Prediction of genetic traits with deep neural networks: how to consider epistasis ?

Keywords: Deep-Learning, biological sytems
Duration: 4 to 6 months
Supervisors: Alexandre Allauzen and Phillipe Nghe (ESPCI)
Contact: allauzen@dauphine.psl.eu

Context

Understanding the effect of mutations is crucial to predict genetic factors in human diseases as well as the evolution of pathogens and drug resistance. However, despite the current accumulation of large biological datasets, we still do not understand how genetic mutations determine biological traits. The main challenge lies in the interactions between mutations, meaning that a single trait is not determined by a single mutation but by a combination of mutations. This phenomenon, called epistasis, is currently studied on a case-by-case basis and begs for artificial intelligence approaches.

Outline

We propose to develop a machine learning approach leveraging recent experimental observations: interactions between mutations are explained to a large extent by non-linear but smooth functions of features (the phenotype). These features provide an intermediate representation of cellular processes. This suggests a two steps process, where the phenotype (features) can be first expressed from the genotype and some knowledge about the environment, and then predict biological traits of interest. This internship focuses on the second step and how to take into account the epistasis phenomenon using deep-learning models.
In the last decades, deep-learning models have shown impressive results on different tasks involving complex and structured data, like sequences and graphs. Beyond convolutional and recurrent architectures, transformers recently achieve new state of the art results for many applications. One goal of this internship is to design a model architecture dedicated to the prediction task. A first goal is to efficiently predict the biological traits and the other is to deal with the epistasis phenomenon and quantify its impact on the model.

In practice

The internship is co-supervised by A. Allauzen (MILES-LAMSADE, Dauphine Université) and P. Nghe (Laboratory of BioChemistry, ESPCI).
The developed approaches will be tested on well-characterized training sets from the Laboratory of BioChemistry, along with other available biological databases, including Escherichia coli bacteria and yeast. Notably, the laboratory has already developed such twostep models of biological interactions using mechanistic modelling of biochemical process, which will allow comparison with machine learning methods to be developed during this internship.
The internship will include a review of the available datasets to assess the knowledges we can extract and use to build the input representation. Then, the choice of neural architectures will focus on the genetic traits prediction and how it relates to the epistasis phenomenon. The experimental setup will rely on the pytorch library (in python) to implement deep-learning model.
If you are interested in, feel free to contact us with a CV.