en calculant pour chacun un score d'hub. Les gènes de plus hauts scores sont les gènes auxquels nous accordons le plus de confiance d'être des hubs du génome complet de la souris. Enfin, un réseau a été construit en sélectionnant parmi les 100 gènes en haut du classement des hubs les gènes les plus fortement corrélés aux groupes de co-régulation associés au trait. Nous avons ainsi sélectionné 71 FTs, 45 associés à HP et 26 à HSPC (Tableau $TF_trait.csv$ en annexe). A titre d'exemples, pour les FTs associés à HSPC on trouve *Ikzf1, Cebpb, Gata1, Notch2, Runx1 et Ezh2*, largement reconnus pour leur rôle dans le maintien de l'auto-renouvellement et/ou la différenciation des CSH. Pour les FT associés à HP on trouve 6 gènes, *Ets1, Fli1, Smad1, Sox17, Sox18, et Sox7*, rapportés pour leur rôle critique dans le processus d'EHT (Figure 6.1).

Ces résultats montrent tout l'intérêt à appliquer HubNed pour l'étude du transcriptome des cellules de mammifères. Cette méthodologie permet non seulement de trouver des groupes de co-régulation associés au trait mais aussi, parmi les hubs inférés pour le RRG global, ceux qui, parce qu'ils sont en amont des réseaux, sont possiblement des régulateurs des processus biologiques caractéristiques des groupes de co-régulation. HubNed associe donc les avantages de deux types d'algorithmes, ceux, tel WGCNA, qui mettent en évidence des modules de co-régulation, et ceux, tels Genie3, qui identifient les FTs en amont des modules. Néanmoins, la comparaison des connectivités (Figure 6.2) montrent qu'à partir de ces données les méthodes prédisent les mêmes hubs dans l'ensemble. D'autres études comparatives se basant sur de nouvelles données permettraient de mieux souligner l'avantage d'utiliser HubNeD par rapport à d'autres algorithmes.

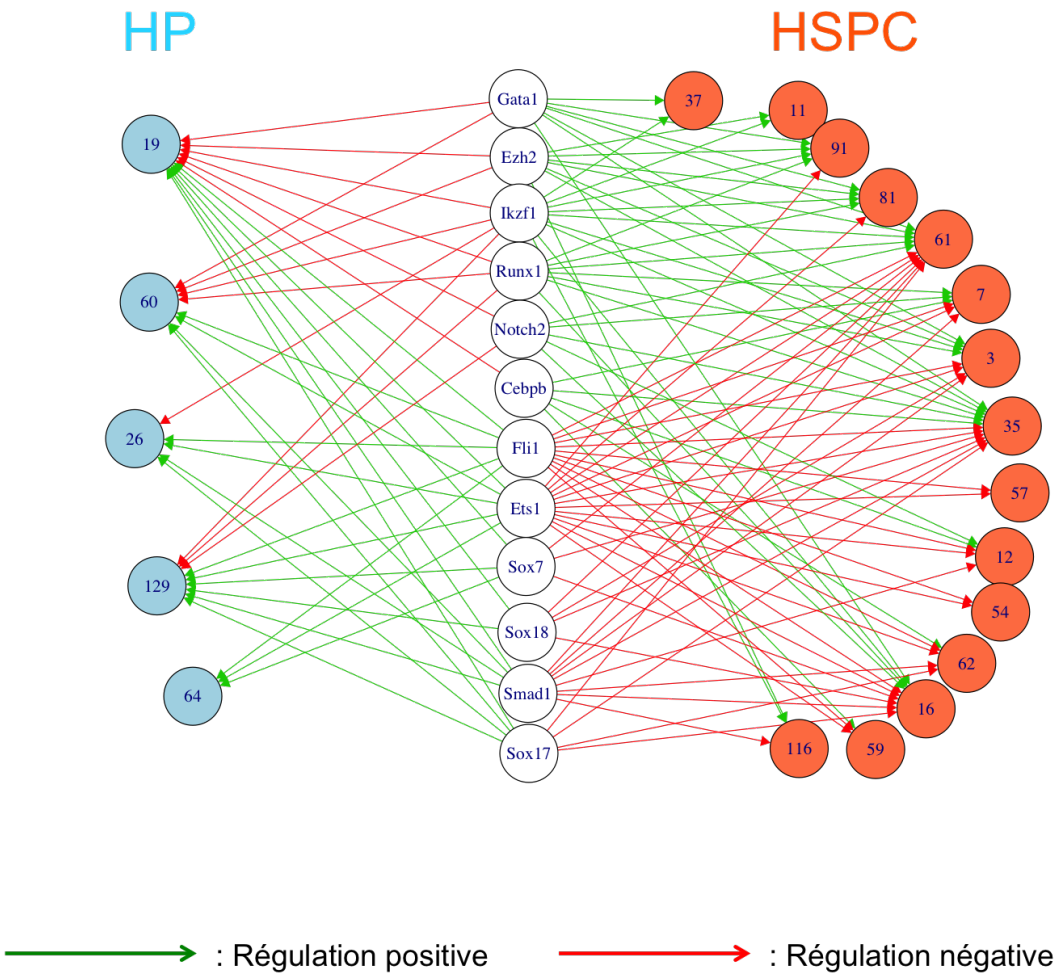


FIGURE 6.1 – **Réseau reconstruit pour les hubs les plus fortement corrélés aux groupes de co-régulation associés au trait HP/HSPC.** Les groupes de co-régulation sont numérotés de 1 à 29 est sont coloriés en oranges pour les groupes associés à HSPC et en bleu pour les groupes associés à HP (ces groupes sont décrits dans la Table1 en annexe). Les arêtes sont en coloriés en vert pour les régulations positives et en rouge pour les régulations négatives. Les familles de fonctions biologiques de certains groupes de co-régulation sont mentionnées à coté des groupes.

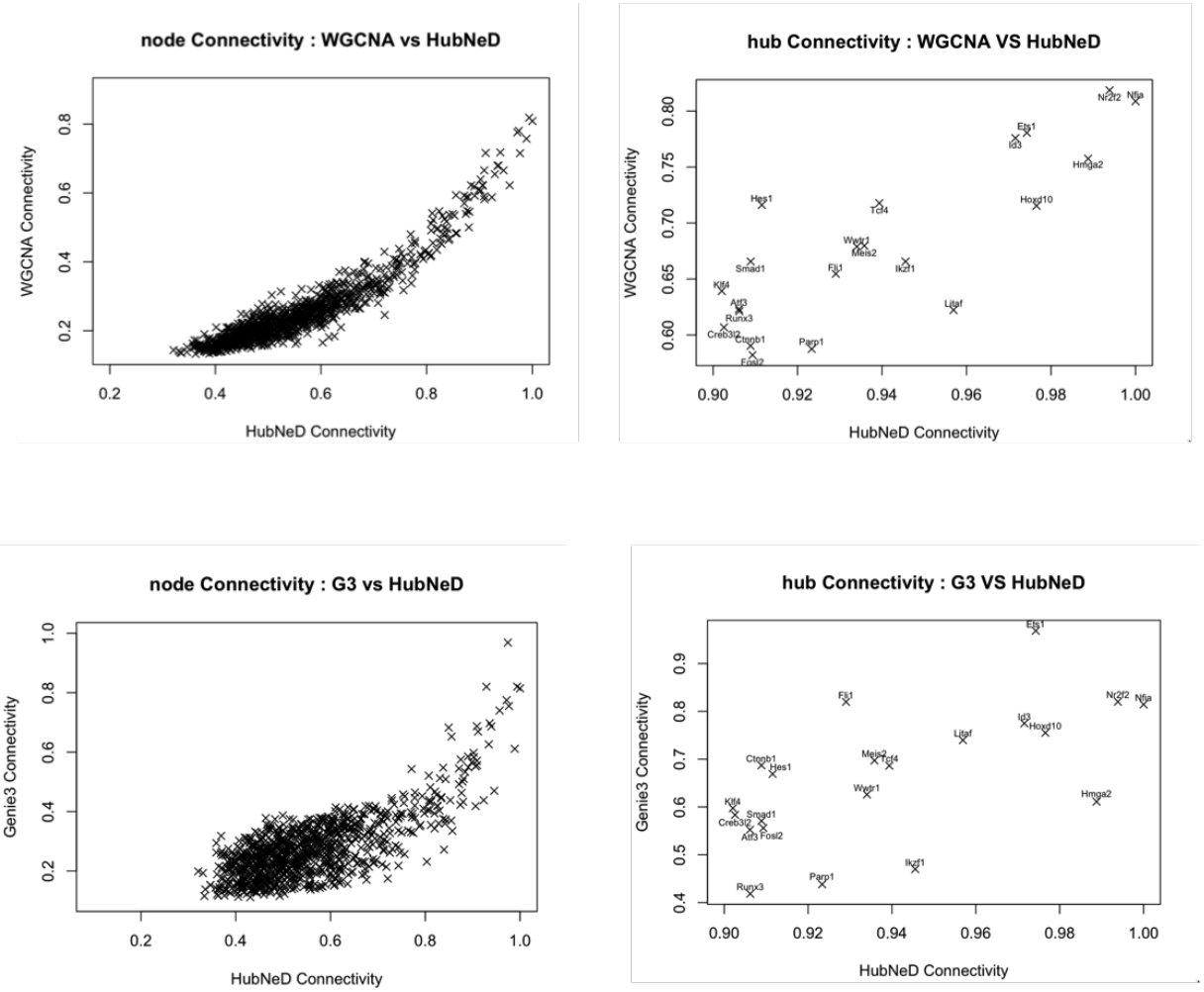


FIGURE 6.2 – **Comparaison des connectivités des noeuds des réseaux reconstruits par HubNeD d'une part et WGCNA ou Genie3 d'autre part.**

Chapitre 7

Conclusions et perspectives

HubNeD a été conçu pour reconstruire des RRGs sur la base de trois hypothèses :

1. Les RRGs sont organisés autour d'un petit nombre de hubs très dominant en terme de connectivité.

2. Les RRGs sont peu denses en arêtes et contiennent une faible proportion de triangles.

3. Les corrélations significatives calculées à partir des données d'expression de gènes (données réelles) sont des signaux de co-regulation.

L'existence de hubs dans les RRGs est l'hypothèse principale pour HubNeD (hypothèse 1). La stratégie originale d'HubNeD est d'inférer les hubs puis de reconstruire le GRN autour d'eux en évitant la formation de triangles (hypothèse 2). Nous avons observé que, dans le contexte des données simulées, les corrélations significatives sont directes ou indirectes (elles correspondent respectivement à des régulations ou à des co-régulations) alors que, lorsqu'elles sont calculées à partir des données réelles, les corrélations significatives sont indirectes. Pour identifier les hubs du système, HubNeD exploite les corrélations les plus significatives pour extraire des clusters

homogènes de co-régulation CHCs (clusters comprenant des gènes susceptibles de partager les mêmes programmes de régulation) (hypothèse 3).

La distinction entre corrélations directes et indirectes, les dernières résultant de la convolution des premiers, constitue une étape essentielle de la reconstruction d'un réseau causal. Le nettoyage des scores de similarité en supprimant les corrélations indirectes est appelée déconvolution. Dans le contexte de la reconstruction des RRGs, cette tâche n'est pas facile et toutes les méthodes existantes ont mis au point des stratégies adaptées pour résoudre ce problème, la plupart du temps en analysant les structures de similarité locales. ARACNE par exemple, l'une des premières méthodes de reconstruction de RRGs, développée en 2006, considère tous les triplets possibles et interdit de lier la paire dont le score d'information mutuelles est le plus faible. Des approches locales similaires sont courantes dans les méthodes de reconstruction des RRGs. Néanmoins, la plupart de ces méthodes supposent que les corrélations les plus élevées sont des signaux de régulation, ce qui génère beaucoup d'erreur de prédiction. En effet, nous avons montré que la correspondance entre les scores de corrélation et les RRGs associés est différente lorsque les scores sont calculés à partir de données d'expression simulées ou réelles. Les méthodes appliquées aux données simulées (DREAM4) permettent d'identifier un sous-ensemble raisonnable des arêtes présentes dans le réseau de référence. Cependant, lorsque ces méthodes sont appliquées à des données d'expression réelles de *E. coli* et de *S. cerevisiae*, les performances de ces méthodes baissent considérablement. La raison de cette diminution importante est que la similarité entre la concentration en ARNm d'un régulateur et la concentration en ARNm d'une de ses cibles est très souvent nettement inférieure à la similarité entre les concentrations en ARNm de deux cibles co-régulées. Cela s'explique par le fait que la concentration en ARNm n'est pas le seul vecteur de contrôle de la régulation, qui peut par exemple être déclenchée par une régulation post-transcriptionnelle ou post-traductionnelle (modifications de la phosphorylation par exemple).

Une fois les hubs déduits, une étape de reconstruction locale du réseau autour de ses hubs, basée sur une déconvolution centrée, pénalise les triangles en privilégiant non pas les arêtes de score les plus élevées mais les arêtes connectées aux hubs. De plus, la déconvolution centrée sur les hubs conduit à un réseau partiellement orienté en forçant les arêtes incidentes aux hubs d'être dirigés vers l'extérieur (Figure 4.2). HubNed s'est avéré nettement meilleur que les autres méthodes pour reconstruire les RRGs d'*E.coli* et de *S. cerevisiae*. L'étape fondamentale de notre approche est une nouvelle méthode de clustering graphique qui capture les corrélations les plus significatives. Elle produit des clusters de co-régulation plus homogènes que ceux produits par les autres méthodes de clustering lorsqu'on les compare sur le système d'*E.coli*. Cependant, dans le contexte du système *S. cerevisiae*, où nous avons montré qu'il n'existait pas de correspondance entre les fortes corrélations calculées à partir des données d'expression et le nombre de régulateurs communs, cette approche n'a pas réussi à extraire des clusters de co-régulation homogènes. Cela indique que soit le RRG de *S. cerevisiae* considéré dans le projet DREAM5 comme réel est trop strict, manquant beaucoup de régulations réelles, soit que la co-expression est moins indicative de la co-régulation chez les organismes eucaryotes en raison du niveau plus élevé des régulations post-transcriptionnelles et post-traductionnelles dans les organismes complexes. En tout cas, malgré l'incapacité de toutes les méthodes d'extraire des clusters de co-régulation homogènes, HubNeD est la seule méthode qui permet une réelle percée dans la reconstruction du réseau de *S. cerevisiae*, indiquant que notre stratégie d'inférence des hubs en amont de la reconstruction du réseau permet d'apporter des éléments de réponses à des problèmes qui jusque là n'avaient aucune solution, même partielle.

Enfin, nous avons décrit dans le dernier chapitre comment HubNeD a été utilisé pour analyser des données transcriptomiques de cellules uniques de la souris. Nous avons ainsi reconstruit des groupes de co-régulation homogènes d'un point de vue fonctionnel. Nous avons alors montré comment un trait biologique peut être intégré pour sélectionner les groupes de co-régulation associés à la fonction biolo-

gique d'intérêt. L'identification des hubs permet ainsi de sélectionner parmi ceux-ci qui affichent les plus fortes corrélations avec les groupes de co-régulation associés au trait.

Bibliographie

[1] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507) :1304–1351, 2001.

[2] International Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011) :931, 2004.

[3] Elizabeth Pennisi. Encode project writes eulogy for junk dna, 2012.

[4] L Pray. Eukaryotic genome complexity. *Nature Education*, 1(1) :96, 2008.

[5] Patrick C Phillips. Epistasis ?the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11) :855, 2008.

[6] Carsten Carlberg and Ferdinand Molnár. *Mechanisms of gene regulation*. Springer, 2014.

[7] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258) :561, 1970.

[8] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804) :651, 2000.

[9] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586) :1551–1555, 2002.

[10] Aljoscha Palinkas, Sascha Bulik, Alexander Bockmayr, and Hermann-Georg Holzhütter. Sequential metabolic phases as a means to optimize cellular output in a constant environment. *PloS one*, 10(3) :e0118347, 2015.

[11] Benno Schwikowski, Peter Uetz, and Stanley Fields. A network of protein–protein interactions in yeast. *Nature biotechnology*, 18(12) :1257, 2000.

[12] Dário Abdulrehman, Pedro Tiago Monteiro, Miguel Cacho Teixeira, Nuno Pereira Mira, Artur Bastos Lourenço, Sandra Costa dos Santos, Tânia Rodrigues Cabrito, Alexandre Paulo Francisco, Sara Cordeiro Madeira, Ricardo Santos Aires, et al. Yeastract : providing a programmatic access to curated transcriptional regulatory associations in saccharomyces cerevisiae through a web services interface. *Nucleic acids research*, 39(suppl_1) :D136–D140, 2010.

[13] Zhanzhi Hu, Patrick J Killion, and Vishwanath R Iyer. Genetic reconstruction of a functional transcriptional regulatory network. *Nature genetics*, 39(5) :683, 2007.

[14] Socorro Gama-Castro, Verónica Jiménez-Jacinto, Martín Peralta-Gil, Alberto Santos-Zavaleta, Mónica I Peñaloza-Spinola, Bruno Contreras-Moreira, Juan Segura-Salazar, Luis Muñiz-Rascado, Irma Martínez-Flores, Heladia Salgado, et al. Regulondb (version 6.0) : gene regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucleic acids research*, 36(suppl_1) :D120–D124, 2008.

[15] Julio A Freyre-Gonzalez and LG Trevino-Quintanilla. Analyzing regulatory networks in bacteria. *Nat Educ*, 3 :24, 2010.

[16] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439) :509–512, 1999.

[17] Réka Albert and Albert-László Barabási. Topology of evolving networks : local events and universality. *Physical review letters*, 85(24) :5234, 2000.

[18] Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9) :5766–5771, 2002.

[19] Gábor Csányi and Balázs Szendrői. Structure of a large social network. *Physical Review E*, 69(3) :036131, 2004.

[20] Reuven Cohen, Keren Erez, Daniel Ben-Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Physical review letters*, 85(21) :4626, 2000.

[21] Simon R Broadbent and John M Hammersley. Percolation processes : I. crystals and mazes. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 53, pages 629–641. Cambridge University Press, 1957.

[22] Uri Alon. Network motifs : theory and experimental approaches. *Nature Reviews Genetics*, 8(6) :450, 2007.

[23] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4) :663–676, 2006.

[24] Peter Langfelder and Steve Horvath. Wgcna : an r package for weighted correlation network analysis. *BMC Bioinformatics*, (1) :559, 2008.

[25] Peter Langfelder and Steve Horvath. Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software*, 46(11) :1–17, 2012.

[26] Antonio Musso and Vukan R Vuchic. *Characteristics of metro networks and methodology for their evaluation.* National Research Council, Transportation Research Board, 1988.

[27] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv :1111.4503*, 2011.

[28] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1) :17–60, 1960.

[29] Sergey Edunov, Carlos Diuk, Ismail Onur Filiz, Smriti Bhagat, and Moira Burke. Three and a half degrees of separation.

[30] Emile F Nuwaysir, Wei Huang, Thomas J Albert, Jaz Singh, Kate Nuwaysir, Alan Pitas, Todd Richmond, Tom Gorski, James P Berg, Jeff Ballin, et al. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome research*, 12(11) :1749–1755, 2002.

[31] Y Tu, G Stolovitzky, and U Klein. Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences*, 99(22) :14031–14036, 2002.

[32] Ryan D Morin, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor J Pugh, Helen McDonald, Richard Varhol, Steven JM Jones, and Marco A Marra. Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing. *Biotechniques*, 45(1) :81–94, 2008.

[33] Yongjun Chu and David R Corey. Rna sequencing : platform selection, experimental design, and data interpretation. *Nucleic acid therapeutics*, 22(4) :271–274, 2012.

[34] Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, Gustavo Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8) :796–804, 2012.

[35] Gökmen Altay and Frank Emmert-Streib. Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*, 26(14) :1738–1744, 2010.

[36] Frank Emmert-Streib, Galina Glazko, Ricardo De Matos Simoes, et al. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. *Frontiers in genetics*, 3 :8, 2012.

[37] Stefan R Maetschke, Piyush B Madhamshettiwar, Melissa J Davis, and Mark A Ragan. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in bioinformatics*, 15(2) :195–211, 2013.

[38] Daniel Marbach, Robert J Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences*, 107(14) :6286–6291, 2010.

[39] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58 :240–242, 1895.

[40] Claude E Shannon, Warren Weaver, and Arthur W Burks. The mathematical theory of communication. 1951.

[41] Ricardo de Matos Simoes and Frank Emmert-Streib. Influence of statistical estimators of mutual information and data heterogeneity on the inference of gene regulatory networks. *PLoS One*, 6(12) :e29279, 2011.

[42] Andy M Yip and Steve Horvath. The generalized topological overlap matrix for detecting modules in gene networks.

[43] Yinyin Yuan, Chang-Tsun Li, and Oliver Windram. Directed partial correlation : inferring large-scale gene regulatory network through induced topology disruptions. *PLoS One*, 6(4) :e16835, 2011.

[44] Alberto De La Fuente, Nan Bing, Ina Hoeschele, and Pedro Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18) :3565–3574, 2004.

[45] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne : an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(1) :S7, 2006.

[46] Thomas M Cover and Joy A Thomas. Elements of information theory 2nd edition. 2006.

[47] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1) :e8, 2007.

[48] Gökmen Altay and Frank Emmert-Streib. Inferring the conservative causal core of gene regulatory networks. *BMC systems biology*, 4(1) :132, 2010.

[49] Ricardo de Matos Simoes and Frank Emmert-Streib. Bagging statistical network inference from large-scale gene expression data. *PLoS One*, 7(3) :e33624, 2012.

[50] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[51] Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. Tigress : trustful inference of gene regulation using stability selection. *BMC systems biology*, 6(1) :145, 2012.

[52] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[53] Patrick E Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology*, 2007 :8–8, 2007.

[54] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8) :1226–1238, 2005.

[55] Alexandre Irrthum, Louis Wehenkel, Pierre Geurts, et al. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9) :e12776, 2010.

[56] Judea Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Elsevier, 2014.

[57] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

[58] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

[59] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464, 1978.

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

[61] Subramani Mani. *A bayesian local causal discovery framework*. PhD thesis, University of Pittsburgh, 2006.

[62] Ka Yee Yeung, Roger E Bumgarner, and Adrian E Raftery. Bayesian model averaging : development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, 21(10) :2394–2402, 2005.

[63] JJ Faith, B Hayete, JT Thaden, I Mogno, J Wierzbowski, G Cottarel, et al. Supplemental website for : Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles, 2007.

[64] Gustavo Stolovitzky, DON Monroe, and Andrea Califano. Dialogue on reverse-engineering assessment and methods. *Annals of the New York Academy of Sciences*, 1115(1) :1–22, 2007.

[65] Robert J Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K Sorger, Leonidas G Alexopoulos, Xiaowei Xue, Neil D Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. Towards a rigorous assessment of systems biology models : the dream3 challenges. *PloS one*, 5(2) :e9202, 2010.

[66] Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver : in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16) :2263–2270, 2011.

[67] Peter Langfelder and Steve Horvath. Fast r functions for robust correlations and hierarchical clustering. *Journal of statistical software*, 46(11), 2012.

[68] Carlos-Filipe Pereira, Betty Chang, Andreia Gomes, Jeffrey Bernitz, Dmitri Papatsenko, Xiaohong Niu, Gemma Swiers, Emanuele Azzoni, Marella FTR de Bruijn, Christoph Schaniel, et al. Hematopoietic reprogramming in vitro informs in vivo identification of hemogenic precursors to definitive hematopoietic stem cells. *Developmental cell*, 36(5) :525–539, 2016.

[69] Ronald A Fisher. Xv. ?the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2) :399–433, 1919.

[70] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools : paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1) :1–13, 2008.