Inférence de réseaux de régulation de gènes au travers de scores étendus dans les réseaux Bayésiens.

Jimmy Vandel [1], **Brigitte Mangin** [1], **Matthieu Vignes** [1], **Damien Leroux** [1], **Olivier Loudet** [2], **Marie-Laure Martin-Magniette** [3,4], **Simon de Givry** [1]

1. *INRA, UR 875 Unité de Biométrie et Intelligence Artificielle, F-31326, Castanet-Tolosan, France.*

2. *INRA, UMR 1318, Institut Jean-Pierre Bourgin, F-78000, Versailles, France.*

3. *INRA, UMR 1165 Unité de Recherche en Génomique Végétale, F-91057, Evry, France.*

4. *INRA, UMR 518 Mathématiques et Informatique Appliquées, F-75231, Paris, France.*

RÉSUMÉ. *L'inférence de réseaux de régulation de gènes s'oriente actuellement vers l'utilisation conjointe d'informations biologiques complémentaires. Nous utilisons ici des données de marqueurs génétiques en plus des classiques données d'expression dans le cadre des réseaux bayésiens statiques discrets. Nous comparons les qualités de différents scores ainsi que l'impact d'un a priori lié à la connectivité des réseaux. Nous proposons et comparons deux modélisations aux approches existantes pour l'inférence de réseaux de régulation. Sur des données simulées, l'un de nos modèles obtient les meilleurs résultats dans le cas d'échantillons de petites tailles. Nous utilisons ce même modèle sur des données réelles d'Arabidopsis thaliana.*

ABSTRACT. *Inferring gene regulatory networks tends to use several biological information. Here we use data from genetic markers and expression data in the framework of discrete static bayesian networks. We compare several scores and also the impact of a network connectivity a priori. We propose and compare two models with existing approaches of gene regulatory network inference. On simulated data one of our models reached better results in the case of small sample size. We use this model on real data in Arabidopsis thaliana.*

MOTS-CLÉS : *apprentissage de structure, réseau Bayésien, régulation de gènes, génétique génomique.*

KEYWORDS: *structure learning, Bayesian network, gene regulation, genetical genomics.*

DOI:10.3166/RIA.x.1-30

1. Introduction

Le gène est l'unité fonctionnelle de l'hérédité, porteur de l'information génétique qui contient les mécanismes de fonctionnement des organismes vivants, d'une génération à l'autre. L'expression d'un gène dans une cellule peut se traduire par la production de protéines, chacune pouvant réguler l'expression d'autres gènes. Une meilleure connaissance du réseau de régulation d'un ensemble de gènes est une aide précieuse pour les biologistes dans l'analyse de caractères complexes comme la susceptibilité à certaines maladies ou la résistance à des stress hydriques et thermiques dans le cas des plantes (Mehrabian *et al.*, 2005 ; Keurentjes *et al.*, 2007).

La reconstruction de réseau de régulation est un problème complexe (Li *et al.*, 2008), en particulier du fait que les données d'expression mesurent la quantité d'un très grand nombre de gènes transcrits sur un petit échantillon. Ce cadre méthodologique découle des données sur les transcrits. En effet, les technologies de type *microarray* permettent d'observer sur une unique puce un très grand nombre de transcrits. Le coût de cette technologie, bien que diminuant continuellement, est encore trop élevé pour envisager un échantillon (de puces) de grande taille. Nous nous inscrivons ici dans le cadre de mesures effectuées en état stationnaire du fonctionnement de la cellule, sans données manquantes, en environnement contrôlé (mêmes facteurs environnementaux pour tous les individus) et excluons les phénomènes épigénétiques (*i.e.* régulation de gènes via des signaux ne provenant pas des gènes). De plus nous posons le postulat que la structure du réseau de régulation est identique pour tous les échantillons mesurés.

Plusieurs approches pour prédire la topologie de ces réseaux ont été proposées allant d'une recherche des relations locales entre gènes à une modélisation globale du réseau au travers des modèles graphiques (Jordan, 1999). Nous exposons en Section 2 un état de l'art de ces méthodes de reconstruction en nous concentrant sur celles implémentées dans des logiciels disponibles. Parmi les modèles graphiques nous nous intéressons plus spécialement aux réseaux bayésiens qui se sont révélés performants de par leur capacité à représenter des relations complexes et à modéliser les indépendances entre les variables qui les composent. Une des techniques d'apprentissage de ces réseaux consiste à parcourir un sous ensemble des réseaux possibles en attribuant pour chacun d'eux un score représentant la qualité de la structure au vu des données puis de sélectionner celui maximisant le score. Nous nous attachons en Section 3 à décrire plus en détail les différents scores proposés pour les réseaux bayésiens discrets et nous présentons un nouveau score adapté à des réseaux de faible connectivité.

Ces dernières années, plusieurs travaux (Jansen, Nap, 2001), (Ghazalpour *et al.*, 2006), (Zhu *et al.*, 2007), (Liu *et al.*, 2008), (Chu *et al.*, 2009), (Chipman, Singh, 2011) ont montré l'intérêt d'exploiter des données de polymorphisme pour mieux reconstruire un réseau de régulation de gènes. Ces données proviennent de la variabilité génétique entre individus et correspondent à l'observation de différents génotypes sur des marqueurs moléculaires pour un ensemble d'individus. Nous nous intéressons plus particulièrement au polymorphisme dit fonctionnel qui est une mutation de l'ADN

ayant pour effet de modifier le niveau d'expression du gène ou la nature de la protéine produite, influant de manière complexe (non linéaire) sur la régulation d'autres gènes. Nous disposons ainsi de deux types de données à la fois continues (expressions des gènes) et discrètes (marqueurs génétiques). Notre modélisation nous conduira à discrétiser les données d'expressions (Section 5.4).

De la même manière que (Chipman, Singh, 2011), nous proposons une modélisation par un réseau bayésien qui intègre explicitement les données de polymorphisme dans ses variables plutôt que via un *a priori* sur la structure du réseau. Nous proposons dans la Section 4, deux modélisations possibles en soulignant leurs avantages et leurs inconvénients respectifs. Nous comparons par la suite ces deux modèles à des approches existantes en Section 6 sur des données simulées décrites en Section 5.1. Puis nous appliquons en Section 7 notre modèle le plus efficace sur des données réelles d'Arabidopsis thaliana présentées en Section 5.2. Enfin nous concluons et présentons les perspectives de ce travail.

2. Etat de l'art des méthodes d'inférence de réseaux de gènes

Parmi les nombreuses méthodes existantes pour reconstruire des réseaux de régulation de gènes à partir de données d'expression, nous pouvons globalement définir 3 classes de méthodes (Bansal *et al.*, 2007). La première d'entre elles s'attache à mesurer la corrélation entre chaque paire de gènes puis détermine par seuillage les relations significatives pour en extraire le graphe final. Ces méthodes requièrent donc de l'utilisateur un seuil dont la valeur ne peut être déterminée *a priori*, de plus le caractère symétrique des corrélations ne permet pas de trouver le sens des relations. En outre cette approche distingue difficilement les régulations directes entre deux gènes de celles s'effectuant par l'intermédiaire d'un troisième gène. Plusieurs méthodes tentent de pallier ce problème notamment celles implémentées dans les logiciels ARACNE (Margolin *et al.*, 2006) et CLR (Faith *et al.*, 2007) qui se basent sur une mesure de l'information mutuelle. Le premier analyse l'ensemble des triplets connexes du graphe final en enlevant pour chacun d'eux la relation la plus faible tandis que le second adapte son seuil en fonction de l'information mutuelle moyenne de l'ensemble des gènes. De son coté le logiciel ParCorA (Fuente *et al.*, 2004) calcule la corrélation partielle de Pearson ou de Spearman, permettant ainsi de repérer les relations indirectes grâce au conditionnement par un ensemble de gènes.

La deuxième classe de méthodes représente le réseau de régulation sous la forme d'un modèle graphique gaussien (GGM) qui présume que les données suivent une loi normale multivariée et privilégie donc une modélisation linéaire des données d'expression des gènes en relation dans le graphe. L'approche traditionnelle nécessite le calcul des corrélations partielles de chaque couple de gènes sachant l'ensemble des autres gènes par inversion de la matrice de covariance Σ. Cependant dans le cas où la taille de l'échantillon tend à être plus petite que le nombre de gènes, Σ n'est plus inversible classiquement. Schäfer et Strimmer (2005) proposent dans leur logiciel GeneNet de coupler le calcul de la pseudo-inverse de Σ à une opération de bagging afin

de stabiliser l'estimation de la matrice des corrélations partielles Π. Puis un test de significativité par rapport à 0 des valeurs contenues dans la matrice Π permet d'établir les relations présentes dans le graphe. D'autres techniques permettent d'estimer la matrice Π en sélectionnant la structure qui maximise une vraisemblance pénalisée afin d'assurer la faible connectivité du graphe. Meinshausen et Bühlmann (2006) effectuent une régression Lasso indépendante pour chaque gène définissant ainsi un système d'équations pour l'ensemble des gènes (*SEM Lasso*) tandis que le logiciel SIMoNe (Chiquet *et al.*, 2009) réalise des régressions Lasso non indépendantes et fait l'hypothèse supplémentaire que le réseau de régulation possède une structure modulaire. Enfin Liu *et al.* (2008) utilisent une modélisation par équations structurales, similaires à des régressions indépendantes, dans le cadre de données de génétiques génomiques.

Certaines approches mélangent plusieurs caractéristiques des méthodes mentionnées précédemment, le logiciel GGMSelect (Giraud *et al.*, 2009) opère une stratégie de boosting et réunit des méthodes de calcul des corrélations par paires et de modélisation par GGM. Le logiciel propose trois familles de graphes, la première basée sur la corrélation de Spearman d'ordre 1 (famille "C01") tandis que les deux autres utilisent la régression Lasso indépendante par gène avec une pénalité classique en l_1 ("LA") ou adaptative ("EW"). Par la suite un critère global dédié dans le cadre des GGM permet de sélectionner le meilleur représentant parmi les graphes des différentes familles.

La dernière catégorie se compose des modèles graphiques probabilistes discrets. Dans le cas de données d'expression issues d'états stationnaires nous nous intéressons plus particulièrement aux réseaux bayésiens statiques modélisés par un graphe dirigé acyclique (DAG) où l'état de chaque noeud de ce graphe est régi par une distribution de probabilités conditionnelles. Contrairement aux deux premières classes de méthodes, les réseaux bayésiens permettent de représenter des régulations non linéaires complexes (hormis les régulations formant des cycles) tout en orientant partiellement celles-ci. Friedman *et al.* (2000) appliquent pour la première fois les réseaux bayésiens sur des données d'expressions puis Zhu *et al.* (2007) y ajoutent des données issues de marqueurs génétiques en tant qu'*a priori* sur les relations. Le logiciel SCT (Chipman, Singh, 2011) modélise les marqueurs génétiques en tant que variables à part entière du graphe, il construit dans un premier temps un ensemble d'arbres causaux stochastiques pour chaque marqueur afin d'obtenir un *a priori* sur les relations guidant ainsi l'apprentissage de la structure par une méthode de Monte Carlo par chaîne de Markov (MCMC). Le but de cet article est donc d'utiliser dans ce même cadre une modélisation des marqueurs similaire à Chipman et Singh (2011) et d'intégrer explicitement des connaissances biologiques sur la nature des polymorphismes. Notre approche utilisant les réseaux bayésiens s'est déjà montrée compétitive face aux régressions Lasso et Dantzig dans le cadre de la compétition DREAM5 (Vignes *et al.*, 2011). Cette compétition internationale permet de comparer chaque année depuis 2006 des méthodes d'apprentissage et d'inférence de systèmes biologiques variés, notamment de réseaux de régulation de gènes (Marbach *et al.*, 2010 ; Huynh-Thu *et al.*, 2010 ; Küffner *et al.*, 2012).

3. Apprentissage de Réseaux Bayésiens

3.1. Les Réseaux Bayésiens

Un réseau bayésien (Naïm *et al.*, 2008) noté $B = (\mathcal{G}, \mathbf{P}_\mathcal{G})$ est composé d'un graphe dirigé sans circuit $\mathcal{G} = (\mathbf{X}, \mathbf{E})$ dont les sommets représentent un ensemble de p variables aléatoires discrètes $\mathbf{X} = \{X_1, \ldots, X_p\}$, reliées par un ensemble $\mathbf{E}$ d'arcs, et d'un ensemble de probabilités conditionnelles $\mathbf{P}_\mathcal{G} = \{P_1, \ldots, P_p\}$ définies par la topologie du graphe : $P_i = \mathbb{P}(X_i|Pa(X_i))$ où $Pa(X_i) = \{X_j \in \mathbf{X} \,|\, \overrightarrow{(X_j, X_i)} \in \mathbf{E}\}$ est l'ensemble des parents de X_i dans $\mathcal{G}$. Un réseau bayésien B représente ainsi une distribution de probabilité sur $\mathbf{X}$ dont la loi jointe se factorise de la manière suivante :

$$\mathbb{P}(\mathbf{X}) = \prod_{i=1}^{p} \mathbb{P}(X_i|Pa(X_i))$$

Les probabilités conditionnelles se définissent au travers de l'ensemble des paramètres θ du modèle tel que

$$\mathbb{P}(X_i = k|Pa(X_i) = j) = \theta_{ijk}$$

Le nombre de paramètres indépendants définissant les probabilités conditionnelles $\mathbf{P}_\mathcal{G}$ est appelé la dimension de B et se note $Dim(B) = \sum_{i=1}^{p} Dim(P_i)$ avec $Dim(P_i) = (r_i - 1)q_i$ où r_i est égal à la taille du domaine de la variable X_i et $q_i = \prod_{X_j \in Pa(X_i)} r_j$ correspond au nombre de configurations possibles pour les parents de X_i.

3.2. Apprentissage à base de score

L'une des approches couramment utilisée pour l'apprentissage de réseaux bayésiens vise à sélectionner le graphe $\mathcal{G}$ maximisant $\mathbb{P}(\mathcal{G}|\mathbf{D})$ avec $\mathbf{D}$ représentant les données observées constituées de n échantillons, chaque échantillon étant l'observation des p variables sans manquant.

D'après le théorème de Bayes nous pouvons écrire

$$\mathbb{P}(\mathcal{G}|\mathbf{D}) = \frac{\mathbb{P}(\mathbf{D}|\mathcal{G})\mathbb{P}(\mathcal{G})}{\mathbb{P}(\mathbf{D})} \propto \mathbb{P}(\mathbf{D}|\mathcal{G})\mathbb{P}(\mathcal{G}) \tag{1}$$

où $\mathbb{P}(\mathbf{D})$ ne dépend pas de la structure.

Afin de ne privilégier aucune structure *a priori*, $\mathbb{P}(\mathcal{G})$ est généralement assimilé à une constante pour l'ensemble des graphes, revenant ainsi à sélectionner le graphe maximisant $\mathbb{P}(\mathbf{D}|\mathcal{G})$.

Ce dernier terme représente la vraisemblance des données pour un graphe considéré, cette probabilité se calcule par intégration sur les paramètres θ :

$$\mathbb{P}(\mathbf{D}|\mathcal{G}) = \int_\theta \mathbb{P}(\mathbf{D}|\mathcal{G}, \theta)\mathbb{P}(\theta|\mathcal{G})d\theta \tag{2}$$

Cette intégrale n'étant pas calculable dans le cas général, plusieurs approches d'estimation ont été proposées, chacune d'elles définissant un score que l'on cherche à maximiser.

3.2.1. Score BIC

Schwarz (1978) utilise l'approximation de Laplace pour estimer l'intégrale [2] . L'expression ainsi obtenue se simplifie sous l'hypothèse d'un grand échantillon, définissant le score *Bayesian Information Criterion* (BIC).

$$BIC(\mathcal{G}) = \log(\mathbb{P}(\mathbf{D}|\mathcal{G}, \theta^{ML})) - \frac{1}{2}\log(n)Dim(B_{\mathcal{G}}) \approx \log(\mathbb{P}(\mathbf{D}|\mathcal{G}))$$

où $B_{\mathcal{G}} = (\mathcal{G}, \theta^{ML})$ représente le réseau bayésien constitué du graphe $\mathcal{G}$ et des paramètres θ^{ML} estimés par maximum de vraisemblance à partir des n observations.

Ce score se décompose intuitivement en un premier terme calculant la log-vraisemblance des données par rapport au modèle ainsi qu'un second terme venant pénaliser les structures complexes.

3.2.2. Score BD

Plutôt que d'effectuer une approximation de la vraisemblance Cooper et Hersovits (1992) considèrent les paramètres $\theta_{\mathbf{ij}}$ comme indépendants et suivant une loi de Dirichlet d'hyper-paramètres $\alpha_{\mathbf{ijk}}$. Sous ces hypothèses l'intégrale [2] s'exprime sous la forme du score *Bayesian Dirichlet* (BD).

$$BD(\mathcal{G}) = \mathbb{P}(\mathbf{D}|\mathcal{G}) = \prod_{i=1}^{p}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(n_{ij}+\alpha_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(n_{ijk}+\alpha_{ijk})}{\Gamma(\alpha_{ijk})}$$

avec n_{ijk}, le nombre d'occurrences de la configuration $(X_i = k, Pa(X_i) = j)$ sur les n échantillons, et $n_{ij} = \sum_{k=1}^{r_i} n_{ijk}$.

Ce score requiert de la part de l'utilisateur une valeur pour l'ensemble des hyper-paramètres $\alpha_{\mathbf{ijk}}$, ce qui se révèle bien souvent impossible en pratique. Cependant sous les conditions d'un *a priori* uniforme sur les paramètres et d'une équivalence pour les structures définissant la même loi jointe (mêmes indépendances conditionnelles), le calcul de ces hyper-paramètres se simplifie tel que $\alpha_{\mathbf{ijk}} = \frac{\alpha}{\mathbf{r_i q_i}}$, où α devient l'unique paramètre du score devenu *Bayesian Dirichlet equivalent uniform* BD_{eu}. Malheureusement ce score se révèle très sensible au choix de ce paramètre (Silander *et al.*, 2007).

3.2.3. Score fNML

Silander *et al.* (2010) proposent une approche basée sur le principe du maximum de vraisemblance normalisé (*Normalized Maximum Likelihood* (NML)) (Shtarkov, 1987).

$$NML(\mathcal{G}) = \frac{\mathbb{P}(\mathbf{D}|\mathcal{G}, \theta^{ML})}{\sum_{\mathbf{D}'}\mathbb{P}(\mathbf{D}'|\mathcal{G}, \theta^{ML})}$$

où la normalisation s'effectue sur l'ensemble des jeux de données possibles de même taille. Bien que le cardinal de cet ensemble soit exponentiel, Kontkanen et Myllymäki (2007) définissent une procédure linéaire en temps afin d'effectuer un calcul local pour chaque variable multinomiale $X_i \in \mathbf{X}$. Silander *et al.* (2010) utilisent ce résultat pour définir le score *factorized Normalized Maximum Likelihood* (fNML).

$$fNML(\mathcal{G}) = \log(\mathbb{P}(\mathbf{D}|\mathcal{G}, \theta^{ML})) - \sum_{i=1}^{p} \sum_{j=1}^{q_i} \log(C_{n_{ij}}^{r_i})$$

avec $C_{n_{ij}}^{r_i}$ le coefficient de normalisation pour la variable X_i (Kontkanen, Myllymäki, 2007).

Ce score s'exprime de façon similaire au critère BIC, avec un terme de pénalité qui dans ce cas ne dépend pas uniquement de la dimension du graphe mais aussi des données à l'instar de BD_{eu} sans avoir de paramètre à régler. Par ailleurs les scores BIC et fNML se révèlent être asymptotiquement équivalents (Silander *et al.*, 2010).

3.3. Critères étendus

En s'inspirant des travaux développés par Chen et Chen (2008) dans le cadre du choix de modèle de régression linéaire, on peut montrer que le choix d'un *a priori* uniforme sur l'ensemble des structures possibles dans l'équation [1], mène à privilégier les réseaux à forte connectivité allant ainsi à l'encontre de la sélection de structures peu denses pourtant caractéristiques des réseaux de régulation connus. L'idée majeure provient du fait que le nombre de réseaux possibles augmente de façon exponentielle avec le degré entrant de chaque noeud du graphe (jusqu'à $n/2$), déséquilibrant les probabilités des classes de modèle présentant les mêmes degrés entrants pour chaque noeud. Afin de corriger ce phénomène, nous proposons d'utiliser un *a priori* uniforme sur ces classes. Pour se faire, suivant la proposition de Chen et Chen (2008), nous définissons la probabilité *a priori* de chaque modèle $\mathcal{G}_i$ par

$$\mathbb{P}(\mathcal{G}_i) \propto \tau(X_i)^{-\gamma}$$

avec $\gamma \in [0, 1]$, $\mathcal{G}_i$ le graphe restreint à X_i composé des variables $\mathbf{X}' = \{X_i \cup Pa(X_i)\}$ reliées par $\mathbf{E}' = \{\overrightarrow{(X_j, X_i)} \,|\, X_j \in Pa(X_i)\}$ et $\tau(X_i) = C_{a_i}^{p-1}$ le nombre de combinaisons de $a_i = |Pa(X_i)|$ parents possibles sur les $p-1$ parents potentiels. La valeur $\gamma = 0$ correspond à un *a priori* uniforme sur l'ensemble des DAGs tandis que la valeur $\gamma = 1$ définit un *a priori* uniforme sur les classes de connectivité. Afin d'obtenir une probabilité sur le graphe global, nous supposons l'indépendance des graphes restreints, $\mathbb{P}(\mathcal{G}) \approx \prod_{i=1}^{p} \mathbb{P}(\mathcal{G}_i)$.

Il apparait évident que cette dernière approximation surestime le nombre réel de réseaux bayésiens présentant la même séquence sur les degrés entrants a_i, $\forall i \in \{1, \ldots, p\}$ en raison de leur caractère acyclique. Il n'existe par exemple aucun réseau bayésien à deux variables dont chacun possède un seul parent, alors que notre approximation comptabilisera un réseau possible (cyclique). La figure 1 montre l'écart entre

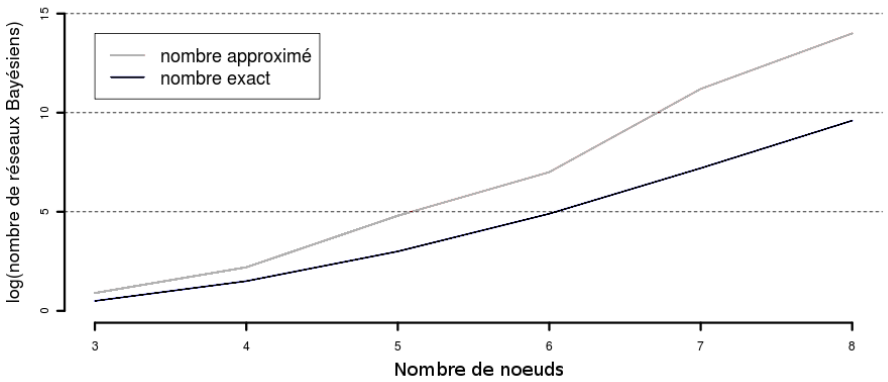


Figure 1. Comparaison du logarithme du nombre réel (noir) et approximé (gris) de DAGs dont la connectivité entrante des noeuds suit une loi de puissance en fonction du nombre de noeuds du réseau.

le nombre réel de réseaux bayésiens possibles de mêmes degrés entrants (Favier *et al.*, 2009) (la distribution des degrés sur les noeuds suit une loi de puissance, avec un nombre d'arcs ≈ nombre de noeuds) et le nombre estimé pour différentes tailles de réseaux. Il convient donc d'utiliser une valeur $\gamma < 1$ ($\gamma = 0.7$) afin d'obtenir réellement un *a priori* uniforme sur les classes de réseaux bayésiens de même degré entrant.

Dans la suite de l'article nous noterons BIC_γ, BDeu_γ et fNML_γ les scores définis précédemment, enrichis du paramètre γ.

4. Modélisation

Nous présentons ici deux modélisations possibles dans le cadre des réseaux bayésiens prenant en compte deux sources d'information : les données de polymorphisme et les niveaux d'expression des gènes. Nous considérons de manière commune aux deux modèles la présence d'un seul polymorphisme fonctionnel par gène (i.e ayant un effet sur son expression ou son intensité de régulation) et supposons que ce polymorphisme représente une mutation ponctuelle de l'ADN appelée *Single Nucleotide Polymorphism* (SNP) identifiable à l'aide d'un marqueur situé à proximité immédiate du gène. On définit donc pour chaque gène $i \in [1, p]$, une variable d'expression de ce gène G_i et une variable de génotype M_i pouvant prendre deux valeurs (normal ou muté [1]).

1. Nous nous plaçons dans le cas d'individus diploïdes homozygotes ayant la même copie d'un gène sur les deux chromosomes homologues. C'est le cas par exemple pour des plantes obtenues par auto-fécondations successives (*Recombinant Inbred Lines (RIL)*).

4.1. Modèle non-fusionné

Ce modèle se compose de $2p$ variables distinctes comme présenté sur la figure 2. Bien que le nombre de variables du modèle soit le double du nombre de gènes, l'ensemble des réseaux possibles peut se réduire grâce à certaines connaissances biologiques. Les premières sont issues de la liaison génétique existante entre les marqueurs se succédant sur le chromosome selon le postulat de non interférence des *crossing − over* et du dogme de la biologie moléculaire ainsi illustrés sur la figure 3. D'autres connaissances relatives à la position de la mutation par rapport au gène peuvent également être utilisées comme nous le montrons sur la figure 4.

En tenant compte de ces informations le nombre de réseaux bayésiens possibles est $(p-1)^{m_{cod}} * N$ avec m_{cod} le nombre de gènes dont la mutation se situe en région codante et N le nombre de DAGs sur les p variables gènes, cette représentation permet en contre partie de représenter au plus près les phénomènes biologiques simulés.

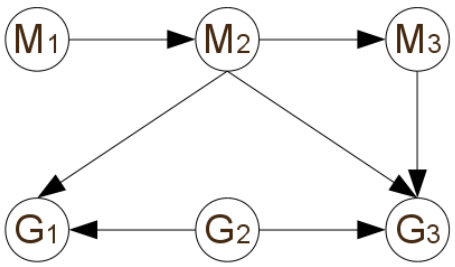


*Figure 2. Exemple d'un réseau de régulation où le gène 2 régule les gènes 1 et 3, $G_2 \rightarrow G_1$ & $G_2 \rightarrow G_3$. Les marqueurs génétiques sont liés suivant leur succession sur le chromosome tel que $M_i \rightarrow M_{i+1}$. Nous pouvons distinguer deux positions pour une mutation. Si celle-ci se situe dans la région promotrice de son gène telle que M_3, la mutation modifiera uniquement le niveau d'expression du gène, nous aurons ainsi $M_3 \rightarrow G_3$. Dans le cas où la mutation a lieu dans la région codante du gène (*exon*) comme pour M_1 et M_2 seule la force de régulation du gène sur les autres gènes sera modifiée et non son niveau d'expression. Dans ce cas pour chaque relation $G_i \rightarrow G_j$ avec dans notre exemple $i = 2$ et $j \in \{1, 3\}$ nous avons aussi $M_i \rightarrow G_j$.*

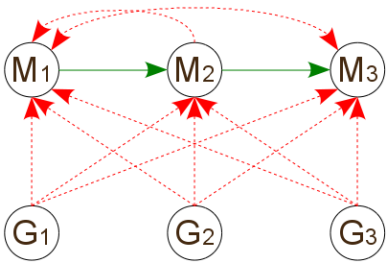


Figure 3. Connaissant la position des marqueurs sur le chromosome, ceux ci se modélisent sous la forme d'une chaine de Markov d'ordre 1 suivant leur ordre (flèches pleines) $M_i \rightarrow M_{i+1}$ tout en interdisant (flèches traitillées) les autres relations entre marqueurs $M_i \nrightarrow M_j \ \forall j \neq i+1$. Nous savons également qu'aucune expression de gène ne peut causer l'apparition d'une mutation $G_i \nrightarrow M_j \ \forall i, j$.

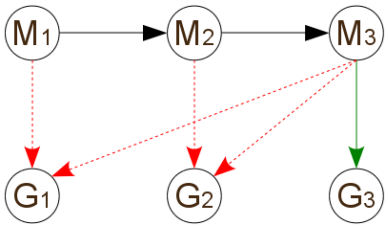


Figure 4. Pour chaque marqueur M_i situé dans la région promotrice de son gène (ici M_3) nous pouvons fixer les arcs $M_i \rightarrow G_i$ et interdire les arcs $M_i \nrightarrow G_j \ \forall j \neq i$. Dans le cas d'un marqueur M_i se situant dans la région codante (ici M_1 et M_2) nous interdisons uniquement les arcs $M_i \nrightarrow G_i$.

4.2. Modèle fusionné

Une modélisation alternative basée sur la fusion des deux variables associées à un gène (M_i et G_i) permet cette fois de ne pas augmenter le nombre de variables du modèle au prix de domaines plus grands. De cette fusion en résulte une nouvelle variable E_i dont le domaine sera le produit cartésien de celui de M_i et G_i comme décrit sur la figure 5

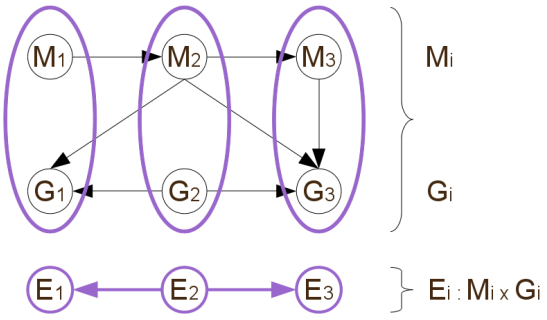


Figure 5. Modèle fusionné

De cette fusion, nous pouvons décomposer le calcul de la log-vraisemblance des données $\log(\mathbb{P}(\mathbf{D}|\mathcal{G}))$ comme suit :

$$\begin{aligned}
\log(\mathbb{P}(\mathbf{D}|\mathcal{G})) &= \log(\prod_{l=1}^{n}\prod_{i=1}^{p}\mathbb{P}(e_i^l|Pa(e_i^l))) \\
&= \log(\prod_{l=1}^{n}\prod_{i=1}^{p}\mathbb{P}(m_i^l, g_i^l|Pa(e_i^l))) \\
&= \log(\prod_{l=1}^{n}\prod_{i=1}^{p}\mathbb{P}(g_i^l|Pa(e_i^l), m_i^l)\mathbb{P}(m_i^l|Pa(e_i^l))) \\
&= \sum_{l=1}^{n}\sum_{i=1}^{p}\log(\mathbb{P}(g_i^l|Pa(e_i^l), m_i^l)) + \sum_{l=1}^{n}\sum_{i=1}^{p}\log(\mathbb{P}(m_i^l|Pa(e_i^l)))
\end{aligned}$$

avec $e_i^l = \{m_i^l, g_i^l\}$, l'observation du génotype et du niveau d'expression du gène i de l'individu l et $Pa(e_i^l)$, les observations des parents dans $\mathcal{G}$ du gène i de l'individu l. Le terme $T = \sum_{l=1}^{n}\sum_{i=1}^{p}\log(\mathbb{P}(m_i^l|Pa(e_i^l)))$ ne dépend que de la liaison génétique entre marqueurs [2]. Du fait de la fusion des variables, l'estimation des paramètres inclut donc une estimation de la liaison génétique conduisant à inférer des relations entre marqueurs que nous ne pourrions pas distinguer des relations entre gènes.

Afin d'ignorer cette dépendance induite par la liaison génétique dans la vraisemblance il convient de calculer $\mathbb{P}'(\mathbf{D}|\mathcal{G}) = \prod_{l=1}^{n}\prod_{i=1}^{p}\mathbb{P}(g_i^l|Pa(e_i^l), m_i^l)$. Ainsi nous pouvons adapter à cette modélisation fusionnée les 3 scores vus précédemment en redéfinissant $r_i' = r_{g_i}$ le domaine de définition de g_i, $q_i' = r_{m_i} * \prod_{e_j \in Pa(e_i)} r_{e_j}$ le nombre de configurations possibles pour les parents de e_i en y intégrant m_i et n'_{ijk}, le nombre d'occurrences de la configuration $(g_i = k, \{Pa(e_i), m_i\} = j)$.

Le principal avantage de cette modélisation se situe dans le nombre réduit de variables composant le graphe permettant d'accélérer la recherche dans l'espace des structures possibles. Cependant l'ajout d'une variable fusionnée en tant que parent équivaut à l'ajout simultané d'une variable d'expression et de génotype, la qualité de l'estimation des paramètres (plus nombreux) s'en trouve ainsi amoindrie. De plus la modélisation fusionnée ne permet pas de prendre en compte simplement les connaissances biologiques présentées en Section 4.1.

2. Les génotypes étant conditionnellement indépendants des niveaux d'expression, nous pouvons remplacer $\mathbb{P}(m_i^l|Pa(e_i^l))$ par $\mathbb{P}(m_i^l|Pa(m_i^l))$, en ne conservant dans $Pa(m_i^l)$ que les observations des génotypes.

5. Données expérimentales

5.1. Simulation des données génomiques et des données génétiques

L'observation de la structure des réseaux de régulation connus à ce jour a permis d'identifier certains éléments caractéristiques, notamment une structure dite *scale-free* s'articulant autour de quelques gènes *hubs* ayant un degré de connectivité élevé tandis qu'une grande majorité de leurs gènes satellites ne sont reliés qu'à un petit nombre de gènes (Barabasi, Oltvai, 2004). Afin de simuler des données réalistes, nous utilisons une collection de réseaux artificiels Web50 de $p = 50$ gènes (http://www.comp-sys-bio.org/AGN/) (Mendes *et al.*, 2003) ayant une structure de type *scale-free* (pouvant contenir des cycles) et un tirage aléatoire équiprobable pour l'orientation des arcs et le type de régulation (activation ou inhibition).

Nous avons ensuite simulé les génotypes d'une population de $n = 500$ individus en ségrégation selon un protocole *backcross*, grâce au logiciel CARTHAGENE (Givry *et al.*, 2005), en disposant 50 marqueurs SNP sur un seul chromosome de taille 10 Morgan (sans erreur de génotypage ni observation manquante). Pour chaque marqueur, nous choisissons de façon équiprobable la position de la mutation dans le gène (région promotrice ou région codante).

La simulation des données d'expression des gènes repose sur une fonction non linéaire de l'évolution de la quantité de gènes transcrits au cours du temps communément admise en biologie en première approximation (Liu *et al.*, 2008). Elle se traduit par une équation différentielle ordinaire prenant en compte les activateurs et les inhibiteurs d'un gène, ainsi que la dégradation naturelle des gènes transcrits au cours du temps :

$$\frac{dG_i}{dt} = V_i \prod_{G_j \in Inh(G_i)} Z_j\left(\frac{1}{1+G_j}\right) \prod_{G_k \in Act(G_i)} Z_k\left(1+\frac{G_k}{G_k+1}\right) - G_i + \theta_i G_i$$

avec V_i, le taux moyen de transcription du gène i (ce paramètre vaut β si la mutation du gène i a lieu en région promotrice et 1 sinon), $Inh(E_i)$ (resp. $Act(E_i)$), l'ensemble des gènes inhibiteurs (resp. activateurs) du gène i d'après le réseau artificiel choisi, Z_j (resp. Z_k), le taux moyen d'inhibition (resp. d'activation) de la transcription du gène i par le gène j (resp. k) (ces paramètres sont égaux à β si la mutation a lieu dans la région codante de ces gènes et 1 sinon) et θ_i, un bruit gaussien de moyenne 0 et d'écart-type 0.1. Le paramètre β représente l'impact qu'une mutation produit sur l'expression ou sur l'effet régulateur d'un gène, cet impact sera d'autant plus important que β diffère de 1. Nous présentons en Section 6 l'effet de ce paramètre sur l'apprentissage du réseau.

Nous utilisons le simulateur de réactions biochimiques COPASI (Hoops *et al.*, 2006) pour établir un état stationnaire des niveaux d'expression de l'ensemble des gènes en prenant comme cinétique l'équation définie ci-dessus appliquée à chaque gène et cela pour chaque individu. Nous ajoutons enfin aux résultats obtenus pour

chaque G_i une erreur expérimentale gaussienne centrée dont la variance est égale à 10% de la variance de G_i, à l'instar de (Liu *et al.*, 2008).

5.2. Présentation des données réelles d'Arabidopsis thaliana

Malgré la quantité importante d'informations biologiques disponibles sur Arabidopsis thaliana, les données de génétique génomique sur des individus apparentés sont rares. Afin de valider notre approche nous avons appliqué notre méthode aux données issues d'une population de 158 RILs (Simon *et al.*, 2008). Ces données contiennent pour chacun de ces individus les mesures d'expression obtenues grâce aux 32359 sondes d'une puce CATMA (Hilson *et al.*, 2004), les génotypes relevés sur 89 marqueurs de type SNP répartis uniformément le long des 5 chromosomes d'Arabidopsis, ainsi que les positions physiques et génétiques des sondes et des marqueurs sur le génome. Les données brutes et normalisées sont disponibles dans la base de données CATdb [3](Gagnot *et al.*, 2008).

Pour étudier nos résultats, nous les comparons à l'analyse eQTL qui permet de définir pour chaque gène, un ou plusieurs marqueurs génétiques influençant le niveau d'expression du gène. Cette analyse statistique établit donc des relations causales d'un marqueur vers un gène. Il est établi que la localisation de ces marqueurs causaux n'est pas exacte et que ces analyses ne s'attachent donc pas seulement au marqueur le plus explicatif localement mais délimitent des régions de confiance autour de lui. Si cette région inclut la position du gène nous parlerons d'eQTL de type *cis*, dans le cas contraire il s'agira d'un eQTL de type *trans*. La recherche d'eQTLs sur ces données a permis d'établir pour 5035 sondes la présence d'un ou plusieurs eQTLs [4].

5.3. Pré-traitements des données réelles

Contrairement aux données simulées, les données réelles présentent souvent des informations manquantes ou redondantes. Nous présentons dans cette section les traitements effectués sur les données d'expression et de génotype afin de pallier ces difficultés.

Nous complétons dans un premier temps les données d'expression grâce à un ensemble de prédicteurs. Pour chaque expression G_m présentant au moins une valeur manquante nous sélectionnons un ensemble de $N = 10$ expressions G_N parmi celles qui ne comportent pas de manquants et dont la covariance est la plus forte avec l'expression à compléter. Puis nous calculons la valeur manquante D^l_m pour l'individu l :

$$D^l_{m/N} = \mu_{G_m} + Cov_{G_m,G_N} Cov^{-1}_{G_N} (D^l_{G_N} - \mu_{G_N})$$

où μ_{G_m} est la moyenne de D_m estimée sur l'ensemble des individus observés pour D_m, Cov_{G_m,G_N} est la covariance de l'expression G_m avec chacune des expressions de G_N,

3. http://urgv.evry.inra.fr/CATdb/; Project: GNP07_RILKIT
4. http://qtlstore.versailles.inra.fr/

$Cov_{G_N}^{-1}$ est l'inverse de la covariance des expressions de G_N et $D_{G_N}^l - \mu_{G_N}$ l'écart pour chaque expression de G_N entre la valeur mesurée pour l'individu l et la moyenne sur l'ensemble des individus.

Une fois ces données complétées, nous conservons uniquement pour l'apprentissage du réseau, les gènes pour lesquels un ou plusieurs eQTLs significatifs ont été détectés, le but étant de définir les informations supplémentaires apportées par notre approche par rapport à l'analyse eQTL. Cette sélection réduit le nombre de sondes de 32359 à 4176.

Les données de génotype comportant également certaines données manquantes, nous utilisons le package R *qtl* (Broman *et al.*, 2003) afin d'inférer, sachant les génotypes observés, à la fois les génotypes manquants aux marqueurs ainsi que les génotypes de $pseudo-marqueurs$ espacés d'1 cM, afin de couvrir plus finement la carte génétique. Nous discrétisons ensuite ces données de génotype en 3 classes, en utilisant les 1^{er} et 3^{me} quartiles. Nous obtenons au total 590 marqueurs incluant les 89 vrais marqueurs initiaux complétés.

5.4. Discrétisation des données d'expression de gènes

Le choix du formalisme des réseaux bayésiens discrets impose une étape de discrétisation des données d'expression. Nous avons choisi d'utiliser une méthode fondée sur la recherche de modèles de mélange gaussien couplée à une recherche par *k-means*, afin d'obtenir une discrétisation avec un faible nombre de classes (au plus 4) qui s'adapte à la forme de la distribution des niveaux d'expression de chaque gène. Les différents algorithmes et fonctions utilisés pour la discrétisation sont disponibles sous Matlab r2009b accompagné de la *Statistics Toolbox*.

Méthode des *k-means* modifiée. Une méthode courante en classification non supervisée est la méthode des *k-means* (Stuart, 1982). Elle partitionne un nuage de points en un nombre prédéfini de classes, en minimisant la variance intra-classe et en maximisant celle inter-classes. Nous recherchons une discrétisation en 3 classes (sous-, normal et sur-exprimés) en évitant d'avoir certaines classes avec trop peu de représentants. Pour cela, nous avons adapté la méthode des *k-means*, à l'instar de (Zhu *et al.*, 2007), en appliquant l'algorithme *k-means* à 5 classes, puis en étudiant les regroupements possibles afin d'assurer une population d'au minimum 5% de sous- et sur-exprimés sans dépasser pour autant 30% de la population totale, afin de garder le caractère anormal de ces états. Dans le cas du premier gène de la figure 6, cette méthode obtient une discrétisation naturelle. C'est le cas pour des gènes ayant une distribution approximativement gaussienne. Lorsque la distribution a plusieurs modes comme dans le cas des 2ème et 3ème gènes de la figure 6, nous appliquons une méthode de modèle de mélange gaussien.

Modèle de mélange gaussien. Nous effectuons dans un premier temps un lissage de l'histogramme grâce à la fonction *fastsmooth*. Puis les modes sont détectés afin de calculer le nombre k de gaussiennes à rechercher. Nous appliquons dans un second temps un algorithme *Expectation Maximization* de modèle de mélange gaussien identifiant

les paramètres des k gaussiennes (McLachlan, 1982) (voir les exemples en figure 7). Au final, le nombre de classes k est limité à 4. Dans le cas d'un mode unique, c'est la méthode modifiée des *k-means* qui est utilisée.

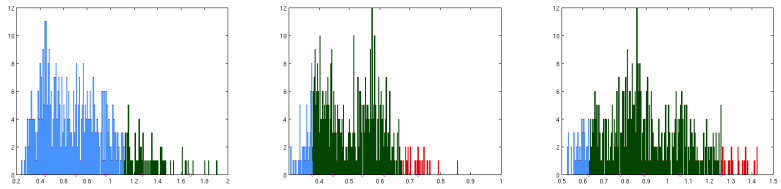


Figure 6. Discrétisation par la méthode modifiée des k-means *appliquée à l'expression respectivement d'un gène du réseau* Web50_001 *et de 2 gènes du réseau* Web50_008 *(*de gauche à droite*, G_3, G_{33} et G_{13}). Chaque figure donne l'histogramme du nombre d'individus en fonction de l'expression du gène. Les différentes couleurs représentent les 2 ou 3 classes trouvées par la méthode.*

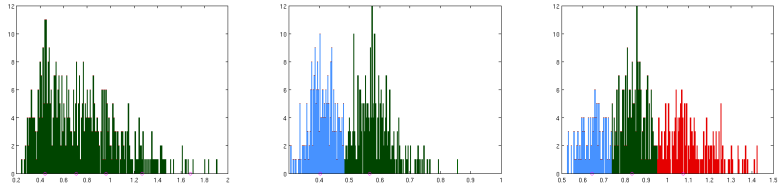


Figure 7. Discrétisation par modèle de mélange gaussien pour les expressions G_3, G_{33} et G_{13}.

6. Résultats sur données simulées

Dans le but d'évaluer les différents concepts abordés dans cet article nous décomposons les résultats obtenus en 4 parties distinctes. Dans un premier temps nous comparons les deux modèles décrits en Section 4 sur la base des critères BIC et fNML, puis nous testons l'intérêt des critères étendus et l'impact des informations biologiques apportées dans le modèle non fusionné. Nous terminons par une comparaison des méthodes existantes pour l'inférence de réseaux de régulations de gènes.

Afin d'inférer les réseaux bayésiens, nous utilisons l'algorithme *Greedy Search* (GS) dont une implémentation est proposée dans le logiciel BANJO (Hartemink, 2005) auquel nous avons adapté le score BDeu à sa version étendue (nous fixons le paramètre $\alpha = 1$) ainsi qu'implémenté les scores BIC et fNML. L'ensemble de ce travail est mis à disposition au travers d'une applet Java [5] intégrant le logiciel BANJO

5. http://carlit.toulouse.inra.fr/genebayesnet/

étendu ainsi que des fonctionnalités de visualisation et de comparaison des réseaux reconstruits dans le cadre des données de génétique génomique (figure 8). Pour des raisons d'espace mémoire nous limitons le nombre maximum de parents à 9 lors de la recherche. Nous initialisons l'algorithme GS avec un graphe vide.

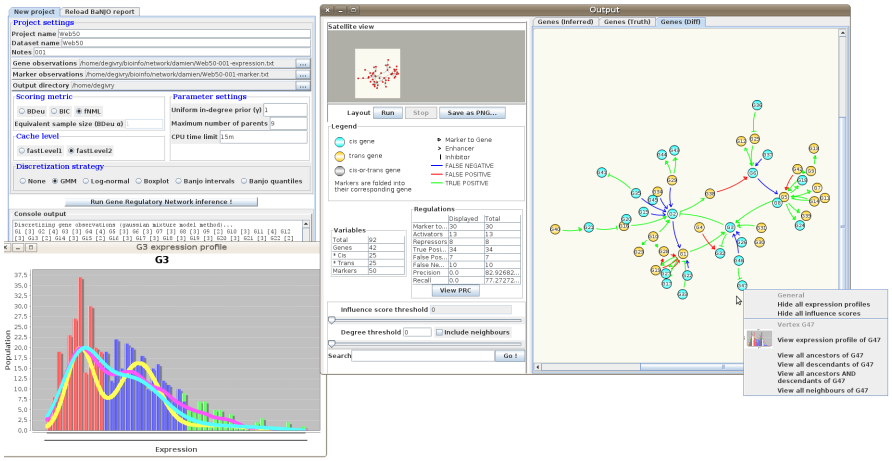


Figure 8. Vue d'écran de l'applet GeneBayesNet intégrant le logiciel BANJO et les scores étendus. Comparaison du réseau reconstruit avec le vrai réseau Web50_001 *pour 500 individus avec le score* $fNML_1$.

A l'exception du modèle fusionné, nous effectuons pour l'ensemble des méthodes testées un post-traitement sur le réseau appris comprenant $2p$ variables car seul nous intéresse le réseau appris des p gènes. Pour cela les relations entre gènes sont conservées et les relation de type $M_i \rightarrow G_j \forall j \neq i$ sont projetées sur $G_i \rightarrow G_j$.

La qualité des réseaux reconstruits est évaluée en terme de sensibilité ($\frac{TP}{TP+FN}$) et de précision ($\frac{TP}{TP+FP}$), avec TP, le nombre d'arcs présents à la fois dans le réseau reconstruit et dans le vrai réseau, FP, le nombre d'arcs prédits à tort dans le réseau reconstruit, et FN, le nombre d'arcs du vrai réseau absents du réseau reconstruit. Vu l'incapacité des différentes méthodes présentées en Section 2 d'orienter de façon complète les relations apprises, nous ne prenons pas en compte l'orientation des arcs dans nos mesures.

Les résultats présentés sont moyennés sur 50 réseaux Web50 (décrits en Section 5.1).

6.1. Comparaison des deux modèles

La qualité des réseaux appris grâce aux deux modélisations nous apparait fortement inégale sur la figure 9. Le modèle non fusionné obtient ainsi de meilleurs résultats pour différentes tailles de population sur les deux métriques hormis pour le

critère $fNML$ avec peu d'individus où le modèle fusionné garde une précision légèrement supérieure mais une sensibilité bien inférieure. Pour 50 et 100 individus, les réseaux sélectionnés par le modèle fusionné avec le critère BIC ne comportent que très peu d'arcs (2 et 5 en moyenne) expliquant ainsi une sensibilité proche de 0. Ce résultat confirme par ailleurs la pénalité élevée que ce critère associe à la complexité du réseau appris. Cette pénalité est d'autant plus forte dans le modèle fusionné, en effet dans ce dernier la variable E_i représente l'ensemble des deux variables M_i et G_i, ainsi là où le modèle fusionné sélectionne E_i en tant que régulateur d'un gène, le modèle non fusionné autorise à ne retenir qu'une seule des deux variables si celle-ci se révèle suffisament explicative. Ce dernier modèle permet ainsi de sélectionner pour une même dimension du réseau bayésien un nombre plus important de régulateurs et d'améliorer la qualité des réseaux appris.

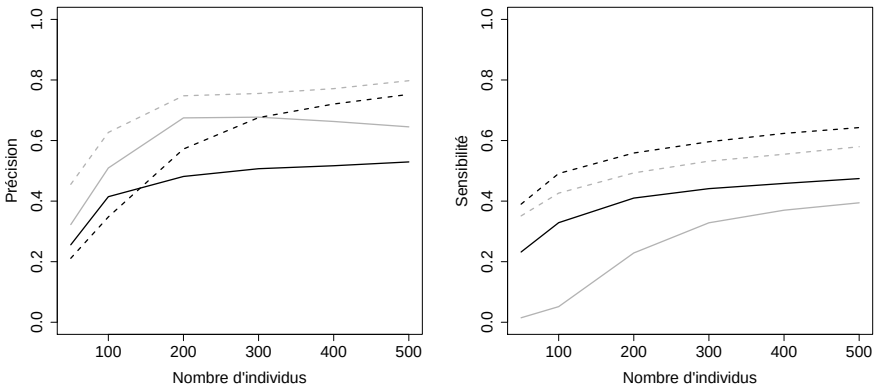


Figure 9. ***Comparaison des modèles.*** *Evolution de la précision (à gauche) et de la sensibilité (à droite) des réseaux appris pour les critères BIC_0 (gris) et $fNML_0$ (noir), en fonction du nombre d'individus. Les modèles fusionné (ligne continue) et non fusionné (ligne traitillée) sont représentés pour chaque critère.*

En nous basant sur ces premières observations, nous présentons par la suite les résultats uniquement sur le modèle non fusionné.

6.2. Impact des critères étendus

Nous pouvons voir sur la figure 10 un comportement similaire des scores BIC et fNML (par souci de clarté nous ne présentons pas les résultats du score BDeu similaire au BIC, hormis pour 50 individus où BDeu se révèle moins performant). Dans le cas d'un petit nombre d'individus la hausse du paramètre γ permet d'améliorer de façon significative la précision des réseaux appris tandis que leur sensibilité se trouve en léger retrait. Ce comportement s'explique du fait qu'une valeur élevée de γ favorise les structures à faible connectivité. Ce phénomène est d'autant plus marqué lorsque

le nombre d'individus décroit, le graphe final comporte alors de nombreuses relations incertaines dont une grande partie représente de fausses relations, augmenter γ permet de ne pas retenir ces arcs peu significatifs améliorant ainsi la précision du réseau final. La perte en sensibilité montre cependant que de vraies régulations figurent parmi ces relation incertaines. Malgré cela l'étude du ratio précision/sensibilité montre l'intérêt d'utiliser une valeur de gamma supérieure à 0. La valeur de ce ratio tend par ailleurs à se réduire lorsque γ se rapproche de 1, impliquant l'existence d'une valeur de γ optimale. Nous vérifions également l'équivalence asymptotique des deux critères aux vus des résultats convergents pour un nombre croissant d'individus (Silander *et al.*, 2010).

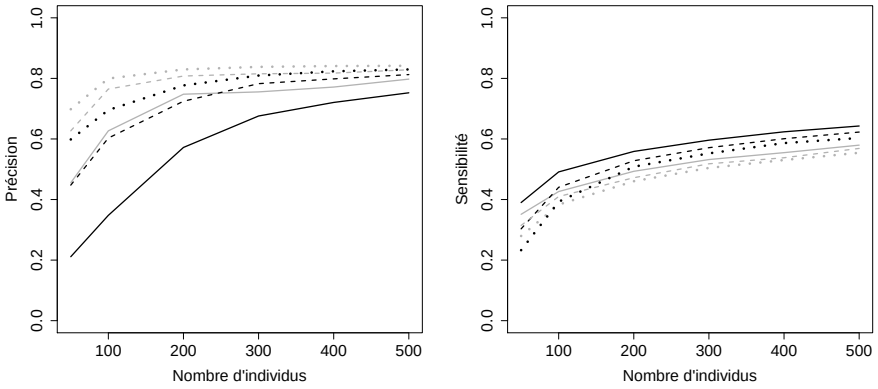


Figure 10. ***Comparaison des critères étendus.*** *Evolution de la précision (à gauche) et de la sensibilité (à droite) des réseaux appris avec les critères étendus BIC (gris) et $fNML$ (noir), en fonction du nombre d'individus. Pour chaqu'un d'eux, 3 valeurs de gamma sont représentés : $\gamma = 0$ (ligne continue), $\gamma = 0.5$ (ligne traitillée) et $\gamma = 1$ (ligne pointillée).*

6.3. Informations biologiques supplémentaires dans le modèle non-fusionné

Grâce à certaines connaissances biologiques nous pouvons restreindre le nombre de structures possibles dans le cas du modèle non-fusionné comme nous l'avons montré en Section 4.1. Nous avons appliqué la même liste de contraintes sur les relations apprises par la méthode SEM Lasso. Dans ce cadre imposer une relation de régulation consiste à effectuer la régression linéaire correspondant à cette relation et de poursuivre la méthode standard en considérant le résidu de cette première régression tandis qu'interdire une relation s'effectue en réduisant la liste des régresseurs possibles. Nous présentons sur la figure 11 l'impact de ces connaissances sur la qualité des réseaux appris. Bien que ces restrictions ne donnent aucune information explicite sur les relations entre gènes, celles-ci orientent favorablement la recherche et améliorent la précision des 3 approches. Cette progression est d'autant plus significative que le

nombre d'individus décroit, seule la méthode SEM Lasso se montre moins sensible avec 50 individus dû à ses mauvaises performances dans le cas de petites populations. De façon similaire à la section précédente le critère BIC et la méthode Lasso affichent une baisse de leur sensibilité, celle-ci reste toutefois limitée pour le critère BIC. Seul le critère fNML réagit positivement à ces restrictions avec une amélioration de la qualité des réseaux sur les deux métriques, faisant de celui-ci un critère de choix pour incorporer des connaissances expertes supplémentaires.

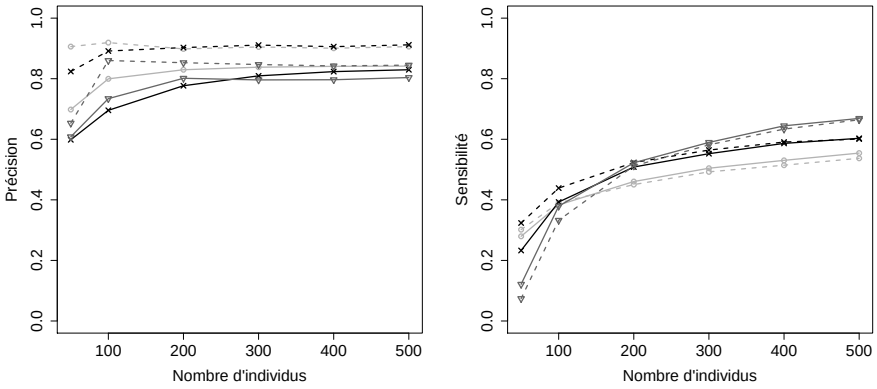


Figure 11. ***Impact des connaissances biologiques.*** *Evolution de la précision (à gauche) et de la sensibilité (à droite) des réseaux appris avec les critères BIC_1 (rond gris clair) et $fNML_1$ (croix noire) ainsi que l'approche SEM Lasso (triangle gris), en fonction du nombre d'individus. Pour chaque approche, 2 tracés présentent l'apprentissage avec (ligne traitillée) et sans (ligne continue) ces restrictions.*

6.4. Comparaison globale des méthodes d'inférence de réseaux

Nous comparons dans les prochaines parties le modèle non fusionné avec les différentes approches présentées en Section 2, nous étudions l'impact du nombre d'individus ainsi que la capacité des méthodes à gérer des données générées sous diverses conditions. Les logiciels ou packages utilisés sont disponibles gratuitement et résumés dans le tableau 1, accompagnés des choix effectués concernant divers paramètres dont la valeur a été définie de manière à obtenir un nombre d'arcs proche de celui du réseau réel. Nous utilisons pour le logiciel GGMSelect les familles "$C01$" et "LA", la famille "EW" ne donnant aucun résultat sur nos jeux de données. Compte tenu des précédents résultats, nous utilisons pour ce comparatif les critères BIC_1, $BDeu_1$, $fNML_1$ et l'approche SEM Lasso en exploitant l'information de position des polymorphismes afin de restreindre les structures apprises possibles. Dans notre comparatif seuls notre modèle non fusionné, l'approche SEM Lasso et le logiciel SCT sont aptes à distinguer les données d'expression et de polymorphisme, les autres méthodes ayant été développées dans le but de traiter uniquement les niveaux d'expression, nous leur fournissons

donc de manière indissociée les deux types de données.

Tableau 1. ***Paramètres des méthodes testées***

Logiciels	***Description***	***Paramètres***
BANJO(v2.2)	Réseaux bayésiens	$\alpha_{BD_{eu}} = 1; \gamma = 1$
SCT(v0.1)	Réseaux bayésiens	$poids = \{1, 2\}; iterations = 10M$
ARACNE	Information mutuelle	$seuil = 0.15$
CLR(v1.2)	Information mutuelle	$seuil = 4$
ParCorA	Corrélation de Spearman	$pseuil = 0.01;\ 1er\ ordre$
SIMoNe(v1.0)	GGM	$nombredemodules = 2$
GGMselect(v0.1)	GGM	$familles\ C01\ \&\ LA$
GeneNet(v1.2.4)	GGM	$seuil = 0.95$
SEM Lasso	Régression Lasso	$\alpha_{meinshausen} = 0.1$

6.5. Nombre d'individus

Le nombre d'individus disponibles pour l'apprentissage de structure reste bien souvent l'un des principaux facteurs limitants. Nous comparons ici le cas favorable d'une population de 500 individus à une situation plus réaliste où seuls 50 individus sont disponibles. En comparant les 2 colonnes de la figure 12 nous pouvons noter de manière générale une grande disparité dans les performances des différentes approches, de même aucune d'entre elles ne semble clairement dominer ce comparatif. Deux groupes de méthodes émergent malgré tout de façon plus marquée lorsque le nombre d'individus augmente. La faible sensibilité des différentes approches avec une population réduite montre une incapacité commune à retrouver un nombre élevé de régulations. Ces difficultés proviennent notamment des structures de gènes en *hub* où un gène est fortement connecté aux autres gènes et dont son expression se trouve dominée par un faible nombre de régulateurs occultant ainsi les autres régulations.

ARACNE et CLR obtiennent pour les différentes configurations des résultats similaires, CLR domine tout de même ce duel avec 500 individus face à ARACNE peu sensible à la taille de la population. Ces performances restent cependant éloignées de celle de ParCorA qui se montre efficace et semble tirer un avantage certain des mesures de corrélations conditionnelles.

SIMoNe obtient les plus mauvais résultats du comparatif dû notamment à une faible précision, ce phénomène peut probablement s'expliquer par le réseau modulaire autour duquel est conçu cette méthode, caractéristique ne correspondant pas à nos réseaux. En effet les noeuds des réseaux sont divisés en deux classes (niveaux d'expression et polymorphismes) mais en raison du nombre de régulations entre gènes dans nos données, les densités intra et inter classes sont très proches et ne permettent pas de distinguer d'éventuels modules (nous avons par ailleurs testés un nombre de modules variant de 2 à 6 sans aucune amélioration significative). La qualité des graphes

résultants de GeneNet, GGMselect et SEM Lasso se dégrade rapidement lorsque la population diminue. Parmi ces 3 méthodes l'approche SEM Lasso domine pour 500 individus tandis que GGMselect tire partie d'une population plus petite, dans les différents tests ces deux approches gardent une précision élevée.

Notre approche apparait comme une des méthodes les plus robustes de même que SCT et ParCorA, si ce dernier obtient d'ailleurs de meilleures performances pour 500 individus, la situation s'inverse pour 50 individus. Au final les trois scores étudiés dans cet article restent proches au niveau de la qualité des réseaux obtenus, $BDeu_1$ domine légèrement les situations avec peu d'individus tandis que $fNML_1$ prend l'avantage lorsque la population croît.

6.6. Impact de la distance génétique entre marqueurs

La position des marqueurs le long du chromosome est un facteur influençant la difficulté d'apprentissage du réseau. En effet la probabilité que l'état d'un marqueur (muté ou non) soit différent de celui qui le précède ou le succède sur le chromosome est directement proportionnel à la distance génétique les séparant, ainsi deux marqueurs espacés d'1cM auront des états contraires sur seulement 1% de la population ce qui les rend potentiellement indissociables dans une population de 50 individus. Ces faibles écarts tendent à perturber l'apprentissage du réseau en augmentant les ressemblances entre variables. Nous comparons les 2 premières configurations de la figure 12 afin d'observer ce phénomène, un espacement aléatoire produit notamment des distances entre marqueurs extrêmement faibles comparé à l'espacement uniforme représentant le cas idéal.

On observe naturellement une diminution des performances pour toutes les méthodes lorsque l'espacement est aléatoire cependant cette baisse n'est pas identique suivant les approches. Alors que la dégradation des performances des méthodes par réseaux bayésiens et par corrélations reste mesurée, les approches basées sur le modèle linéaire montrent de fortes irrégularités. GeneNet perd sensiblement en précision avec peu d'individus alors que SIMoNe se dégrade pour une grande population, seul SEM Lasso et GGMselect restent faiblement perturbés.

6.7. Impact des mutations

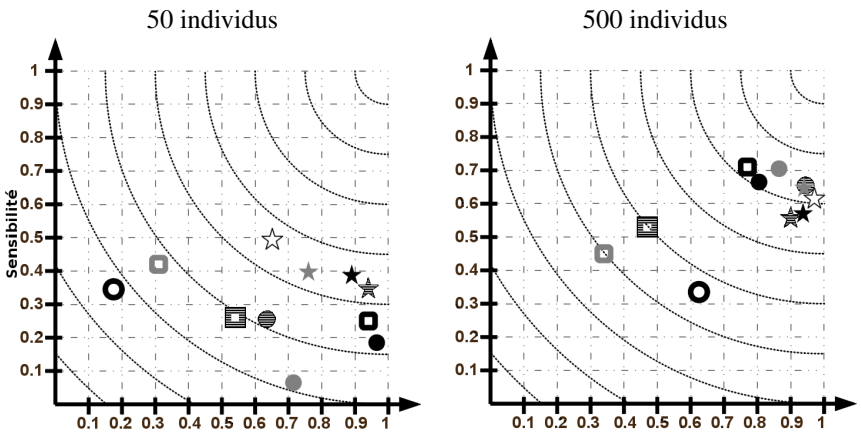
Dans le cas de données simulées, il est souvent difficile de quantifier le réalisme de celles-ci, l'équation régissant l'expression d'un gène présentée en Section 5.1 n'échappe pas à cet aspect. Afin de couvrir un panel de données possibles sans remettre en cause la forme de l'équation, il convient de faire varier son unique paramètre β. Les configurations (b) et (c) de la figure 12 présentent deux valeurs possibles de ce paramètre, comme mentionné précédemment une valeur se rapprochant de 1 implique une réduction de l'effet des mutations sur l'expression des gènes.

Au niveau des différentes méthodes, on observe une stabilité générale des méthodes de corrélations ainsi que de SCT tandis que notre approche se dégrade de fa-

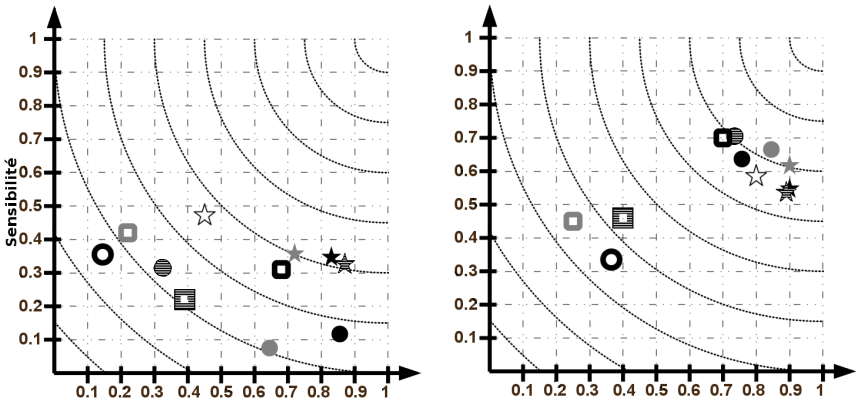
çon plus marquée tout en restant compétitive pour 50 individus. Les méthodes SEM Lasso et GGMselect se dégradent pour 500 individus (fortement pour ce dernier) tandis que pour 50 individus les performances augmentent légèrement. Seul SIMoNe se détériore pour les deux tailles d'échantillon. Les résultats de GeneNet sont les plus surprenants avec une précision divisée par deux pour 500 individus, alors que cette même baisse pour 50 individus se trouve compensée par une hausse de la sensibilité gardant le même ratio précision/sensibilité.

En réalité pour une valeur de β proche de 1, il est certes difficile d'identifier des régulations provenant des marqueurs, mais pour une valeur trop éloignée de 1 (ici 0.25) le niveau d'expression des gènes se trouve dominé par quelques marqueurs masquant ainsi les autres régulations (similaire aux difficultés liées aux *hubs*). Ce phénomène s'amplifie d'autant plus pour les gènes dont les marqueurs sont situés en région promotrice.

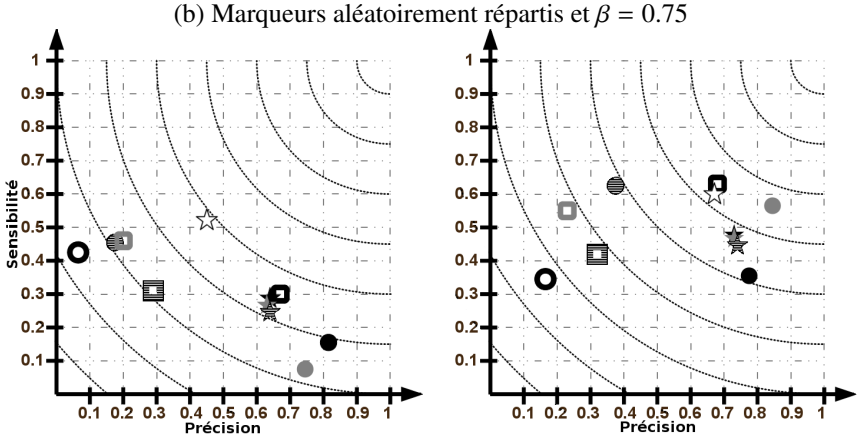
Malgré ces résultats hétérogènes, certains points ressortent de ces comparatifs. Les différents scores utilisés pour sélectionner les réseaux bayésiens avec l'algorithme GS obtiennent des résultats similaires et se montrent particulièrement stables dans les différentes configurations testées de même que SCT. Les méthodes de corrélations apparaisent également robustes, seules les méthodes SEM Lasso et GGMselect montrent un caractère semblable parmi les modèles linéaires tandis que les autres méthodes se révèlent particulièrement instables. Parmi toutes ces méthodes, ParCorA et les approches par réseaux bayésiens (GS et SCT) obtiennent au final les meilleurs résultats cumulés de cette comparaison sur données simulées. Cependant SCT nécessite le réglage de 8 paramètres différents dont la pondération de l'*a priori* à laquelle la méthode se révèle très sensible. Bien que nous utilisions ici une pondération adaptée connaissant les vrais réseaux, le réglage des différents paramètres lors de l'analyse de données réelles peut s'avérer difficile.



(a) Marqueurs uniformément répartis et $\beta = 0.75$



(b) Marqueurs aléatoirement répartis et $\beta = 0.75$



(c) Marqueurs aléatoirement répartis et $\beta = 0.25$

Figure 12. ***Précision (axe horizontal) et sensibilité (axe vertical) pour 50 (colonne de gauche) et 500 individus (colonne de droite) dans 3 configurations. (a)****Les marqueurs sont espacés uniformément sur le chromosome avec un paramètre* β *quantifiant l'impact de la mutation égal à 0.75.* ***(b)****Les marqueurs sont espacés aléatoirement avec le même* β*.* ***(c)****Les marqueurs sont espacés aléatoirement et les mutations ont un impact plus fort (*$\beta = 0.25$*). Pour chaque figure les méthodes sont classées en 3 catégories:* ***1-Réseaux de corrélations*** *ARACNE (carré gris), CLR (carré gris zébré), ParCorA (carré noir);* ***2-Modèles linéaires*** *SIMoNe (cercle noir vide), GeneNet (cercle gris zébré), GGMSelect (cercle noir plein), SEM Lasso (cercle gris);* ***3-Réseaux bayésiens*** *SCT(étoile noire vide) et BANJO avec* BIC_1 *(étoile grise zébrée),* $BDeu_1$ *(étoile noire), et* $fNML_1$ *(étoile grise).*

7. Résultats sur données réelles

7.1. Adaptations de l'algorithme Greedy Search

L'exploration du voisinage faite par l'algorithme étant quadratique dans le nombre de variables, il est nécessaire d'effectuer en amont de notre recherche heuristique une pré-sélection des parents potentiels pour chaque variable expression (Goldenberg, Moore, 2004). Pour chacune de ces variables nous comparons la situation sans parent à l'ajout d'un parent en terme de score local $BDeu_1$ et nous conservons uniquement les parents qui améliorent ce score. Cette technique similaire aux méthodes dites *hybrides* (Tsamardinos *et al.*, 2006) permet un passage à l'échelle dans le cas de données réelles en ne considérant qu'un sous-ensemble de parents possibles pour chaque variable, réduisant ainsi l'espace de recherche.

La non correspondance d'un marqueur par gène empêche également d'appliquer toutes les restrictions de la recherche telles que présentées en section 4.1. Seules les restrictions illustrées sur la figure 3 sont conservées. Par ailleurs, pour des raisons d'espace mémoire et de temps de calcul, nous limitons le nombre de parents maximum à 4.

7.2. Analyse du réseau Bayésien reconstruit

Notre méthode apprend sur ces données un réseau peu dense constitué de 4766 variables et 6137 arcs soit un rapport nombre d'arcs/ nombre de variables d'environ $1,3$. Le réseau est constitué d'une composante connexe majeure regroupant 4004 variables, les autres variables forment des composantes de dimension inférieure à 5. Seules 284 relations sont de type $M \rightarrow E$ tandis que 5853 relations représentent des régulations entre variables d'expression.

L'analyse des degrés entrants et sortants du réseau appris (Tableau 2), montre que 278 sondes possèdent 3 parents ou plus (maximum de 4 parents atteint pour 41 sondes)

tandis que certaines sondes diffusent fortement allant jusqu'à des degrés sortant de 55.

Tableau 2. Nombre de variables d'expression et de marqueur suivant leurs degrés entrant et sortant.

Degré	0	1	2	3	4	5	6	7	8	9	10+
Degré entrant des gènes	263	2008	1627	237	41	-	-	-	-	-	-
Degré sortant des gènes	2164	844	465	260	118	98	58	46	26	29	68
Degré sortant des marqueurs	457	67	29	15	9	7	2	1	3	-	-

7.3. Comparaison avec l'analyse eQTL

La figure 13 illustre deux configurations où le réseau bayésien propose un modèle explicatif plus précis sur les régulations entre gènes. Dans ces deux situations un même eQTL a été détecté pour les gènes G_1 et G_2 suggérant ainsi l'existence d'une relation entre ces deux gènes, ce qui est confirmé par la structure du réseau bayésien appris.

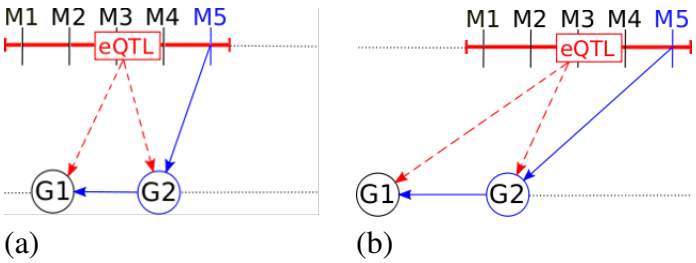


Figure 13. Figures représentant sur deux niveaux des variables marqueur et expression positionnées par rapport à la séquence génomique. Chaque eQTL détecté pour l'un des deux gènes G_1 et G_2 est représenté en traitillé rouge, l'incertitude sur la position exacte d'un eQTL amène à définir une région de confiance pouvant s'étendre sur plusieurs marqueurs. Tandis qu'en bleu continu figure les arcs appris dans le réseau bayésien. (a)Situation où le réseau bayésien contient un chemin orienté composé de 2 gènes débutant par un marqueur M_5 situé dans la région de confiance de l'eQTL détecté en cis pour G_1 et G_2. (b)Situation où le réseau bayésien contient un chemin orienté composé de 2 gènes débutant par un marqueur M_5 situé dans la région de confiance de l'eQTL détecté en cis pour G_2 et en trans pour G_1.

Nous avons étudié pour chaque arc détecté par l'analyse eQTL, qu'elle était la configuration proposée par le réseau bayésien. Nous parlerons d'arc expliqué directement par le réseau bayésien lorsque celui-ci est présent dans le réseau et indirectement

si cet arc est représenté par un chemin orienté dans le réseau. Par exemple, l'arc détecté pour G_2 sur la figure 13 est expliqué directement tandis que l'explication de l'arc détecté pour G_1 est indirecte.

Pour notre comparaison, seuls les 1269 eQTLs les plus significatifs (FDR de 1%) ont été retenus.

	arcs détectés par l'analyse eQTL	arcs expliqués par le réseau bayésien	(direct/indirect)
cis-eQTL	938	637	(202/435)
trans-eQTL	331	229	(68/161)

Figure 14. Comparaison du nombre d'arcs détectés par l'analyse eQTL et expliqués par le réseau bayésien en fonction du type d'effet de l'eQTL.

Nous pouvons voir sur le tableau 14 que le réseau bayésien permet d'expliquer respectivement 67% et 69% des eQTLs détectés de type *cis* et *trans*, parmi lesquels plus de 68% sont expliqués de manière indirecte dans le réseau. Le réseau bayésien propose un nombre plus important d'explications impliquant des régulations entre gènes que d'interactions directes marqueurs-gènes. Cette observation justifie pleinement l'intérêt d'un modèle non fusionné des gènes et des marqueurs afin de saisir au mieux la complexité des interactions.

Afin d'étudier les plus fortes relations de notre réseau, nous classons les arcs appris en fonction de l'amélioration produite en terme de score $BDeu_0$ par leur ajout lors de l'algorithme GS puis nous nous intéressons aux seules relations entre gènes situés sur des chromosomes différents pour éviter tout biais d'interprétation dû à leur proximité. Enfin nous effectuons une recherche de *gene ontology* commune pour les 10 premières relations d'après la base de données *TAIR* (Swarbreck *et al.*, 2008). Cette analyse préliminaire fait ressortir clairement une fonction commune des gènes $AT4G25050$ et $AT1G15820$ reliés dans notre réseau, connus pour participer tous deux aux processus biologiques de réponse à la lumière chez Arabidopsis (Anderssona *et al.*, 2001 ; Bonaventure, Ohlrogge, 2002). Cette relation s'inscrit dans une configuration similaire à la figure 13b où un même eQTL est détecté en tant que *cis* pour le gène $AT1G15820$ et en *trans* pour le gène $AT4G25050$.

8. Conclusion

Nous avons proposé dans cet article d'intégrer aux scores BIC, BD_{eu} et $fNML$ un *a priori* sur les structures de manière à privilégier des graphes peu denses. Nous avons validé l'intérêt de ces scores étendus pour l'apprentissage de réseaux bayésiens statiques sur des données simulées. Nous avons également développé deux modèles permettant de prendre en compte des données biologiques variées (données d'expression et de polymorphisme) dont un modèle dit *non fusionné* pouvant de plus intégrer des connaissances biologiques supplémentaires facilitant l'apprentissage du réseau.

Le modèle fusionné s'est révélé moins performant que son homologue non fusionné mais permet en contre-partie de réduire l'espace de recherche. Nous avons par la suite comparé le modèle non fusionné à diverses méthodes disponibles d'apprentissage de réseau de régulation à partir de données d'expressions et obtenus les meilleurs résultats pour notre approche par réseaux bayésiens avec la plus petite population simulée et la plus réaliste compte tenu des possibilités expérimentales actuelles. Finalement nous avons utilisé ce modèle non fusionné sur des données réelles d'Arabidopsis thaliana et avons comparé notre réseau appris avec une analyse eQTL.

Il est important de noter que les diverses méthodes testées ici, autres que les réseaux bayésiens (dont le logiciel SCT) et SEM Lasso, ne sont pas conçues pour traiter différemment les données de polymorphisme et considèrent donc les variables marqueurs comme des variables d'expression. La majorité de ces méthodes nécessitent un paramétrage qui se révèle en pratique bien difficile à effectuer pouvant faire varier la qualité des résultats obtenus de façon significative. En conséquence, afin de ne désavantager aucune de ces méthodes, différents paramétrages ont été testés et seuls les meilleurs résultats obtenus figurent dans nos comparaisons. De plus nous sommes conscients que le choix de réseaux à faible connectivité est un facteur non négligeable dans les performances des différentes méthodes et notamment les approches par réseaux bayésiens qui sont davantage pénalisées, du à l'acyclicité des graphes reconstruits. Cependant une récente étude sur des réseaux simulés de grande taille (1000 gènes) avec des connectivités variées a montré l'interêt de l'approche par réseaux bayésiens (Vignes *et al.*, 2011).

De nombreuses perspectives sont ouvertes suite à cette étude, notamment celle de s'intéresser aux relations causales dans ces réseaux afin d'obtenir une orientation des régulations détectées (partial DAG), ainsi que l'étude d'autres méthodes d'optimisation à base de score que l'algorithme *Greedy Search*. Au vu de la faible sensibilité des différentes approches, un travail s'attachant à mieux retrouver des relations difficiles autour des *hubs* doit également être mené sans pour autant détériorer la précision des méthodes.

Bibliographie

Anderssona J., Waltersb R., Hortonb P., Janssona S. (2001). Antisense inhibition of the photosynthetic antenna proteins cp29 and cp26: Implications for the mechanism of protective energy dissipation. *Plant Cell*, vol. 13, p. 1193-1204.

Bansal M., Belcastro V., Ambesi-Impiombato A., Bernardo D. di. (2007). How to infer gene networks from expression profiles. *Mol Syst Biol*, vol. 3, n° 78.

Barabasi A., Oltvai Z. (2004, FEB). Network biology: Understanding the cell's functional organization. *NATURE REVIEWS GENETICS*, vol. 5, n° 2, p. 101-115.

Bonaventure G., Ohlrogge J. (2002). Differential regulation of mrna levels of acyl carrier protein isoforms in arabidopsis. *Plant Physiology*, vol. 128, p. 223-235.

Broman K., Wu H., Sen S., Churchill G. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, vol. 19, p. 889-890.

Chen J., Chen Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, vol. 95, n° 3, p. 759-711.

Chipman K., Singh A. (2011). Using stochastic causal trees to augment bayesian networks for modeling eqtl datasets. *BMC Bioinformatics*, vol. 12, n° 1.

Chiquet J., Smith A., Grasseau G., Matias C., Ambroise C. (2009). Simone: Statistical inference for modular networks. *Bioinformatics*, vol. 25, n° 3, p. 417-418.

Chu J., Weiss S., Carey V., Raby B. (2009). A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. *BMC Systems Biology*, vol. 3, n° 55.

Cooper G., Hersovits E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, vol. 9, p. 309–347.

Faith J. J., Hayete B., Thaden J. T., Mogno I., Wierzbowski J., Cottarel G. *et al. (2007). Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol*, vol. 5.*

Favier A., De Givry S., Jégou P. (2009). Exploiting problem structure for solution counting. In Proceedings of the 15th international conference on principles and practice of constraint programming*, p. 335–343.*

Friedman N., Linial M., Nachman I., Peer D. (2000). Using bayesian networks to analyse expression data. Journal of computational biology*, vol. 7, n° 3, p. 601-620.*

Fuente A. de la, Bing N., Hoeschele I., Mendes P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics*, vol. 20, n° 18, p. 3565-3574.*

Gagnot S., Tamby J.-P., Martin-Magniette M.-L., Bitton F., Taconnat L., Balzergue S. et al. (2008). Catdb: a public access to arabidopsis transcriptome data from the urgv-catma platform. *Nucleic Acids Research*, vol. 36, n° suppl 1, p. D986-D990.

Ghazalpour A., Doss S., Zhang B., Wang S., Plaisier C., Castellanos R. *et al. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight.* PLOS GENETICS*, vol. 2, n° 8, p. 1182-1192.*

Giraud C., Huet S., Verzelen N. (2009). Graph selection with ggmselect. *Rapport technique. Ecole Polytechnique.*

Givry S. de, Bouchez M., Chabrier P., Milan D., Schiex T. (2005). Carthagene: multipopulation integrated genetic and radiated hybrid mapping. Bioinformatics*, vol. 21, n° 8, p. 1703-1704.*

Goldenberg A., Moore A. (2004). Tractable learning of large bayes net structures from sparse data. In Proceedings of the twenty-first international conference on machine learning*, p. 44–51.*

Hartemink A. (2005). Reverse engineering gene regulatory networks. Nature Biotechnology*, vol. 23, p. 554-555.*

Hilson P., Allemeersch J., Altmann T., Aubourg S., Avon A., Beynon J. et al. (2004). Versatile gene-specific sequence tags for arabidopsis functional genomics: Transcript profiling and reverse genetics applications. *Genome Research*, vol. 14, n° 10b, p. 2176-2189.

Hoops S., Sahle S., Gauges R., Lee C., Pahle J., Simus N. *et al. (2006). Copasi–a complex pathway simulator.* Bioinformatics*, vol. 22, n° 24, p. 3067-3074.*

Huynh-Thu V. A., Irrthum A., Wehenkel L., Geurts P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS ONE*, vol. 5.*

Jansen R., Nap J. (2001). Genetical genomics : the added value from segregation. Trends in genetics*, vol. 17, n° 7, p. 388-391.*

Jordan M. (Ed.). (1999). Learning in graphical models*. MIT Press.*

Keurentjes J. J. B., Fu J., Terpstra I. R., Garcia J. M., Ackerveken G. van den, Snoek L. B. et al. (2007). Regulatory network construction in arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences*, vol. 104, n° 5, p. 1708-1713.

Kontkanen P., Myllymäki P. (2007). A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, vol. 103, n° 6, p. 227 - 233.

Küffner R., Petri T., Tavakkolkhah P., Windhager L., Zimmer R. (2012). Inferring gene regulatory networks by anova. *Bioinformatics*, vol. 28, n° 10, p. 1376–1382.

Li H., Xuan J., Wang Y., Zhan M. (2008). Inferring regulatory networks. *Frontiers in Bioscience*, vol. 13, p. 263-275.

Liu B., Fuente A. de la, Hoeschele I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *GENETICS*, vol. 178, n° 3, p. 1763-1776.

Marbach D., Prill R. J., Schaffter T., Mattiussi C., Floreano D., Stolovitzky G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *PNAS*, vol. 107, p. 6286-6291.

Margolin A., Nemenman I., Basso K., Wiggins C., Stolovitzky G., Favera R. *et al. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics*, vol. 7, n° Suppl 1.*

McLachlan G. J. (1982). Classification pattern recognition and reduction of dimensionality. In, p. 199–208. Elsevier.

Mehrabian M., Allayee H., Stockton J., Lum P. Y., Drake T. A., Castellani L. W. et al. (2005). Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat Genet*, vol. 37, n° 11, p. 1224-1233.

Meinshausen N., Bühlmann P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, vol. 34, n° 3, p. 1436-1462.

Mendes P., Sha W., Ye K. (2003). Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, vol. 19, p. ii122-ii129.

Naïm P., Leray P., Pourret O., Wuillemin P.-H. (2008). *Réseaux bayésiens* (3rd éd.). Eyrolles.

Schäfer J., Strimmer K. (2005). An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, vol. 21, n° 6, p. 754-764.

Schwarz G. (1978). Estimating the dimension of a model. *The annals of statistics*.

Shtarkov Y. M. (1987). Universal sequential coding of single messages. *Problems of Information Transmission*, vol. 23, n° 3, p. 175–186.

Silander T., Kontkanen P., Myllymäki P. (2007). On sensitivity of the map bayesian network structure to the equivalent sample size parameter. In *Proc. of uai-07*, p. 360–367. Vancouver, Canada.

Silander T., Roos T., Myllymäki P. (2010). Learning locally minimax optimal bayesian networks. *International Journal of Approximate Reasoning*, vol. 51, n° 5, p. 544 - 557.

Simon M., Loudet O., Durand S., Bérard A., Brunel D., Sennesal F.-X. *et al. (2008). Qtl mapping in five new large ril populations of arabidopsis thaliana genotyped with consensus snp markers.* Genetics*, vol. 178, p. 2253-2264.*

Stuart P. (1982). Least squares quantization in pcm. IEEE Transactions on Information Theory.

Swarbreck D., Wilks C., Lamesch P., Berardini T., Garcia-Hernandez M., Foerster H. et al. (2008). The arabidopsis information resource (tair): gene structure and function annotation. *Nucleic Acids Research*, vol. 36, p. 1009-1014.

Tsamardinos I., Brown L., Aliferis C. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Mach. Learn.*, vol. 65, p. 31–78.

Vignes M., Vandel J., Allouche D., Ramadan-Alban N., Cierco-Ayrolles C., Schiex T. *et al. (2011, 12). Gene regulatory network reconstruction using bayesian networks, the dantzig selector, the lasso and their meta-analysis.* PLoS ONE*, vol. 6.*

Zhu J., Wiener M. C., Zhang C., Friedman A., Minch E., Lum P. Y. et al. (2007). Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLOS COMPUTATIONAL BIOLOGY*, vol. 3, n° 4, p. 692-703.