INTERNSHIP PROPOSAL

Laboratory name : Gulliver
Supervisors : Olivier Rivoire
e-mails : olivier.rivoire@espci.fr
Web pages : https://www.gulliver.espci.fr, http://statbio.net
Internship location : ESPCI, 10 rue Vauquelin, 75005 Paris

Physics-based statistical models of protein sequences

Since the discovery of the genetic code more than half a century ago, understanding the relationship between the amino acid sequence of a protein and its function (specific binding, catalysis, regulation, information transmission, etc.) has remained an open problem in biology. The traditional approach to this problem combines structural biology (3D protein structures determined by X-ray crystallography or NMR) and computer modeling, effectively solving the 'sequence→3D structure' problem but leaving the 'sequence→function' problem largely open. Consequently, we do not fully comprehend the function of many proteins and cannot design new protein sequences with desired functions.

A completely different approach is to look at the problem from the standpoint of Evolution, the dynamic process by which natural proteins are formed, and to analyze how evolution encodes function into protein sequences. This data-driven approach infers statistical models of the sequence→function relationship from datasets of protein sequences, arising either from natural evolution or from experimental evolution performed in the lab. These models are constructed using tools from statistical physics (such as Potts models) and/or machine learning (for example, Transformers). They exist in sequence space rather than physical space and are generative, enabling the proposal of new sequences of functional proteins.

A current challenge is instilling physical interpretability into these models, to infer how sequences relate to different physical properties of the proteins (such as catalysis versus thermal stability), and to design proteins with new combinations of physical properties. Recent studies have demonstrated how this can be achieved by inferring a biophysical model from experiments of artificial evolution, with parameters learned through machine learning methods [1]. For instance, our group conducts experimental selections of antibodies [2,3] and has shown how the data from such experiments can be used to build a physics-based statistical model [4]. This model enables the design of antibodies with new specificity profiles, i.e., the ability to bind certain molecules while avoiding others.

The goal of the internship will be to develop these statistical models using sequence data from experimental or natural evolution. The project requires good skills in statistical physics and an interest in biological questions. Proficiency in machine learning would be a plus but is not necessary as long as the candidate is prepared to learn about these approaches. The project will take place in an interdisciplinary team of physicists and biologists working theoretically and experimentally on related projects.

References:

[1] Kinney, J. B., McCandlish, D. M. (2019). Massively parallel assays and quantitative sequence–function relationships. Annual review of genomics and human genetics, 20, 99-127.

[2] Boyer, S., Biswas, D., Kumar Soshee, A., Scaramozzino, N., Nizak, C., Rivoire, O. (2016). Hierarchy and extremes in selections from pools of randomized proteins. Proceedings of the National Academy of Sciences, 113(13), 3482-3487.

[3] Schulz, S., Boyer, S., Smerlak, M., Cocco, S., Monasson, R., Nizak, C., Rivoire, O. (2021). Parameters and determinants of responses to selection in antibody libraries. PLoS computational biology, 17(3), e1008751.

[4] Fernandez-de-Cossio-Diaz, J. Uguzzoni, G. Ricard, K. Anselma, A. Nizak, C. Pagnani, A. Rivoire, O. (2023). Inference and design of antibody specificity: from experiments to models and back. preprint.