Optimisation par colonie de fourmis pour la sélection de variables par construction stochastique de couverture de Markov - Application pour la médecine de précision

Clément Niel[1], Christine Sinoquet[1]

[1] LS2N, UMR CNRS 6004, Université de Nantes, 2 rue de la Houssinière, BP 92208, 44322 Nantes, France
{clement.niel,christine.sinoquet}@univ-nantes.fr

Mots-clés : *métaheuristique, sélection de variables, couverture de Markov, optimisation, colonie de fourmis, médecine de précision*

Introduction

L'usage intensif des technologies de génotypage à haut débit a ouvert l'ère de la médecine de précision. Dans ce contexte, les études d'association génétique s'appuient sur les données de génotypage pour générer des connaissances fines sur les liens entre génotypes et pathologie complexe (ou phénotype). L'épistasie caractérise la situation où une combinaison de marqueurs génétiques (i.e. un *pattern* d'épistasie) détermine le phénotype, alors que chacun d'entre eux présente un effet individuel nul ou faible sur la pathologie étudiée. Cette situation met en échec les méthodes classiques utilisées par les études d'association génétique. Nous apportons dans cet article une contribution originale, complémentaire aux approches proposées dans la littérature [1].

Nous énonçons le problème de la découverte de *patterns* d'épistasie comme un problème de sélection de variables, et proposons de le résoudre au moyen d'une stratégie d'apprentissage de couvertures de Markov (ACM). Nous présentons une première approche innovante, SMMB (Multiple Stochastic Markov blankets), qui combine une stratégie ACM et une approche ensembliste, pour traiter des données imparfaites d'une part (i.e. sans hypothèses sur celles-ci), et sans hypothèse sur le nombre de variables impliquées dans le *pattern* d'épistasie à découvrir. Nous présentons ensuite la variante SMMB-ACO, qui intègre une stratégie d'optimisation de type colonie de fourmis. Nous comparons SMMB et SMMB-ACO avec trois autres méthodes, sur données simulées et réelles. Nous montrons que l'amélioration, par optimisation de type colonie de fourmis, des performances de SMMB, déjà bien classée, fait de SMMB-ACO une approche prometteuse.

Cadre méthodologique

La méthode SMMB s'inscrit dans plusieurs cadres méthodologiques : réseaux Bayésiens, apprentissage statistique et optimisation combinatoire. Reposant sur le concept de couverture de Markov, la méthode SMMB utilise la diversification et la robustesse apportées par les méthodes d'apprentissage ensemblistes.

Dans un réseau Bayésien construit sur un ensemble de variables V, la couverture de Markov (CM) d'une variable cible T ($T \notin V$), $M(T)$, est un ensemble minimal de variables qui rend T indépendante de toutes les autres variables de V : $\forall X \in V \setminus M(T),\ X \perp\!\!\!\perp T \mid M(T)$ (toute variable hors de $M(T)$ est indépendante de T, conditionnellement à $M(T)$). Dans la littérature, plusieurs variantes de l'algorithme pionnier IAMB [2] ont été proposées, pour tenter d'apprendre efficacement une CM optimale à partir d'un ensemble vide. Les variations portent sur la conception et l'intrication d'une phase "forward", qui incorpore des variables candidates dans la CM en cours de construction, et d'une phase "backward", destinée à en éliminer les faux positifs. Le test d'indépendance conditionnelle mentionné ci-dessus est un ingrédient essentiel de ces méthodes.

Algorithmes SMMB et SMMB-ACO

L'algorithme SMMB est une approche originale ; c'est la première variation stochastique proposée pour le calcul d'une couverture de Markov : de multiples couvertures sont calculées sur des versions perturbées du jeu de données, obtenues par échantillonnage des variables. Une couverture "consensus" est ensuite construite. Pour assurer le passage à l'échelle requis par les études d'association génétique, et ne pas biaiser l'apprentissage des CM en y incorporant une variable fortement dépendante avec le

phénotype, la méthode SMMB se démarque des autres approches en ajoutant des groupes de variables à la CM en cours de construction, au lieu d'ajouter les variables une à une.

L'algorithme SMMB-ACO permet de restreindre encore l'espace de recherche des combinaisons de variables explorées. Dans ce cadre, un échantillon de variables est assigné à chacune des fourmis, qui en apprend une couverture de Markov. A tous les niveaux de l'algorithme SMMB, l'échantillonnage des variables est réalisé selon une loi uniforme. Dans SMMB-ACO, l'échantillonnage est guidé au moyen d'une distribution de probabilité, calculée grâce à une stratégie d'optimisation de type colonie de fourmis.

Evaluation et résultats

Les algorithmes SMMB et SMMB-ACO sont implémentés en C++, et parallélisés grâce à la librairie OpenMP. Les méthodes SMMB et SMMB-ACO ont été comparées à trois autres méthodes, BEAM [3], DASSO-MB [4] et AntEpiSeeker [5], sur des jeux de données simulées et réelles. Comme SMMB, BEAM repose également sur un cadre Bayésien, mais met en œuvre une méthode Monte-Carlo par chaîne de Markov. DASSO-MB réalise l'apprentissage déterministe d'une unique couverture de Markov. AntEpiSeeker utilise en préliminaire l'adaptation directe d'un algorithme d'optimisation par colonie de fourmis, pour détecter des ensembles de variables statistiquement associés avec la variable cible ; puis, sous l'hypothèse que le *pattern* d'épistasie à découvrir implique *r* variables, tous les sous-ensembles de taille *r* de chacun de ces ensembles sont ensuite examinés exhaustivement.

SMMB requiert la spécification de 6 paramètres. SMMB-ACO utilise 6 autres paramètres supplémentaires. Ces paramètres, la méthode suivie pour leur réglage, ainsi que l'indication des valeurs spécifiées seront décrites dans une version étendue.

SMMB-ACO mise à part, SMMB est la meilleure méthode dans 50% des 12 conditions simulées (100 jeux d'essais simulés pour chaque condition), ou bien elle est classée dans les deux meilleures pour 4 autres conditions. SMMB-ACO améliore les performances de SMMB dans 100% des cas, et se classe meilleure méthode dans 7 conditions sur 12. Sur données réelles (23 chromosomes présentant de 5754 à 38 867 variables, et génome entier comportant 469 612 variables), SMMB-ACO améliore également les performances de SMMB. Pour le génome entier, au moins 6 des 7 *patterns* d'épistasie publiés par ailleurs sont respectivement identifiés par 100%, 90%, 80% et 70% des 10 exécutions réalisées pour BEAM, SMMB-ACO, AntEpiSeeker et SMMB. Les performances de SMMB-ACO sont très proches de celles de la meilleure méthode (BEAM), qui est de 2.8 à 4.5 fois plus lente (e.g. 53h versus 19h sur le génome entier), et SMMB-ACO est la méthode la plus rapide après la méthode déterministe DASSO-MB, peu performante (e.g. 17h versus 19h sur le génome entier), qui ne retrouve que 5 *patterns*.

Conclusion et perspectives

Nos travaux ont montré que l'intégration d'une stratégie inspirée de l'optimisation par colonie de fourmis à l'algorithme SMMB améliore encore la performance de ce dernier, et en fait une méthode compétitive. De plus, à l'échelle d'un génome complet, la méthode SMMB-ACO s'est révélée significativement plus rapide que la meilleure méthode, et de performance proche de cette dernière. Nos prochains travaux s'attacheront à diminuer encore la complexité temporelle de SMMB-ACO, tout en permettant de générer plusieurs *patterns* d'épistasie optimaux différents s'il en existe.

Références

[1] C. Niel, C. Sinoquet, C. Dina, and G. Rocheleau. A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, 6 :285, 2015.

[2] I. Tsamardinos, C. F. Aliferis, and A. R. Statnikov. Algorithms for large scale Markov blanket discovery. In *proceedings of the 16th International FLAIRS Conference*, 376–380, 2003.

[3] Y. Zhang and J. S. Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39 :1167–1173, 2007.

[4] B. Han, M. Park, and X.-W. Chen. A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinformatics*, 11(Suppl.3) :S5, 2010.

[5] Y.Wang, X. Liu, K. Robbins, and R. Rekaya. AntEpiSeeker : detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. *BMC Research Notes*, 3 :117, 2010.