Réseaux bayésiens hiérarchiques avec variables latentes pour la modélisation des dépendances entre SNP: une approche pour les études d'association pangénomiques

Raphaël Mourad, Christine Sinoquet, Philippe Leray

Réseaux bayésiens hiérarchiques avec variables latentes pour la modélisation des dépendances entre SNP: une approche pour les études d'association pangénomiques.

Raphaël Mourad*, Christine Sinoquet**, Philippe Leray*

*LINA, UMR CNRS 6241, Ecole Polytechnique de l'Université de Nantes,
la Chantrerie, rue Christian Pauc, BP 50609, 44306 Nantes Cedex 3, France
{raphael.mourad,philippe.leray}@univ-nantes.fr
http://www.lina.univ-nantes.fr
**LINA, UMR CNRS 6241, Université de Nantes,
2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03, France
christine.sinoquet@univ-nantes.fr
http://www.lina.univ-nantes.fr

Résumé. Découvrir la base génétique des maladies génétiques communes constitue un enjeu de santé publique. Cependant cette tâche présente de nombreuses difficultés comme la dimension des données à traiter et l'identification des mutations causales. Dans cette perspective, la modélisation des dépendances entre marqueurs génétiques à l'aide de réseaux bayésiens hiérarchiques offre de multiples avantages: une réduction de la dimension des données à l'aide de variables latentes et l'identification des marqueurs causaux grâce à la propriété d'indépendance conditionnelle.

1 Introduction

Dans le contexte des maladies génétiques communes comme le diabète, l'hypertension artérielle ou l'asthme, les études d'association pangénomiques (GWAS) sont développées afin de localiser les facteurs génétiques causaux. A cette fin, des données de grande dimension doivent être analysées, comme par exemple plus d'un million de SNP (i.e., de marqueurs génétiques). Le grand nombre de tests statistiques à réaliser engendre la détection d'un nombre important de fausses associations, diminuant aussi la puissance statistique. Une approche intéressante consiste à tirer parti de la forte corrélation existante entre les marqueurs proches, autrement dit du déséquilibre de liaison (LD), afin de réduire la dimension des données. Dans cette optique, différentes approches ont été développées comme les méthodes fondées sur l'inférence d'haplotypes (Schaid (2004)), sur les blocs haplotypiques (Pattaro et al. (2008)) et sur la sélection de tags SNP (Stram (2004)). En outre, dans la situation où des régions d'intérêt sur l'ADN ont déjà été localisées, il est nécessaire d'identifier ensuite les éventuels marqueurs causaux vis-à-vis de la maladie. Cette étape est appelée cartographie fine (fine mapping) et consiste à analyser tous les marqueurs génétiques à disposition dans la région d'intérêt. Dans cette perspective, des études se basent sur l'analyse des haplotypes et des tags SNP (Pittman

et al. (2005)). Une modélisation du LD à l'aide de forêts de modèles hiérarchiques latents (FHLCMs, Forest of Hierarchical Latent Class Models) apparaît comme une solution pour le traitement de données de grande dimension et pour la cartographie fine d'une région d'intérêt ; notamment, grâce à la capacité des FHLCMs à synthétiser l'information pour réduire la dimension et à distinguer l'influence de chaque variable afin d'identifier les mutations causales.

En section 2, nous rappelons la définition des modèles latents et des modèles hiérarchiques latents. En section 3, nous présentons les avantages d'une modélisation du LD par les FHLCMs, ainsi que l'approche proposée pour les GWAS. Enfin, en section 4, nous concluons sur l'intérêt des FHLCMs et discutons des perspectives soulevées par notre approche.

2 Modèle latent et modèle hiérarchique latent

Dans la suite de l'article, nous nous restreignons aux variables discrètes (latentes ou observées). Les réseaux bayésiens sont des modèles graphiques probabilistes (Naïm et al. (2007)). Ils sont définis par un graphe orienté sans circuit (la structure) représentant les relations de dépendance dans le groupe de variables étudiées et par une distribution de probabilités conditionnelles associée à chaque variable (les paramètres). Les modèles latents (LCMs, Latent Class Models) forment une classe particulière de réseaux bayésiens : toutes les variables observées (VO) sont dépendantes d'une unique variable latente (VL) (Figure $1(a)$). Les modèles latents sont généralement utilisés pour la classification non supervisée. Cependant, ces modèles reposent sur une hypothèse souvent fausse : l'indépendance locale (Zhang et Kocka (2004)), c'est-à-dire que les VO sont toutes mutuellement indépendantes conditionnellement à la VL. Les modèles hiérarchiques latents (HLCM, Hierarchical Latent Class Models) généralisent les modèles latents et ne basent plus sur cette hypothèse. Leur structure est celle d'un arbre dont les feuilles sont des VO et les noeuds internes sont des VL (par exemple Figure $1(b)$).

3 Modélisation du LD et approche proposée pour les GWAS

Dans le génome humain, les dépendances entre SNP (Single Nucleotide Polymorphisms) présentent une structure caractéristique, appelée structure en blocs haplotypiques (International HapMap Consortium (2003)). Des régions à forte corrélation entre SNP sont séparées par de petites régions à faible corrélation, appelées hotspots de recombinaison. Des méthodes ont été développées, notamment les méthodes basées sur les blocs haplotypiques, afin de tirer parti de cette caractéristique (voir Section 1). Cependant, les frontières entre blocs haplotypiques ne sont pas toujours bien définies. De ce fait, les méthodes basées sur l'inférence de blocs haplotypiques ne parviennent généralement pas à capturer toutes les dépendances entre SNP. La Figure $1(c)$ illustre le déséquilibre de liaison à l'aide d'une séquence de 50 kilobases (kb). Nous constatons que le découpage en haploblocs ne prend pas en compte toutes les dépendances entre SNP, mais seulement les dépendances fortes entre SNP contigus (blocs). Grâce à leur structure hiérarchique, les HLCMs pourraient offrir une modélisation plus fine et plus souple du LD que les méthodes basées sur les haploblocs : plus fine, car de nombreuses dépendances entre SNP peuvent être prises en compte, qu'elles soient directes ou indirectes ; plus souple, car des dépendances fortes entre SNP voisins et des dépendances plus faibles entre

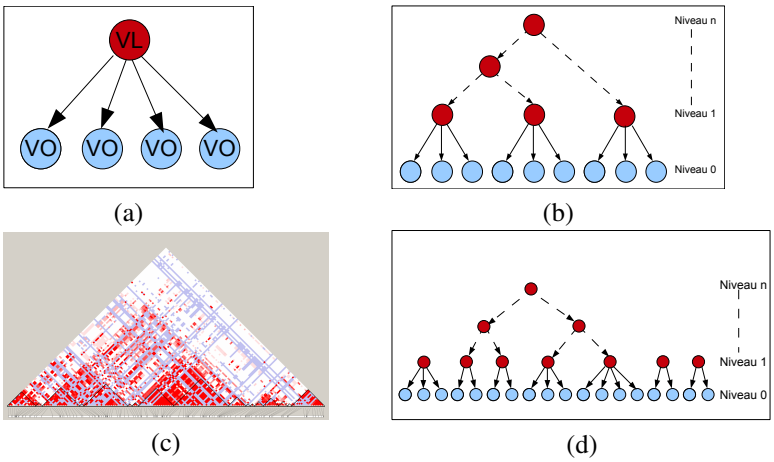

FIG. 1 – *(a) Modèle latent. Les variables observées sont colorées en bleu (couleur claire), tandis que les variables latentes sont colorées en rouge (couleur foncée). (b) Modèle hiérarchique latent. (c) LD plot (matrice des dépendances statistiques entre paires de marqueurs génétiques). Génome humain, chromosome 2, région de 50kb [234 357kb - 234 407kb]. (d) Forêt de modèles hiérarchiques latents.*

SNP plus éloignés sont modélisées. Différents niveaux de dépendance sont ainsi modélisés. Cependant, les HLCMs présentent l'inconvénient de contraindre tous les SNP à être dépendants entre eux, directement ou indirectement, s'écartant ainsi de la structure réelle du LD. En effet, le LD s'observe rarement entre SNP séparés de plus de 500kb. Pour résoudre ce problème, nous proposons l'emploi de forêts de HLCMs (FHLCMs), présentées en Figure $1(d)$, qui généralisent les HLCMs et ne contraignent plus tous les SNP à être dépendants entre eux.

La modélisation du LD par les FHLCMs offre de nombreuses possibilités pour les études d'association pangénomiques. Tout d'abord, les valeurs manquantes des variables latentes peuvent être imputées et ces variables peuvent être utilisées, comme les haploblocs ou les tags SNP, pour la réduction de dimension. Les FHLCMs peuvent être perçus alors comme un outil de data mining. En effet, grâce à la structure hiérarchique du modèle, l'utilisateur peut commencer par les niveaux les plus élevés du modèle, puis en descendant de niveau, il peut "zoomer" sur des régions d'intérêt. L'utilisateur dispose ainsi de plusieurs niveaux de réduction de dimension des données. Il peut ainsi déployer, par exemple, une stratégie descendante lors de la recherche d'association SNP-maladie : les tests d'association avec le phénotype maladie peuvent être d'abord réalisés avec les variables latentes des plus hauts niveaux, puis lorsque certaines régions à fortes associations sont ciblées, l'utilisateur peut descendre de niveau progressivement afin d'idenfitier de plus en plus finement le ou les SNP associés à la maladie. Par ailleurs, dans la problématique d'identification des mutations causales, les FHLCMs devraient permettre de distinguer les SNP directement associés (vrais positifs), i.e. les marqueurs causaux, des SNP indirectement associés à la maladie par le LD (faux positifs). Pour cela, des tests d'association SNP-maladie conditionnellement à la variable parente (latente) permettraient de

distinguer l'influence d'un SNP de celle des autres SNP présents dans le FHCLMs. Outre ces deux applications des FHLCMs, ces modèles pourraient être aussi employés comme outils de visualisation du LD à l'aide de leur graphe, ou pour la modélisation des dépendances entre marqueurs génétiques liées à la structure de la population. Récemment, nous avons développé un premier algorithme d'apprentissage de FHLCMs pour la première application proposée, la réduction de dimension de données de GWAS (soumis).

4 Conclusions et perspectives

Dans le cadre des GWAS, l'emploi de FHLCMs offrent de nombreuses possibilités, notamment pour la réduction de dimension des données génétiques et l'identification des mutations causales. De prochaines études permettront d'évaluer, sur des données simulées et réelles, l'efficacité de cette approche pour la localisation des mutations causales.

Références

International HapMap Consortium (2003). The international hapmap project. *Nature 426*(6968), 789–796.

Naïm, P., P.-H. Wuillemin, P. Leray, O. Pourret, et A. Becker (2007). *Réseaux bayésiens* (3 ed.).

Pattaro, C., I. Ruczinski, D. M. Fallin, et G. Parmigiani (2008). Haplotype block partitioning as a tool for dimensionality reduction in snp association studies. *BMC Genomics 9*, 405.

Pittman, A., A. Myers, P. Abou-Sleiman, K. M. Fung, H., L. Marlowe, J. Duckworth, D. Leung, D. Williams, L. Kilford, N. Thomas, C. Morris, D. Dickson, N. Wood, J. Hardy, A. Lees, et R. de Silva (2005). Linkage disequilibrium fine mapping and haplotype association analysis of the tau gene in progressive supranuclear palsy and corticobasal degeneration. *Journal of Medical Genetics 42*, 837–846.

Schaid, D. J. (2004). Evaluating association of haplotypes with traits. *Genetic Epidemiology 27*, 348–364.

Stram, D. O. (2004). Tag snp selection for association studies. *Genetic Epidemiology 27*, 365–374.

Zhang, N. L. et T. Kocka (2004). Efficient learning of hierarchical latent class models. In *Proceedings of the 16th IEEE ICTAI*, pp. 585–593.

Summary

Discover the genetic basis of common genetic diseases represents a public health issue. However this task presents several difficulties such as high data dimensionality and identification of the causal mutations. For this purpose, the modelling of dependencies between genetic markers using hierarchical bayesian networks offers several possibilities: data dimension reduction through latent variables and identification of causal markers through the conditional independence property.