Détection d'outliers à partir de l'analyse en composantes principales : application en génétique des populations.

Keurcien Luu [1] & Michael G. B. Blum [2]

[1] *Laboratoire TIMC-IMAG UMR 5525, Université Grenoble-Alpes, CNRS, keurcien.luu@imag.fr*
[2] *Laboratoire TIMC-IMAG UMR 5525, Université Grenoble-Alpes, CNRS, michael.blum@imag.fr*

Résumé. Notre objectif est de détecter à partir de données génomiques quels sont les gènes qui permettent aux individus de s'adapter à leur environnement. Compte tenu du volume important des données de séquençage, le défi computationnel est de proposer une méthode statistique adaptée au volume des données. Nous proposons une approche rapide, basée sur l'analyse en composantes principales. Le principe est de considérer comme gènes candidats ceux qui sont excessivement corrélés avec les composantes principales. Pour ce faire, nous calculons pour chaque marqueur génétique un vecteur qui mesure l'association entre un marqueur génétique et les composantes principales et nous utilisons la distance de Mahalanobis pour trouver les vecteurs atypiques. En utilisant un jeux de données humains comprenant un peu plus d'un millier d'individus et des centaines de milliers de marqueurs génétiques, nous montrons que cette approche permet de détecter un exemple bien connu d'adaptation biologique chez l'homme.

Mots-clés. génétique des populations, détection d'outliers, distance de Mahalanobis, analyse en composantes principales.

Abstract. Our goal is to use genomic data to detect which genes are involved in biological adaptation to local environment. Given the large volume of sequencing data, the computational challenge is to propose a statistical method that scales to the size of data. We propose a fast approach for detecting selection based on principal component analysis. The principle is to consider as candidates the genes that are excessively related with the principal components. To detect these genes, we compute, for each genetic marker, a vector that measures the association between this genetic marker and principal components. We then use the Mahalanobis distance to find outlier genes. Using a human dataset including more than a thousand of people and hundreds of thousands of genetic markers, we show that this approach can detect a well known example of biological adaptation in humans.

Keywords. population genetics, outlier detection, Mahalanobis distance, principal component analysis.

1 Introduction

Le séquençage à bas coût rend possible le séquençage massif d'individus dans beaucoup d'espèces. Les données génomiques obtenues au sein de populations d'une même espèce permettent la détection des gènes qui confèrent aux individus la capacité de s'adapter à leur environnement. La détection de gènes sous sélection repose principalement sur le principe que la différenciation génétique (distance génétique) entre populations est plus grande pour les gènes sous sélection que pour les autres dits neutres. En effet, l'augmentation rapide de la fréquence d'un allèle au sein d'une population peut se produire s'il confère un avantage sélectif (résistance à une maladie, capacité à vivre en altitude, ...) face aux pressions environnementales. Cette approche dite « outlier approach » revient à considérer comme gènes candidats ceux qui présentent un excès de différences de fréquences d'allèles entre les populations [1]. Un défaut de cette approche est de regrouper au préalable les individus dans des populations définies. Par exemple, dans le cas du jeu de données que nous allons analyser, il n'existe pas de clusters d'individus bien définis ce qui rend difficile de regrouper des individus dans des populations. Récemment des méthodes basées sur l'Analyse en Composantes Principales (ACP) s'affranchissant de cette contrainte ont été proposées ([2,3]). Nous présentons ici une nouvelle approche basée sur l'ACP.

2 Méthodes

Nous supposons dans la suite que les marqueurs génétiques sont bi-alléliques, c'est-à-dire que pour chaque marqueur génétique, seuls deux allèles existent. Pour un individu et un locus donné (localisation sur le génome), un marqueur génétique peut donc être codé par un 0, un 1 ou un 2, selon que le chromosome porte zéro, une ou deux mutations sur l'ensemble des deux copies. Ces marqueurs bialléliques sont appelés SNP (Single Nucleotide Polymorphism) en anglais. Nous construisons alors la matrice G de génotypes à partir des génotypes de chaque individu constituant le jeu de données, en disposant les individus en lignes, et les SNP en colonnes. Cette matrice comporte typiquement de $n = 100$ à $n = 10,000$ individus tandis que le nombre de SNPs p peut être beaucoup plus grand. Une première étape consiste à normaliser la matrice des génotypes, SNP par SNP

$$\tilde{G}_{ij} = \frac{G_{ij} - p_j}{\sqrt{2 \times p_j(1-p_j)}},$$

où G_{ij} est le génotype du i-ème individu au j-ème SNP, et p_j désigne la fréquence allélique de l'allèle de référence du j-ème SNP $p_j = \sum_i G_{ij}/(2n)$. Cette normalisation est justifiée par une approximation binomiale pour les G_{ij} [4]. A partir de la matrice de génotypes normalisés, on réalise une analyse en composantes principales, qui permet de réduire la

dimension des données et de ne garder que les K axes principaux. On note la valeur des scores pour ces K axes $X_1, \ldots, X_K$. Ces scores mesurent la structure des populations. Nous effectuons ensuite une régression linéaire multiple pour chacun des marqueurs $j = 1, \ldots, p$:

$$G_j = \sum_{k=1}^{K} \beta_k^j X_k + \epsilon_j$$

où G_j est le vecteur des génotypes pour le j-ème SNP, β_k^j est le coefficient de régression correspondant au j-ème SNP et à la k-ème composante principale et ϵ_j est le vecteur des résidus. Chaque coefficient de régression correspond à un facteur multiplicatif près à la corrélation entre le j-ème marqueur et la k-ème composante principale. Les coefficients de régression sont ensuite normalisés en prenant en compte la variance des résidus de sorte à obtenir K z-scores $z = (z_1, \ldots, z_K)$ pour chacun des p marqueurs génétiques.

Il est possible de tester si les marqueurs sont outliers pour chacune des composantes principales [2]. L'approche composante par composante peut simplifier l'interprétation des résultats. En revanche, un SNP peut ne pas être considéré comme un outlier pour chaque axe, et cependant être un outlier pour la distribution jointe des K z-scores.

Afin de prendre en compte la distribution jointe des K z-scores, nous utilisons une approche classique d'analyse multivariée pour détecter les outliers. La statistique de test pour détecter les outliers est la distance D de Mahalanobis définie par

$$D^2 = (z - \bar{z})^T \Sigma^{-1} (z - \bar{z}),$$

où Σ est la matrice $(K \times K)$ de covariance des z-scores et $\bar{z}$ est le vecteur des K moyennes des z-scores. La matrice de variance-covariance est estimée avec l'estimateur robuste de Gnanadesikan-Kettenring orthogonalisée [5].

A un facteur multiplicatif λ près, la distance de Mahalanobis est distribuée suivant une loi du χ^2 à K degré de liberté $D^2/\lambda \sim \chi^2_K$. Le paramètre λ, appelé facteur d'inflation génomique, est estimé en divisant la médiane observée des distance par la médiane attendue pour une loi du χ^2 à K degrés de liberté [6]. Une fois calculé les p-valeurs, nous utilisons le calcul des q-valeurs pour contrôler le taux de fausses découvertes [7].

3 Résultats

Nous avons analysé des données de SNPs collectés chez des individus européens dont les 4 grand parents proviennent de la même région géographique que l'individu échantilloné [8]. Les données contiennent 447245 SNPs obtenus chez 1385 individus. La structure de populations est captée par les deux premières composantes principales. La première composante principale capte un gradient sud-nord qui est la direction suivant laquelle la différenciation varie le plus vite en Europe [8,9]. Pour ces données, il n'y a pas de clusters d'individus bien définis rendant l'approche basée sur l'ACP particulièrement utile. En

utilisant un seuil de taux de fausses découvertes à 1%, on trouve parmi les gènes candidats le gène LCT qui permet de dégrader le lactose à l'âge adulte. Ce gène est bien connu comme exemple d'adaptation biologique [10]. L'allèle qui confère la possibilité de digérer le lactose à l'âge adulte varie suivant un gradient sud-nord ; il est plus présent en Europe du Nord dans les populations qui ont une forte tradition d'élevage. Le fait que nous retrouvions comme candidat cet exemple archétypal de l'adaptation biologique chez l'homme confirme l'intérêt de l'approche basée sur l'analyse en composantes principales.

Bibliographie

[1] Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics : from genotyping to genome typing. Nature reviews genetics, 4(12), 981-994.
[2] Galinsky, K. J., Bhatia, G., Loh, P. R., Georgiev, S., Mukherjee, S., Patterson, N. J., & Price, A. L. (2015). Fast principal components analysis reveals independent evolution of ADH1B gene in Europe and East Asia. bioRxiv, 018143.
[3] Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E., & Blum, M. G. (2016). Detecting genomic signatures of natural selection with principal component analysis : application to the 1000 Genomes data. Molecular biology and evolution, msv334.
[4] Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. PLoS genet, 2(12), e190.
[5] Maronna, R. A., & Zamar, R. H. (2012). Robust estimates of location and dispersion for high-dimensional datasets. Technometrics.
[6] Devlin, B., & Roeder, K. (1999). Genomic control for association studies. Biometrics, 55(4), 997-1004.
[7] Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100(16), 9440-9445.
[8] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., ... & Stephens, M. (2008). Genes mirror geography within Europe. Nature, 456(7218), 98-101.
[9] Jay, F., Sjödin, P., Jakobsson, M., & Blum, M. G. (2013). Anisotropic isolation by distance : the main orientations of human genetic differentiation. Molecular biology and evolution, 30(3), 513-525.
[10] Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., ... & Ibrahim, M. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nature genetics, 39(1), 31-40.