INTERNSHIP / PhD PROPOSAL

Model-driven Directed Evolution

Biological systems result from Evolution by natural selection. While the principles are known since Darwin, we do not have yet a general quantitative theory of Evolution. One difficulty is that natural Evolution is neither repeatable nor controllable. Our approach is to study fundamental evolutionary questions at the molecular scale, using proteins that we evolve experimentally in a controlled and reproducible way.

In experiments, we typically handle populations of $\sim 10^{12}$ proteins, which we screen to retrieve those that have desired functional properties (binding to a chosen molecule, catalysis of a chemical reaction,….). We then make copies of the selected proteins to reconstitue a large population and introduce mutations to produce new variants, before repeating a new cycle. The technology to perform these steps is well developed and has had many successes (it won a Nobel prize in 2018). Yet, these successes are critically dependent on the choice of the starting point, as we are currently not able to navigate long distances in sequence space.

The basic difficulty is the enormity of this space: given the 20 natural amino acids, even small proteins of size 100 have 20^{100} different variants! Current approaches typically start from a functional protein and introduce new mutations at random to make a population of variants. But most of these variants (sometimes all) show no response to the selective pressure. The population of sequences may also be trapped at local maxima of the fitness landscape. We propose to develop a more efficient approach that uses data derived from the experiments to guide the search towards promising variants.

This approach is based on two recent advances: (i) high-throughput sequencing technology that allows us to read out the composition of our populations of proteins at different stages of evolution; (ii) statistical methods from machine learning and statistical physics that allow us to infer from these large data-sets a model of the relation between sequence and function (the so-called fitness landscape). The idea is to use this model to predict and produce beneficial mutations.

The internship will consist in testing one iteration of this strategy using data that we have already obtained. The focus will be on the inference of a model from the data using methods from statistical physics (inverse Potts models) and machine learning. The predictions made by the model will be checked experimentally in the lab. The candidate will have the opportunity to follow or participate to these experiments.

We expect good skills in statistical physics or/and machine learning and an interest for biological questions. The project will take place in an interdisciplinary team of physicists and biologists working theoretically and experimentally on related projects.

References:

- M. Goldsmith, D. Tawfik (2012). Directed enzyme evolution: beyond the low-hanging fruit. *Current opinion in structural biology 22:* 406.
- S. Boyer, D. Biswas, A. K. Soshee, N. Scaramozzino, C. Nizak, O. Rivoire (2016). Hierarchy and extremes in selections from pools of randomized proteins. *PNAS* 113: 3482.
- S. Schulz, S. Boyer, M. Smerlak, S. Cocco, R. Monasson, C. Nizak, O. Rivoire (2019). Parameters and determinants of responses to selection in antibody libraries. bioRxiv:712539.

Contacts: clement.nizak@espci.fr , olivier.rivoire@college-de-france.fr