Estimation du risque génétique des maladies complexes basée sur les haplotypes

Félix Balazard [1] [2]

[1] *Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie, 4 Place Jussieu, 75005 Paris*

[2] *Inserm U1169, Hôpital Bicêtre, 78 Rue du Général Leclerc, 94270 Le Kremlin-Bicêtre*
felix.balazard@inserm.fr

Résumé. Les études d'association à l'échelle du génome ont découvert des milliers d'associations entre variants génétiques et des maladies. En utilisant les mêmes données, On peut essayer de prédire le risque de maladie. L'information de phase est une structure biologique importante qui a rarement été utilisée. Nous proposons ici une méthode d'apprentissage statistique en plusieurs étapes qui tente d'utiliser cette information. Elle capture des interactions locales dans des haplotypes courts and combine linéairement les résultats. Nous montrons qu'elle obtient de meilleurs résultats. Toutefois, une variation de notre méthode qui n'utilise pas l'information de phase obtient des performances similaires. Le code source est disponible à l'adresse https://github.com/FelBalazard/Prediction-with-Haplotypes .

Mots-clés. Génétique, apprentissage supervisé, GWAS

Abstract. Genome-wide association studies (GWAS) have uncovered thousands of associations between genetic variants and diseases. Using the same datasets, prediction of disease risk can be attempted. Phase information is an important biological structure that has seldom been used in that setting. We propose here a multi-step machine learning method that aims at using this information. It captures local interactions in short haplotypes and combines linearly the results. We show that it outperforms standard linear models on some GWAS datasets. However, a variation of our method that does not use phase information obtains similar performance. Source code is available at https://github.com/FelBalazard/Prediction-with-Haplotypes .

Keywords. Genetics, supervised learning, GWAS, bioinformatics

1 Les études d'association à l'échelle du génome

Les études d'association à l'échelle du génome ou GWAS (Genome Wide Association Studies) ont consisté à génotyper des milliers de patients et de contrôles sains et à les comparer. Ceci a permis d'établir des milliers d'associations entre SNPs (Single Nucleotide Polymorphism ou polymorphisme d'un unique nucléotide) et maladies.

La première grande étude de ce type a été mené par le Wellcome Trust Case Control Consortium (2007) : 2000 patients pour chacune de 7 maladies différentes, 3000 contrôles partagées et 500000 SNPs répartis sur l'ensemble du génome. Nous illustrerons notre méthode sur ces données.

2 La prédiction du risque génétique

Des méthodes classiques d'apprentissage supervisée ont déjà été appliquées à ce genre de données. Une étape de présélection est souvent nécessaire pour des raisons computationelles. Une machine à vecteurs de support obtient ainsi un AUC de 0,83 dans un jeu de données indépendant pour le diabète de type 1 dans Wei et al. (2009). Dans un très grand jeu de données, la régression lasso obtient un AUC de 0,85 pour la maladie de Crohn dans Wei et al. (2013).

Les données génétiques ont une structure particulière qu'il est intéressant d'utiliser afin d'améliorer les performances prédictives. Ainsi Botta et al. (2014) proposent une variation de l'algorithme des forêts aléatoires afin de prendre en compte la distance sur les chromosomes. Ils obtiennent d'excellents résultats qui n'ont malheureusement pas été répliqués dans une nouvelle cohorte.

3 L'information de phase

L'information additionnelle que nous considérerons dans cette présentation est l'information de phase : les allèles mineurs de deux SNPs hétérozygotes adjacents sont-ils sur le même chromosome ou sur le chromosome homologue ? La séquence sur chaque chromosome de la paire est appelée un haplotype. Ceci est illustrée par la figure 1.

Une fois l'information de phase récupérée, nous avons donc deux fois les mêmes variables avec des valeurs différentes. De plus, l'information que nous pouvons espérer récupérer sont des interactions à l'intérieur de chaque haplotype. Notre approche a donc été de traiter chaque haplotype comme une observation distincte et de capturer des interactions locales grâce une méthode d'apprentissage non-linéaire, les forêts aléatoires. Ceci est fait pour un grand nombre d'endroits du génome associés, même faiblement, à la maladie. Les résultats sont ensuite combinés linéairement en utilisant la régression lasso.

Bibliographie

[1] Wellcome Trust Case Control Consortium (2007) *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls*, Nature, Londres
[2] Wei Z. , Kang W. et al (2009) *From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes*, PLoS Genetics,

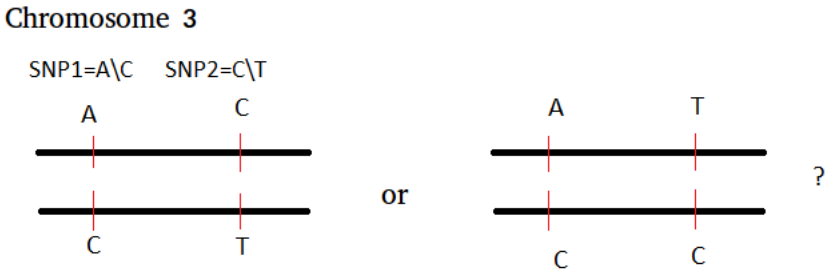

Figure 1: Les valeurs des deux SNPs ne permettent pas de distinguer entre les deux paires d'haplotypes possibles.

San Francisco
[3] Wei Z. et al et International IBD Genetics Consortium (2013) *Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease*, The American Journal of Human Genetics, Philadelphia
[4] Botta V. , Louppe G. , Geurts P. et Wehenkel L. (2014) *Exploiting SNP Correlations within Random Forest for Genome-Wide Association Studies*, PLoS One, San Francisco