Project Title: Alternative splicing-inspired protein design

Project Supervisors:
Elodie Laine
Sorbonne Université, Paris
elodie.laine@sorbonne-universite.fr
Hugues Richard
Robert Koch Institute , Berlin - Sorbonne Université, Paris
hugues.richard@sorbonne-universite.fr

Summary:
Eukaryotes have evolved a transcription machinery that can augment the protein repertoire without increasing the genome size. It produces several mRNA transcripts from the same gene, by choosing different initiation/termination sites and/or by splicing different exons. Alternative splicing (AS) concerns almost all multi-exon genes in vertebrates and it has been suggested that two protein isoforms may have completely different cellular partners and may adopt different 3D folds. Hence, the generative potential of this mechanism is fascinating.

In recent years, we have developed a couple of computational methods, ThorAxe [1] and PhyloSofS [2], to assess the impact of AS on protein 3D structures and interaction in evolution. Our methods exploit publicly available gene annotations and RNA-seq data. They construct evolutionary splicing graphs resuming both intra- and inter-species transcript variability, transcripts phylogenetic forests representing plausible evolutionary scenarios leading to present-day transcripts, and 3D structures of transcript isoforms. We have applied them on well-documented genes, among which several therapeutic targets, and we are now scaling up to the whole human genome.

The goal of this project is to develop a probabilistic model that will learn from the functional AS events observed today in nature to generate new protein functional diversity. The main challenge will be to design an artificial system able to learn the underlying *rules* of (functional) AS and to generalise to any protein sequence. We will particularly focus on variational auto-encoders (VAE) and deep neural network-powered autoregressive models (DNNAM), which have proven very powerful to predict the outcomes of mutations and insertions/deletions and design protein sequences with desired properties. The model will learn from the set of AS events detected by ThorAxe on the Human genome scale. To ensure the good quality of the training data, we will rely on evolutionary conservation. The rationale is that the AS-induced variations selected over millions of years of evolution comply with physical and environmental constraints and thus are likely functional.

References:
1. Zea DJ, Laskina S, Richard H and Laine E (*in preparation*) https://github.com/PhyloSofS-Team/thoraxe.
2. Ait-hamlat A, Zea DJ, Labeeuw A, Polit L, Richard H and Laine E (2020) *J Mol Biol* https://github.com/PhyloSofS-Team/PhyloSofS

Expected skills:
The candidate should have: background in computer sciences or in bioinformatics, knowledge in artificial intelligence, high-throughput computing and very good programming skills ; some knowledge in biology (transcriptomics, structural bioinformatics) are a plus. Teamwork skills are essential for the achievement of the project.

Environment:
The candidate will benefit from the inter-disciplinary environment of the Laboratory of Computational and Quantitative Biology (www.lcqb.upmc.fr). The lab is located close to the Seine river, in Paris.

Possibility of funding:
The student will be provided with a monthly stipend of around 550 euros during up to six months.

Applications:
Please send an email containing your CV to Elodie Laine (elodie.laine@sorbonne-universite.fr) and Hugues Richard (hugues.richard@sorbonne-universite.fr).