M2 internship - Deep learning to model genetic pleiotropy to understand the human genetic architecture

Keywords

statistical genetics; pleiotropy; complex traits and diseases; post-genomic data; semi-supervised and supervised learning; penalized methods; neural networks

Description

One striking observation today in the field of human genetics is that as Research advances to understand the genetic architecture of complex traits and to apprehend the etiology of heritable diseases, new paradigms keep emerging revealing more and more of the complexity of biological models. One particular concept seems to resurge: pleiotropy. Indeed, humans have a very low number of protein-coding genes given their organism complexity. Organism complexity can thus stem from the combination of fine-tuned expression of a few genes and their extreme level of interconnection rather than a significant increase in gene number. Therefore, pleiotropy, occurring when one genetic element (e.g. variant, gene) has independent effects on several traits, is thought to play a central role in the genetic architecture of human complex (i.e. caused by more than 1 gene) traits and diseases.

We have already developed several statistical methods to study pleiotropy which have gathered much attention in the human genetics community. On the one hand, in Verbanck et al. (2018), we have shown that horizontal pleiotropy which tended to be neglected was found in almost 50% of causal relationships inferred from Mendelian randomization and could lead to false positives and biased causal estimates. On the other hand, we have developed a proof-of-concept paper measuring pleiotropy at the level of variants mainly focused on horizontal pleiotropy, Jordan, Verbanck, and Do (2019), which has confirmed the existence of widespread pleiotropy across the human genome.

On a different note, machine learning and specifically deep learning have shown high performance in the prediction field, for example with image pattern recognition. We want to reroute these deep-learning methods for genetics, and specifically we have the objective to **build a comprehensive framework to provide a genome-wide map of pleiotropy using machine learning.**

The internship will be dedicated to explore semi-supervised and supervised methods to classify the pleiotropy of genetic variants, using labeled pleiotropic data from the methods the team has been developing.

In human genetics, and especially to study pleiotropy, the major issue is to obtain labeled data since the ground truth is unknown. However, it has been shown that semi-supervised learning strategies have already been applied, with high gain in classification performance (Ratsaby and Venkatesh (1995), Cozman, Cohen, and Cirelo (2003)).

Therefore, we have already developed a strategy to partially label genetic variants for pleiotropy using Gaussian Mixture models (Darrous, Mounier, and Kutalik 2021; Morrison et al. 2020). Thus, we will explore this first strategy of developing a **semi-supervised learning framework** in case of Gaussian Mixture models.

A second approach will explore supervised learning, namely **Convolutional Neural Networks** (CNN) that are commonly applied to analyze images. In CNN architecture, the receptive fields overlap with each other and do convolutions between the kernel and the data: this is analogous to the sliding window approach, a traditional method in genetics, with genomic intervals “sliding” across the genome. Furthemore, the block architecture of CNNs is comparable to LD-blocks (dependence structure between alleles), one of the major obstacle of mapping pleiotropy.

The frameworks Keras and/or Tensorflow (reachable through R and Python) make powerful deep learning tools available, and will be mainly used to develop the frameworks.

The successful candidate:

- will have a master of data science linked to statistics or artificial intelligence, candidates with more theoretical background however showing strong interest in life science applications are also welcome;
- will be enthusiastic about transdisciplinary research and open science at the interface between data science and genetics;
- will show a clear interest to use applied science methodology to benefit biological understanding;
- will have good programming skills, preferentially R and/or Python;
- can have a background in biology or genetics;
- should be open-minded and willing to work as a team with other lab members;
- will speak decent English since we are closely collaborating with Mount Sinai Hospital in New York City, USA.

Scientific environment

Starting date: February 2023

The 6-months Master 2 internship will be supervised by Dr Marie Verbanck who is Assistant Professor (Maître de Conférences) at the Faculté de Pharmacie at Université de Paris within the UR 7537 BioSTM unit (Biostatistique, Traitement et Modélisation des données biologiques). BioSTM’s mission is to develop cutting-edge statistical methodologies to answer real-life biological problems with an emphasis on reproducible science and open research.

How to apply

To apply, please send a concise email describing your research interests and experience as well as an up-to-date CV to Marie Verbanck (marie.verbanck@u-paris.fr). Name and contact for references will be appreciated.

References related to the internship

Cozman, Fabio Gagliardi, Ira Cohen, and Marcelo Cesar Cirelo. 2003. “Semi-Supervised Learning of Mixture Models.” *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, 8.

Darrous, Liza, Ninon Mounier, and Zoltán Kutalik. 2021. “Simultaneous Estimation of Bi-Directional Causal Effects and Heritable Confounding from GWAS Summary Statistics.” *Nature Communications* 12 (1): 7274. https://doi.org/10.1038/s41467-021-26970-w.

Jordan, Daniel M., Marie Verbanck, and Ron Do. 2019. “HOPS: A Quantitative Score Reveals Pervasive Horizontal Pleiotropy in Human Genetic Variation Is Driven by Extreme Polygenicity of Human Traits and Diseases.” *Genome Biology* 20 (1): 222. https://doi.org/10.1186/s13059-019-1844-7.

Morrison, Jean, Nicholas Knoblauch, Joseph H. Marcus, Matthew Stephens, and Xin He. 2020. “Mendelian Randomization Accounting for Correlated and Uncorrelated Pleiotropic Effects Using Genome-Wide Summary Statistics.” *Nature Genetics*, May, 1–8. https://doi.org/10.1038/s41588-020-0631-4.

Ratsaby, Joel, and Santosh S. Venkatesh. 1995. “Learning from a Mixture of Labeled and Unlabeled Examples with Parametric Side Information.” In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, 412–17. COLT ’95. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/225298.225348.

Verbanck, Marie, Chia-Yen Chen, Benjamin Neale, and Ron Do. 2018. “Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization Between Complex Traits and Diseases.” *Nature Genetics*, April, 1. https://doi.org/10.1038/s41588-018-0099-7.