Postdoctoral Position – Multi-locus genome-wide association studies
ANR project SCAPHE

Context

Differences in how patients experience disease can be explained in great part by their genomic differences. Enabling precision medicine hence requires identifying genomic features associated with disease risk, prognosis or response to treatment. This is often achieved using genome-wide association studies (GWAS), which look for associations between single nucleotide polymorphisms (SNPs) and an observable trait (phenotype). However, for many complex traits, the SNPs they uncover account for little of the known heritable variation, a phenomenon referred to as the "missing heritability" problem.

ANR project SCAPHE ("Methods for discovering SNP Combinations Associated with a PHEnotype") builds on the hypothesis that this is due to the effect of non-additive interactions between SNPs, together with a lack of robustness stemming from the relatively small sample sizes. This last issue can be alleviated by integrating biological networks to GWAS. SCAPHE proposes to develop *novel machine learning algorithms for GWAS*, integrating biological networks and modeling non-additive SNP effects, to robustly detect SNP combinations associated with a phenotype.

This postdoctoral position will be funded for up to 24 months as part of SCAPHE. The starting date is flexible, but should be no later than December 2019. The project will take place under the supervision of Chloé-Agathe Azencott (http://cazencott.info).

Research topic

Among the causes for missing heritability, the failure to account for *joint effects* between multiple loci (epistasis) has garnered interest in recent years. Addressing this issue, however, requires solving the statistical difficulties posed by the wide gap between the number of features that can be measured (hundreds of thousands) and that of samples for which they can be collected (a few thousands). One way to address this problem is to reduce the dimensionality of the space of solutions by means of structural constraints. Those can in particular be given by biological networks. Several methods have been developed to that end in recent years [Azencott et al., 2013; Azencott, 2016], but only account for individual or additive effects between features.

While most epistasis detection methods focus on quadratic effects between pair of loci, we are interested here in more complex effects, also known as *higher-order interactions.* Several methods have been proposed in recent years to address this question [Suzumura et al., 2017; Llinares-López et al., 2018; Drouin et al., 2019].

The goal of this postdoctoral project is to integrate network information to methods for higher-order interaction detection in GWAS data. Relevant research questions for this postdoctoral project include:

- the applicability of the approach proposed by Drouin et al. (2019) to human GWAS data;
- robustness/stability (i.e. recovering the same features with slightly different samples);
- statistical significance or false discovery rate control for the selected features.

The postdoctoral researcher is, however, encouraged to develop their own line of research within the provided context.

Lab

The project will take place in the Centre for Computational Biology (CBIO — http://cbio.ensmp.fr), a joint laboratory between Mines ParisTech, one of the most prominent French engineering schools, and Institut Curie, a major hospital and research facility dedicated to cancer. CBIO benefits from an exceptional scientific environment with immediate access to experts and collaborators in biology and medicine, enabling a stimulating interdisciplinary exchange. The laboratory is located in the centre of Paris, both in Mines ParisTech and in the nearby Institut Curie.

Prerequisites

- PhD in machine learning or statistics;
- Proficiency in at least one programming language;
- Motivation to work on genomics and bioinformatics applications. Prior experience in these domains is welcome but not mandatory.

How to apply

Send your CV, a cover letter, and 2 contact references by email at chloe-agathe.azencott@mines-paristech.fr.

Relevant reading:

Azencott, Chloé-Agathe, et al. "Efficient network-guided multi-locus association mapping with graph cuts." *Bioinformatics* 29.13 (2013): i171-i179.

Azencott, Chloé-Agathe. "Network-Guided Biomarker Discovery." *Machine Learning for Health Informatics*. Springer, Cham, 2016. 319-336.

Dernoncourt, David, Blaise Hanczar, and Jean-Daniel Zucker. "Analysis of feature selection stability on high dimension and small sample data." *Computational statistics & data analysis* 71 (2014): 681-693.

Drouin, Alexandre, et al. "Interpretable genotype-to-phenotype classifiers with performance guarantees." *Scientific reports* 9.1 (2019): 4071.

Llinares-López, Felipe, et al. "CASMAP: detection of statistically significant combinations of SNPs in association mapping." *Bioinformatics* (2018).

Suzumura, Shinya, et al. "Selective inference for sparse high-order interaction models." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.