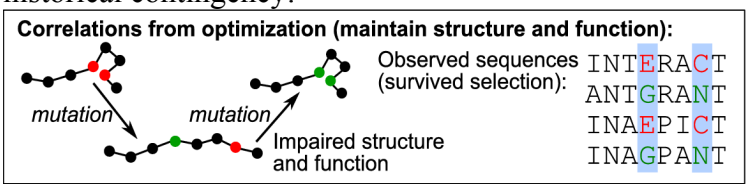
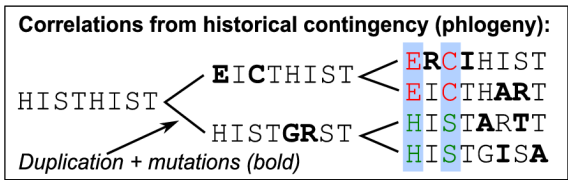
Understanding how optimization and phylogeny shape protein sequences
Internship and funded PhD

Advisor: Anne-Florence Bitbol, EPFL, Lausanne, Switzerland (currently at Laboratoire Jean Perrin, Sorbonne Université, Paris, France); contact: anneflorencebitbol@gmail.com

Proteins and multi-protein complexes play crucial roles in our cells, as enzymes, molecular motors, receptors, and more. The amino-acid sequence of a protein encodes its function, including its structure and its possible interactions. In evolution, random mutations affect the sequence, while natural selection acts at the level of function. Shedding light on the sequence-function mapping of proteins is central to a systems-level understanding of cells, and has far-reaching applications in synthetic biology and drug targeting. The current explosion of available sequence data enables data-driven approaches to discover the principles of protein operation. In particular, methods inspired by statistical physics and information theory have allowed to gain insight in protein structure [1], function [2] and interactions [3,4] starting just from sequence data. The basic idea is that amino acids that possess related functional roles often evolve in a correlated way.

In alignments of homologous protein sequences, which have significant similarity due to shared ancestry, correlations exist between certain amino-acid sites. These correlations can arise both from functional optimization, as homologs tend to maintain similar structures and functions, and from historical contingency:





We aim to establish a full decomposition of protein sequence covariation, dissecting signatures from functional optimization, and from evolutionary history, i.e. phylogeny. This will improve our understanding of the sequence-function relationship of proteins. We also aim to make new predictions for protein-protein interactions from sequence data, and to understand whether real proteins are mechanically optimized.

Several directions are possible, depending on the background and tastes of the applicant:
- Theoretical directions involve developing statistical physics based simulations to generate controlled synthetic data, and employing analytical calculations to make sense of the synthetic data.
- Data-driven directions involve analyzing real protein sequence data, as well as protein structure data.

Practical information:
The internship and PhD will take place at EPFL (École Polytechnique Fédérale de Lausanne) in Lausanne, Switzerland, in the new group of Anne-Florence Bitbol that will start on February 1st, 2020. The PhD position is fully funded with a competitive salary. PhD applicants will need to be admitted to an EPFL graduate school, either EDPY (Physics) or EDCB (Quantitative and Computational Biology).

References:
[1] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. 106, 67 (2009).
[2] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, Cell 138, 774 (2009).
[3] A.-F. Bitbol, R. S. Dwyer, L. J. Colwell, and N. S. Wingreen, Proc. Natl. Acad. Sci. U.S.A. 113, 12180 (2016).
[4] A. F. Bitbol, PLoS Comput. Biol. 14, e1006401 (2018).