Reconstruction de réseau de régulation de gène à l'aide de données génomiques et de données génétiques.

Jimmy Vandel[1], Simon de Givry[1], Brigitte Mangin[1], and Matthieu Vignes[1]

Unité de Biométrie et d'Intelligence Artificielle
INRA, BP 52627, 31326 CASTANET-TOLOSAN cedex, FRANCE
jimmy.vandel@toulouse.inra.fr

Résumé : Une récente approche pour reconstruire un réseau de régulation de gène consiste à exploiter un grand nombre d'individus génétiquement différents sur lesquels on mesure les données (génomiques) d'expression de l'ensemble des gènes. La variabilité génétique influe directement sur la régulation des gènes, permettant de retrouver un plus grand nombre de régulations. L'information génétique (génotypes) d'un individu peut être mesurée expérimentalement à l'aide de marqueurs. Nous étudions l'exploitation conjointe des données d'expression et des données génétiques dans le formalisme des réseaux bayésiens discrets qui permet de capturer des dépendances non linéaires entre gènes et marqueurs. L'apprentissage de leur structure est faite par une recherche gloutonne classique dont le graphe initial est obtenu par une analyse statistique des dépendances linéaires entre l'expression d'un gène et les génotypes aux marqueurs. Nous montrons sur des données simulées l'amélioration offerte par l'utilisation des données génétiques sur la qualité des réseaux reconstruits.

Mots-clés : réseau bayésien discret, apprentissage automatique de la structure, réseau de régulation de gène, génomique génétique.

1 Introduction

Le gène est l'unité fonctionnelle de l'hérédité, porteur d'une information 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.

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 de gènes transcrits d'un très grand nombre de gènes différents seulement à travers un petit échantillon.

Ce cadre méthodologique découle des données sur les transcrits des gènes. En effet, les technologies de type *micro-array* permettent d'observer sur une unique puce un très grand nombre de transcrits. Le coût de cette technologie, bien que diminuant chaque mois, est encore trop élevé pour envisager un échantillon (de puces) de grande taille.

Les premières approches permettant de prédire la topologie d'un réseau se sont attachées à décrire et quantifier les relations locales entre deux gènes quelconques du réseau, le réseau global étant directement construit par seuillage de l'estimateur choisi pour quantifier la relation locale (Ghazalpour *et al.*, 2006; Margolin *et al.*, 2006; Opgen-Rhein & Strimmer, 2007). Les approches globales de reconstruction de réseau sont fondées sur des modèles graphiques (Jordan, 1999). Parmi celles-ci l'estimation de modèles graphiques gaussiens (Chu *et al.*, 2009; Giraud *et al.*, 2009), les équations structurelles (Liu *et al.*, 2008) et les réseaux bayésiens discrets (Friedman, 2004; Zhu *et al.*, 2007; Auliac *et al.*, 2008). Nous avons choisi d'explorer cette dernière approche qui permet de modéliser des interactions complexes (non linéaires).

Ces dernières années, plusieurs travaux (Jansen & Nap, 2001; Ghazalpour *et al.*, 2006; Zhu *et al.*, 2007; Liu *et al.*, 2008; Chu *et al.*, 2009) ont montré l'intérêt d'exploiter des données de polymorphisme pour mieux reconstruire un réseau de régulation de gène. 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 dans la région promotrice du gène ou dans sa région codante ayant pour effet de modifier le niveau d'expression du gène (cas en région promotrice) ou la nature de la protéine produite (cas en région codante), influant de manière complexe (non linéaire) sur la régulation d'autres gènes.

Par rapport à Zhu *et al.* (2007), nous proposons une modélisation par un réseau bayésien qui intègre explicitement les données de polymorphisme dans ses variables plutôt que via des a-priori sur la structure du réseau. Ce modèle est présenté en Section 2. L'algorithme de reconstruction est une recherche gloutonne maximisant un score adapté à notre choix de modélisation (Section 3). Le réseau initial de cette recherche est obtenu par une analyse statistique des dépendances entre l'expression d'un gène et les génotypes aux marqueurs (Section 4).

Nous nous plaçons dans le cadre d'une approche à l'échelle du génome. C'est à dire que l'on dispose d'une carte génétique dense permettant d'observer toutes les mutations fonctionnelles chez un individu et que les mesures d'expression se font sur tous les gènes de l'organisme étudié. Les mesures d'expression sont effectuées en état stationnaire du fonctionnement de la cellule, en environnement contrôlé (mêmes facteurs environnementaux pour tous les individus) et l'on exclue les phénomènes *épigénétiques* (*i.e.* régulation de gènes via des signaux ne provenant pas de la séquence). Ainsi l'apprentissage de la structure s'effectue dans le cadre de données complètement observées sans variables latentes.

Nous présentons des résultats expérimentaux en Section 7 obtenus sur des données simulées décrites en Section 5 et qui ont fait l'objet d'une discrétisation (pour les niveaux d'expression) présentée en Section 6.

2 Modélisation par un réseau bayésien avec des variables de génotype et d'expression fusionnées

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 peut se factoriser de la manière suivante :

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

Le nombre de paramètres indépendants définissant les tables de 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.

Nous présentons d'abord une modélisation par un réseau bayésien distinguant deux types de variable (modèle dit *non fusionné*) : des variables discrètes représentant les génotypes aux marqueurs et des variables continues représentant les niveaux d'expression des gènes. Du fait du formalisme retenu, les variables d'expression sont discrétisées. Cette discrétisation est présentée en Section 6.

Notre modèle fait l'hypothèse simplificatrice de l'existence d'au plus un polymorphisme fonctionnel, c'est à dire ayant un effet sur la régulation des gènes, présent dans chaque gène. On fait l'hypothèse que ce polymorphisme est lié à une mutation ponctuelle de l'ADN appelée *Single Nucleotide Polymorphism* (SNP) et que cette mutation est 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 de génotype M_i pouvant prendre deux valeurs (normal ou muté[1]) et une variable d'expression E_i dont le nombre de valeurs dépend de la discrétisation (limité à 4).

Les différents cas de régulation entre gènes sont résumés sur un exemple de réseau bayésien décrit en Figure 1.

Cette modélisation introduit deux types d'arcs que nous ne souhaitons pas apprendre, ceux entre marqueurs dus à la liaison génétique et ceux reliant l'expression d'un gène à un marqueur qui représentent un non sens biologique. Afin d'éviter le parcours de graphes comportant ces arcs superflus, nous choisissons d'effectuer une fusion des deux variables associées à un gène (M_i et E_i) en une nouvelle variable G_i dont le domaine est le produit cartésien de M_i et E_i (voir Figure 2). Cette fusion a aussi pour avantage de travailler sur un nombre de variables plus petit, accélérant ainsi la recherche d'une bonne structure, sans perte de précision sur la reconstruction du réseau de régulation.

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

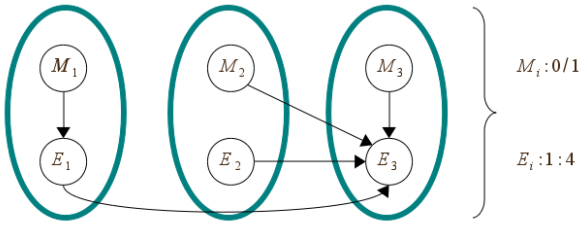

FIG. 1 – Modèle non fusionné. Trois exemples de régulation pour le gène 3 : (i) la valeur du génotype M_3 (0 : normal ou 1 : muté) a un effet sur le niveau d'expression E_3 (i.e. une mutation dans la région promotrice de ce gène active/inhibe son expression), (ii) le gène 1 régule le gène 3 (i.e. la mutation dans la région promotrice du gène 1 active/inhibe son expression qui à son tour vient activer/inhiber l'expression du gène 3), (iii) le gène 2 régule le gène 3 en fonction de son génotype M_2 et de son niveau d'expression E_2 (i.e. *dépendance non linéaire* de (M_2, E_2) sur E_3 due à la mutation dans la région codante du gène 2 combinée à son expression qui active/inhibe l'expression du gène 3).

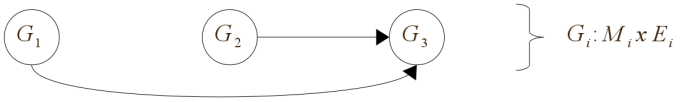

FIG. 2 – Modèle fusionné.

Au final, le réseau de régulation de gène est obtenu en prenant le graphe partiellement orienté représentant de la classe d'équivalence de Markov du réseau bayésien trouvé (Chickering, 2002).

3 Apprentissage de la structure à partir d'un score BIC modifié

Il existe deux grandes familles d'apprentissage d'un réseau bayésien fondées soit sur la recherche d'indépendances conditionnelles soit sur l'optimisation d'un score (Naïm *et al.*, 2008). C'est cette deuxième famille que nous avons retenue et qui repose sur l'exploration de l'espace des graphes possibles pour trouver celui qui maximise un score à partir de données complètement observées. Dans cette étude, nous avons choisi le score *Bayesian Information Criterion* (BIC) (Schwartz, 1978) qui combine un terme de vraisemblance des données et une pénalité sur la complexité du modèle. L'expression du score BIC d'un graphe dirigé sans circuit $\mathcal{G}$ prend donc la forme suivante :

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

où $\mathbf{D}$ représente les données pour n échantillons et $B_{\mathcal{G}} = (\mathcal{G}, \mathbf{P}_{\mathcal{G}}{}^{MV})$ est le réseau bayésien défini par $\mathcal{G}$ et des probabilités conditionnelles $\mathbf{P}_{\mathcal{G}}{}^{MV}$ estimées suivant le principe

de maximum de vraisemblance (Naïm *et al.*, 2008).

Dans notre modèle fusionné, le score BIC peut se réécrire par :

$$\begin{aligned}
BIC(\mathcal{G}) &= \log(\prod_{l=1}^{n}\prod_{i=1}^{p}\mathbb{P}(g_i^l|Pa(g_i^l))) - \frac{1}{2}\log(n)Dim(B_{\mathcal{G}}) \\
&= \log(\prod_{l=1}^{n}\prod_{i=1}^{p}\mathbb{P}(m_i^l, e_i^l|Pa(g_i^l))) - \frac{1}{2}\log(n)Dim(B_{\mathcal{G}}) \\
&= \log(\prod_{l=1}^{n}\prod_{i=1}^{p}\mathbb{P}(e_i^l|Pa(g_i^l), m_i^l)\mathbb{P}(m_i^l|Pa(g_i^l))) - \frac{1}{2}\log(n)Dim(B_{\mathcal{G}}) \\
&= \sum_{l=1}^{n}\sum_{i=1}^{p}\log(\mathbb{P}(e_i^l|Pa(g_i^l), m_i^l)) + \sum_{l=1}^{n}\sum_{i=1}^{p}\log(\mathbb{P}(m_i^l|Pa(g_i^l))) - \frac{1}{2}\log(n)Dim(B_{\mathcal{G}})
\end{aligned}$$

avec $g_i^l = \{m_i^l, e_i^l\}$, l'observation du génotype et du niveau d'expression du gène i de l'individu l et $Pa(g_i^l)$, les observations des parents dans $\mathcal{G}$ du gène i de l'individu l. Le terme $T = \sum_{l=1}^{n}\sum_{i=1}^{p}\log(\mathbb{P}(m_i^l|Pa(g_i^l)))$ ne dépend que de la liaison génétique entre marqueurs et de la structure du réseau appris[2]. Du fait de la fusion des variables, l'estimation des paramètres inclut une estimation de la liaison génétique conduisant à des arêtes que nous ne souhaitons pas apprendre. Par la suite, nous travaillons avec un score BIC modifié (correspondant à la vraisemblance conditionnellement aux génotypes) où nous avons supprimé la dépendance induite par la liaison génétique dans la vraisemblance (en soustrayant par le terme T) et également ajouté un terme lié à la modification de la dimension du réseau bayésien résultant :

$$BIC'(\mathcal{G}) = BIC(\mathcal{G}) - T + \frac{1}{2}\log(n)\sum_{i=1}^{p} q_i$$

avec q_i, le nombre de configurations possibles pour les parents de G_i dans $\mathcal{G}$.

Le terme T se calcule par programmation dynamique dans le modèle classiquement attendu de la liaison génétique qui est une chaîne de Markov non homogène d'ordre 1 portant sur les génotypes des marqueurs ordonnés suivant leur position sur les chromosomes[3]. Les probabilités de transition dépendent de la distance génétique entre marqueurs consécutifs qui est supposée connue (les distances sont fournies par une carte génétique de l'espèce).

A partir de ce score BIC modifié, la recherche d'un graphe de score maximum est un problème d'optimisation NP-difficile (Chickering & Heckermann, 1996). Les meilleures méthodes exactes sont limitées à une trentaine de sommets (Silander & Myllymäki, 2006). Pour cette raison l'utilisation d'heuristiques et d'algorithmes de parcours intelligent d'un sous-ensemble des graphes possibles est nécessaire.

---

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

[3]Ainsi, les termes $\mathbb{P}(m_i^l|Pa(m_i^l))$ se simplifient en ne conservant dans le conditionnement que les marqueurs *flanquants* (marqueurs le plus proche à gauche et à droite) de m_i^l présents dans $Pa(m_i^l)$.

Nous utilisons dans cet article la recherche gloutonne *Greedy Search* (GS) (Chickering, 2002), qui considère toutes les modifications locales du graphe courant (via des opérations d'inversion, d'ajout, ou de suppression d'arcs sans introduire de circuit) et sélectionne dans ce voisinage celle qui maximise un score, en itérant ce processus jusqu'à obtenir un maximum local.

Le point de départ de cet algorithme est un élément important afin de réduire le risque de convergence vers un maximum local de mauvaise qualité. Nous proposons de partir d'un graphe initial obtenu par une analyse statistique des dépendances entre l'expression d'un gène et les génotypes aux marqueurs décrite en Section 4.

4 Construction d'un graphe initial par analyse statistique des dépendances entre expression et génotypes

L'analyse statistique, appelée dans les expérimentations en Section 7 *analyse eQTL* (*expression Quantitative Trait Loci*), a été effectuée à l'aide du logiciel MCQTL (Jourjon *et al.*, 2005), dans l'objectif de sélectionner pour chaque niveau d'expression de gène E_i (non discrétisé) l'ensemble $\mathcal{S}_i$ des marqueurs dont les génotypes expliquent au mieux la variabilité de E_i. Cette sélection de marqueurs s'est opérée dans un modèle de régression linéaire très classique dans le contexte de la cartographie des QTL (Haley & Knott, 1992) où la variable à expliquer est E_i et les variables explicatives l'ensemble de tous les génotypes/marqueurs. L'heuristique pour sélectionner l'ensemble $\mathcal{S}_i$ est une méthode itérative combinant une méthode *forward* et une méthode *backward* adaptées au cadre de variables explicatives ordonnées sur une carte génétique. La méthode *forward* est classique et consiste à ajouter, au modèle courant, le marqueur dont le critère statistique est le plus élevé, tout en restant supérieur à une valeur seuil. La méthode *backward* est adaptée à l'ordonnancement des marqueurs. Chaque marqueur du modèle courant est remis successivement en cause : le marqueur lui-même ou un marqueur voisin, qui possède le critère statistique le plus élevé tout en restant supérieur à un seuil, reste dans le modèle. La méthode itérative s'arrête lorsque l'ensemble $\mathcal{S}_i$ n'évolue plus. Le voisinage d'un marqueur est défini par la carte génétique. Le critère statistique utilisé est le test classique de Fischer. La valeur seuil a été calculée par permutation aléatoire pour contrôler une erreur de type I du test de 20%.

L'ensemble $\mathcal{S}_i$ ainsi obtenu a été augmenté pour chacun de ses marqueurs, des deux marqueurs flanquants, afin de relaxer légèrement le modèle retenu au final par MCQTL. De plus, le nombre de parents d'un gène entrant dans le réseau a été limité à 9 marqueurs (pour des raisons d'espace mémoire des tables de probabilité conditionnelle), c'est-à-dire 3 groupes de 3 marqueurs, sur la base du test de Fisher.

Du fait de notre hypothèse de bijection entre marqueurs et gènes, nous pouvons remplacer les marqueurs par les variables G_j correspondantes dans chaque $\mathcal{S}_i$ et définir ainsi un graphe initial dans le modèle fusionné. Ce graphe peut contenir des circuits qui doivent être supprimés. La suppression du nombre minimum d'arcs dans un graphe orienté quelconque pour le rendre sans circuit (*minimum arc set problem*) est un problème NP-difficile (Festa *et al.*, 2009). Nous utilisons une approche heuristique exploitant la matrice d'accessibilité qui définit pour toute paire orientée de sommets l'exis-

tence d'un chemin dans le graphe. La diagonale de cette matrice permet de localiser aisément la présence de circuits dans le graphe ainsi que l'ensemble C des sommets impliqués dans au moins un circuit. Il suffit alors de choisir un élément de C qui ait le plus grand nombre d'arcs sortants vers d'autres sommets de C et de supprimer un à un les arcs vers ces sommets, en testant de manière itérative si chacune de ces suppressions permet de produire un graphe sans circuit. Le graphe obtenu après cette étape est utilisé comme graphe initial pour la recherche gloutonne.

5 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 choix aléatoire sur 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 (de Givry *et al.*, 2005), en disposant aléatoirement 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). Cette position doit être commune pour un gène à tous les individus étant donné que chaque mutation provient d'un des parents fondateurs.

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 (Liu *et al.*, 2008). Elle se traduit par une équation différentielle ordinaire prenant en compte les activateurs et les inhibiteurs d'un gène et aussi la dégradation naturelle des gènes transcrits au cours du temps :

$$\frac{dE_i}{dt} = V_i \prod_{E_j \in Inh(E_i)} Z_j\left(\frac{1}{1+E_j}\right) \prod_{E_k \in Act(E_i)} Z_k\left(1+\frac{E_k}{E_k+1}\right) - E_i + \theta_i E_i$$

avec V_i, le taux moyen de transcription du gène i (ce paramètre vaut 0.75 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 0.75 si la mutation a lieu dans la région codante de ces gènes et 1 sinon) et θ_i, un bruit qui suit une loi normale de moyenne 0 et d'écart-type 0.1.

Nous utilisons le simulateur de réactions biochimiques COPASI (Hoops *et al.*, 2006) pour établir un état stable 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 E_i un *bruit expérimental* de 10 % proportionnel à la variance de E_i, à l'instar de (Liu *et al.*, 2008).

6 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.

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 de classes prédéfini, 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 3, 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 3, nous appliquons une méthode de modèle de mélange gaussien.

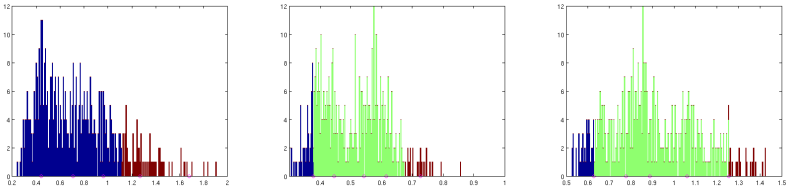

FIG. 3 – Discrétisation par la méthode modifiée des *k-means* appliquée à l'expression respectivement d'un gène du réseau Web50_001 et de 2 gènes du réseau Web50_008 (*de gauche à droite*, E_3, E_{33} et E_{13}). Chaque figure donne l'histogramme du nombre d'individus en fonction de l'expression du gène. Les différentes couleurs représentent les 2 ou 3 classes trouvées par la méthode.

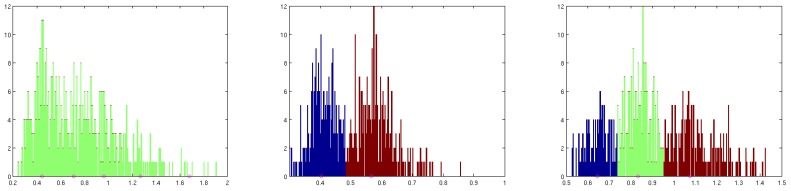

FIG. 4 – Discrétisation par modèle de mélange gaussien pour les expressions E_3, E_{33} et E_{13}.

Modèle de mélange gaussien.

Nous effectuons dans un premier temps un lissage de l'histogramme. 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 de k gaussiennes (McLachlan, 1982). Au final, le nombre de classes k est limité à 4 (voir les exemples en Figure 4). Dans le cas d'un mode unique, c'est la méthode modifiée des *k-means* qui est utilisée.

7 Résultats expérimentaux

La recherche gloutonne GS provient de la librairie Matlab d'algorithmes d'apprentissage *Bayes Net Toolbox Structure Learning Package*[4] *v1.4c* (François & Leray, 2006). Le score BIC a été modifié comme indiqué en Section 3. Le nombre maximum de parents trouvés par l'algorithme GS est limité à 9 par gène (pour des raisons d'espace mémoire). La discrétisation par modèle de mélange gaussien utilise la *Statistics Toolbox* de Matlab r2009b.

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 présents uniquement dans le réseau reconstruit, et FN, le nombre d'arcs présents uniquement dans le vrai réseau. Nous ne prenons pas en compte l'orientation des arcs dans ces mesures.

Les résultats présentés sont moyennés sur 50 réseaux Web50_0XX (décrits en Section 5). Afin de justifier de l'intérêt de l'utilisation des données génétiques pour la reconstruction de réseaux de régulation de gène, nous comparons les résultats de l'algorithme GS dans trois situations : (i) *Exp*, modèle bayésien discret avec uniquement les variables d'expression E_i, (ii) *Exp+Geno+MWST*, modèle bayésien fusionné (décrit en Section 2) avec un graphe initial obtenu par un arbre couvrant de score BIC modifié maximum (François & Leray, 2006), (iii) *Exp+Geno+eQTL*, modèle fusionné avec un graphe initial obtenu par l'analyse eQTL (décrite en Section 4).

[4] http ://www.cs.ubc.ca/ murphyk/Software/BNT/bnt.html et http ://bnt.insa-rouen.fr/.

	Exp	*Exp+Geno+MWST*	*Exp+Geno+eQTL*
précision	0.26	0.52	0.61
sensibilité	0.23	0.39	0.48
nombre d'arcs $TP + FP$	45	37	40

L'introduction des données génétiques permet de doubler la précision et de faire un gain de 70% en sensibilité. L'initialisation du graphe par l'analyse eQTL améliore d'environ 20% par rapport à une initialisation par arbre couvrant. Bien que le score BIC du graphe initial issu de l'analyse eQTL soit inférieur à celui de l'arbre couvrant (résultats non reportés par manque de place), sa qualité en terme de précision et de sensibilité est bien meilleure. L'algorithme GS améliore ensuite légèrement la précision, principalement en retirant des arêtes, parfois au détriment de la sensibilité. Enfin, la valeur du score BIC final pour l'approche *Exp+Geno+eQTL* est la plupart du temps meilleur que pour l'approche *Exp+Geno+MWST*.

A noter que nous avons obtenu des résultats de qualité légèrement inférieure en remplaçant GS par l'algorithme K2 (Cooper & Hersovits, 1992). De même, une méthode consistant à tirer aléatoirement x arêtes parmi les $\frac{p(p-1)}{2}$ arêtes possibles aurait une précison inférieure à 4% quelque soit la valeur de x.

Les résultats de notre approche *Exp+Geno+eQTL* sont cependant de qualité nettement inférieure à ceux reportés dans Liu *et al.* (2008) qui sont obtenus dans le cadre des modèles d'équations structurelles (*Structural Equation Modeling*, SEM). Les SEM sont équivalentes à un modèle graphique gaussien pour les relations entre les niveaux d'expression de gènes aux quelles s'ajoutent les régressions linéaires qui modélisent les liens entre les niveaux d'expression de gènes et les génotypes aux marqueurs. La méthode de Liu *et al.* (2008) consiste à créer un graphe de départ foisonnant en testant par régressions linéaires successives un grand nombre de couples (marqueur, gène) pouvant expliquer le niveau d'expression d'un gène donné. A partir de ce graphe de départ, comportant tous les liens jugés significatifs, une sélection de modèle de type *Occam's window model selection* (Madigan & Raftery, 1994) est menée sur le score BIC de la vraisemblance du modèle SEM. Liu *et al.* (2008) obtiennent une précision et une sensibilité supérieures à 80%. Néanmoins cet écart de performance comparé à nos résultats pourrait s'expliquer en partie par des différences dans la génération des données simulées. En effet, l'article ne précise pas la carte génétique employée et nous avons observés une meilleure reconstruction si les marqueurs sont placés de manière équidistante.

8 Conclusion

Nous avons présenté l'application au problème de la reconstruction de réseau de gène d'une méthode d'apprentissage de structure dans le cadre des réseaux bayésiens. La prise en compte de données supplémentaires (données de génotypes) directement dans le modèle a permis d'améliorer la qualité des réseaux reconstruits par rapport aux seules données d'expression. L'approche statistique étudiant le lien entre les données de génotype et les données d'expression (analyse eQTL) semble complémentaire à l'approche réseau bayésien (régression linéaire sur des variables continues versus maximisation

de la vraisemblance sur des variables discrètes avec des dépendances pouvant être non linéaires). L'intégration de l'approche statistique au travers d'un graphe initial fourni à la recherche gloutonne est une manière assez triviale de combiner les deux approches. D'autres combinaisons sont à explorer, notamment celle introduisant un *a priori* sur les structures intégré à un score *Bayesian Dirichlet* (Cooper & Hersovits, 1992). De même, une comparaison sur d'autres jeux de données (en faisant varier les paramètres du simulateur) et avec d'autres méthodes de reconstruction (Margolin *et al.*, 2006; Opgen-Rhein & Strimmer, 2007; Liu *et al.*, 2008; Chu *et al.*, 2009; Giraud *et al.*, 2009) est à faire.

Le cas de données réelles avec un grand nombre de SNP au voisinage d'un gène risque de remettre en cause notre hypothèse d'un unique polymorphisme fonctionnel par gène et pose la difficulté d'un très grand nombre de paramètres à apprendre. Dans des travaux futurs, nous envisageons d'étudier des algorithmes d'apprentissage de la structure de réseau bayésien plus performants à l'instar de (Hartemink, 2005; Auliac *et al.*, 2008) et d'examiner d'autres fonctions de score plus robustes à un échantillon de petite taille (Silander *et al.*, 2008).

Références

Auliac C., Frouin V., Gidrol X. & D'alche-Buc F. (2008). Evolutionary approaches for the reverse-engineering of gene regulatory networks : a study on a biologically realistic dataset. *BMC Bioinformatics*, **9**, 91–104.

Barabasi A. & Oltvai Z. (2004). Network biology : Understanding the cell's functional organization. *NATURE REVIEWS GENETICS*, **5**(2), 101–115.

Chickering D. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*.

Chickering D. & Heckermann D. (1996). Learning bayesian networks is NP-complete. *In learning from data : Al and Statistics*.

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*, **3**(55).

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

de Givry S., Bouchez M., Chabrier P., Milan D. & Schiex T. (2005). CARTHAGENE : multipopulation integrated genetic and radiated hybrid mapping. *Bioinformatics*, **21**(8), 1703–1704.

Festa P., Pardalos P. M. & Resende M. G. C. (2009). Feedback set problems. In *Encyclopedia of Optimization*, p. 1005–1016. Springer.

François O. & Leray P. (2006). Étude comparative d'algorithmes d'apprentissage de structure dans les réseaux bayésiens. *Journal électronique d'intelligence artificielle*, **5**(39).

Friedman N. (2004). Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*, **303**(5659), 799–805.

Ghazalpour A., Doss S., Zhang B., Wang S., Plaisier C., Castellanos R., Brozell A., Schadt E., Drake T., Lusis A. & Horvath S. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLOS GENETICS*, **2**(8), 1182–1192.

GIRAUD C., HUET S. & VERZELEN N. (2009). *Graph selection with GGMselect*. Rapport interne, Ecole Polytechnique.

HALEY C. & KNOTT S. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69**(4), 315–324.

HARTEMINK A. (2005). Reverse engineering gene regulatory networks. *Nature Biotechnology*, **23**, 554–555.

HOOPS S., SAHLE S., GAUGES R., LEE C., PAHLE J., SIMUS N., SINGHAL M., XU L., MENDES P. & KUMMER U. (2006). COPASI–a COmplex PAthway SImulator. *Bioinformatics*, **22**(24), 3067–3074.

JANSEN R. & NAP J. (2001). Genetical genomics : the added value from segregation. *Trends in genetics*, **17**(7), 388–391.

M. JORDAN, Ed. (1999). *Learning in Graphical Models*. MIT Press.

JOURJON M.-F., JASSON S., MARCEL J., NGOM B. & MANGIN B. (2005). MCQTL : multi-allelic QTL mapping in multi-cross design. *Bioinformatics*, **21**(1), 128–130.

LI H., XUAN J., WANG Y. & ZHAN M. (2008). Inferring regulatory networks. *Frontiers in Bioscience*, **13**, 263–275.

LIU B., DE LA FUENTE A. & HOESCHELE I. (2008). Gene network inference via structural equation modeling in genetical genomics experiments. *GENETICS*, **178**(3), 1763–1776.

MADIGAN D. & RAFTERY A. (1994). Model selection and accounting for model uncertainty in graphical model using occam's window. *Jounal of the American Statistical Association*, **89**(428), 1535–1546.

MARGOLIN A. A., NEMENMAN I., BASSO K., WIGGINS C., STOLOVITZKY G., FAVERA R. D. & CALIFANO A. (2006). ARACNE : An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, **7**(1), S7.

MCLACHLAN G. J. (1982). *Classification pattern recognition and reduction of dimensionality*, chapter The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis, p. 199–208. Elsevier.

MENDES P., SHA W. & YE K. (2003). Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, **19**, ii122–ii129.

NAÏM P., LERAY P., POURRET O. & WUILLEMIN P.-H. (2008). *Réseaux bayésiens*. Eyrolles, 3rd edition.

OPGEN-RHEIN R. & STRIMMER K. (2007). From correlation to causation networks : a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC SYSTEMS BIOLOGY*, **1**.

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

SILANDER T. & MYLLYMÄKI P. (2006). A simple approach for finding the globally optimal bayesian network structure. In *Proc. of UAI-06*, p. 445–452, Cambridge, MA.

SILANDER T., ROOS T., KONTKANEN P. & MYLLYMÄKI P. (2008). Factorized normalized maximum likelihood criterion for learning Bayesian network structures. In *4th European Workshop on Probabilistic Graphical Models*, Hirtshals, Denmark.

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

ZHU J., WIENER M. C., ZHANG C., FRIDMAN A., MINCH E., LUM P. Y., SACHS J. R. & SCHADT E. E. (2007). Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLOS COMPUTATIONAL BIOLOGY*, **3**(4), 692–703.