Séléction de tagSNP : une approche PLNE

David Allouche, Simon de Givry, Thomas Schiex

MIA INRA, 31320 Castanet Tolosan, {allouche,degivry,tschiex}@toulouse.inra.fr

Mots-Clés : *TagSNP, programmation linéaire en nombre entier, set covering.*

1 Introduction

Le problème de la sélection de tagSNP apparaît en génétique et analyse du polymorphisme. Un SNP (Single Nucleotide Polymorphism) est une variation ponctuelle dans le génome d'individus d'une même espèce. Il se caractérise par une altération d'un seul nucléotide (A,T,C, ou G) dans la séquence. Sur les 4 milliards de nucléotides que compte le génome humain, on dénombre environ 10 millions de SNP. Ils expliquent jusqu'à 90% des variations génétiques humaines. Bien que les avancées technologiques aient permis de diminuer le coût, le criblage de la totalité des SNP d'un individu ou d'une population n'est pas envisageable économiquement. La sélection de tagSNP contourne ce problème. Il s'agit d'une forme de compression d'information avec perte consistant à extraire un sous-ensemble de SNP tel que les SNP sélectionnés (appelés tagSNP) capturent l'essentiel de l'information génétique, afin de les utiliser pour le criblage et l'analyse statistique d'une population de grande taille [2]. L'approche consiste dans un premier temps à mesurer la corrélation entre paire de SNP déterminée dans une petite population. Un tagSNP pourra être considéré comme *représentatif* d'un autre SNP si les deux SNP sont suffisamment corrélés dans l'échantillonnage de départ. Cette relation est capturée par la mesure r^2 entre deux SNP, qui en pratique devra être supérieure à un seuil θ (souvent fixé à $\theta = 0.8$) pour être considérée comme significative [1].

2 Présentation du problème

Nous considérons un graphe non orienté $G = (V, E)$ dans lequel chaque sommet $u \in V$ est un SNP et les arêtes $(u, v) \in E$ sont pondérées par la mesure r^2 entre paires de SNP, seules les arêtes dont la pondération est supérieure à θ étant retenues. Le problème tagSNP se réduit alors à un problème de couverture d'ensembles (set covering, NP-dur). En pratique, le nombre de solutions optimales peut être très important et les outils spécialisés du domaine tel que FESTA [3] optimisent, via des approches incomplètes, en plus du nombre de tagSNP, des critères secondaires : entre deux tagSNP, une mesure r^2 faible est préférée afin de maximiser la dispersion des tagSNP ; entre un non-tagSNP et un de ses représentants, une mesure r^2 élevée est préférée pour maximiser la représentativité des tagSNP.

3 Modèle PLNE 0/1

Soit $G = (V, E)$ une instance, à chaque sommet $u \in V$, une variable de décision booléenne x_u est introduite et vaut 1 si le SNP est sélectioné comme tagSNP. A chaque arête $(u, v) \in E$ est associée trois variables de décision booléenne : $R_{u,v} = 1$ si le SNP u est représenté par v ; et vice-versa pour

$R_{v,u} = 1$; $D_{u,v} = 1$ si u et v sont sélectionés comme tagSNP. Concernant la fonction objective (1), une fonction linéaire sur les variables x_u avec un coût élémentaire M si la variable est à 1. Le critère de représentativité associe à chaque paire de variables $R_{u,v}$ et $R_{v,u}$ un coût non nul symétrique $C_{rep}(u,v)$ égal à $\lfloor 100.\frac{1-r^2_{u,v}}{1-\theta} \rfloor$. Enfin, pour capturer la dispersion entre tagSNP liés dans le graphe, un coût $C_{dis}(u,v)$ égal à $\lfloor 100.\frac{r^2_{u,v}-\theta}{1-\theta} \rfloor$ est ajouté si $D_{u,v} = 1$. Afin de garder à ces critères leur caractère secondaire, nous utilisons simplement une valeur de coût M (le coût de sélection d'un SNP) supérieur à la somme des critères secondaires dans l'instance considérée.

$$Min \quad \sum_{u \in V} M.x_u + \sum_{u,v \in E} (C_{rep}(u,v).(R_{u,v} + R_{v,u}) + C_{dis}(u,v).D_{u,v}) \tag{1}$$

$$\begin{aligned}
Subject\ to: \qquad & x_u + \sum_{(u,v) \in E} x_v \geq 1 \quad \forall u \in V && (2)\\
& x_u + \sum_{(u,v) \in E} R_{u,v} = 1 \quad \forall u \in V && (3)\\
& x_v \geq R_{u,v} \text{ et } x_u \geq R_{v,u} \quad \forall (u,v) \in E && (4)\\
& x_u + x_v \leq D_{u,v} + 1 \text{ et } D_{u,v} \leq x_u \text{ et } D_{u,v} \leq x_v \quad \forall (u,v) \in E && (5)
\end{aligned}$$

Les contraintes (2) impliquent qu'un SNP doit nécessairement être tagSNP ou représenté par au moins un de ses voisins. De plus, chaque SNP u est soit un tagSNP, soit il a un *meilleur* représentant v indiqué par $R_{u,v} = 1$ (3) et dans ce cas v doit être un tagSNP (4) ; et pour la dispersion, la sélection d'une paire de tagSNP $(u,v) \in E$ équivaut à $D_{u,v} = 1$ (5).

4 Conclusion

Le modèle ci-dessus a été expérimenté avec `cplex v11` sur les instances dérivées de données obtenues sur le chromosome humain numéro 1, fournies par Steve Qin [3]. Deux valeurs de seuil, $\theta = 0.8$ et 0.5 ont été essayées. Ces instances sont relativement faciles. En 3 minutes, leur résolution à l'optimalité diminue de 21% à 3% le nombre de tagSNP sélectionnés par rapport à deux heuristiques proposées par FESTA [3]. Les résultats obtenus par l'approche PLNE surpasse également ceux obtenus dans le cadre des CSP pondérés [4] via le solver `toulbar2`. Pour $\theta = 0.8$, ce dernier résoud optimalement toutes les instances, bien que 50 fois moins rapide que `cplex`. Néanmoins il constitue une alternative intéressante pour les utilisateurs ne désirant pas investir dans un logiciel commercial.

Références

[1] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, 74(1) :106–120, 2004.

[2] J.N. Hirschhorn and M.J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2) :95–108, 2005.

[3] Z. S. Qin, S. Gopalakrishnan, and G. R. Abecasis. An efficient comprehensive search algorithm for tagsnp selection using linkage disequilibrium criteria. *Bioinformatics*, 22(2) :220–225, 2006.

[4] Marti Sanchez, David Allouche, Simon de Givry, and Thomas Schiex. Russian doll search with tree decomposition. *ijcaii*, pages 603–608, 2009.