Feature selection for biomarker discovery: a novel graph-guided approach to classification

Vivien Goepp and Chloé-Agathe Azencott

CBIO, Mines ParisTech & Institut Curie

Internship project

Detecting interaction between features in genome-wide association studies

Genome-wide association studies (GWAS) are a type of genetic study which have become popular for discovering genomic regions associated to complex diseases like cancer or diabetes. They consist in genotyping a large number of single-nucleotide mutations (or *single nucleotid polymorphisms*, SNPs) in patients and healthy individuals. The mapping between genotype and phenotype is done using classification: the SNPs estimated as relevant correspond to disease-causing mutations. However, biological discoveries made using GWAS have stalled in the past years.

We believe this is due to the inapropriate statistical tools used to detect the causal mutations. Indeed, genomic mutations are believed to have strong interaction effects, while the traditional association test only considers the main effects. Estimating these (potentially high-order) interaction effects is key to detecting new biologically relevant mutations.

Several approaches have been taken to detect interactions (Yung et al., 2011; Kam-Thong et al., 2012), but they suffer limitations due to the sheer computational burden. This internship project offers to develop a new approach.

Set covering machine

This internship aims at adapting a classification method called the set covering machine (Marchand and Shawe-Taylor, 2002) to the setting of feature selection in GWAS data. The set covering machine (SCM) algorithm performs classification of binary data. It learns a classification function as a boolean conjunction of the binary features. (In GWAS, the binary feature is the presence/absence of mutation at a given SNP.)

In the case where there are many features, but only a small subset of those have an effect on the phenotype, this method is shown to perform particularly well. It has been successfully applied to a setting close to GWAS data, for detecting mutations linked to antibiotic resistance (Drouin et al., 2016).

The goal of this internship is to extent this method and to apply it to GWAS data. More precisely, we wish to include *a priori* information about the *proximity* between the SNPs, and to modify the SCM so as to favor choosing SNPs that are closely connected. This *a priori* information is provided through a graph whose nodes are the SNPs and whose edges represent a biological link between the SNPs (for instance, if they are included in the same gene, or if they code for proteins that are in biological interaction). Such networks have been used sucessfully to guide SNP detection (Azencott et al., 2013; Climente-Gonzalez et al., 2020).

Goals of the internship

- Understand the publicly available implementation of SCM[1].
- Extend the implementation to include a graph-guided penalty.
- Apply the method to simulated data and compare it to state-of-the-art methods.

Profile of the candidate

- Student in first or second year of a master in mathematics, computer science, bioinformatics, or any related field
- Solid computer science skills and proficiency in python.
- Interest in genomics and computational biology

Information

- The student will work under the supervision of Vivien Goepp and Chloé-Agathe Azencott. Depending on the sanitary restrictions, the intern would work remotely or in the lab.
- Internship is scheduled to start in February 2021, but preferences can be discussed.
- Duration of the internship: 5-6 months.
- Internship compensation: ∼ 550 €/month .
- To apply, send an email with CV and motivation letter to `vivien.goepp@mines-paristech.fr` and `chloe-agathe.azencott@mines-paristech.fr.`. Please contact us if you have any question regarding the internship.

References

C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, and K. M. Borgwardt. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29 (13):i171–i179, July 2013. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btt238.

H. Climente-Gonzalez, C. Lonjou, F. Lesueur, D. Stoppa-Lyonnet, N. Andrieu, and C.-A. Azencott. Biological networks and GWAS: Comparing and combining network methods to understand the genetics of familial breast cancer susceptibility in the GENESIS study. *biorxiv preprint*, 2020.

A. Drouin, S. Giguère, M. Déraspe, M. Marchand, M. Tyers, V. G. Loo, A.-M. Bourgault, F. Laviolette, and J. Corbeil. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*, 17 (1):754, Dec. 2016. ISSN 1471-2164. doi: 10.1186/s12864-016-2889-6.

T. Kam-Thong, C.-A. Azencott, L. Cayton, B. Pütz, A. Altmann, N. Karbalai, P. G. Sämann, B. Schölkopf, B. Müller-Myhsok, and K. M. Borgwardt. GLIDE: GPU-Based Linear Regression for Detection of Epistasis. *Human Heredity*, 73(4):220–236, 2012. ISSN 0001-5652, 1423-0062. doi: 10.1159/000341885.

M. Marchand and J. Shawe-Taylor. The set covering machine. *Journal of Machine Learning Research*, 3(Dec):723–746, 2002.

L. S. Yung, C. Yang, X. Wan, and W. Yu. GBOOST: A GPU-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics*, 27(9): 1309–1310, 2011.

[1] `https://github.com/aldro61/kover`