Inferring interaction partners and evolutionary constraints from protein sequences

Anne-Florence Bitbol, Sorbonne Université

Proteins and multi-protein complexes play crucial roles in our cells. 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. Hence, 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 has inspired data-driven approaches to discover the principles of protein operation. At the root of these approaches is the observation that amino-acid residues which possess related functional roles often evolve in a correlated way.

First, I will present two novel methods to predict protein-protein interactions from sequence data. One method is based on the maximum-entropy inference approach that has already allowed to infer protein structures from sequences, and the other one is based on information theory. These methods accurately identify which proteins are functional interaction partners among the paralogous proteins of two families, starting from sequence data alone. They also provide signatures of the existence of interactions between protein families. I will further discuss the role of correlations arising from the shared evolutionary history of interacting partners in the success of these methods.

Then, I will propose a simple interpretation of the origin of the "sectors" of collectively correlated amino acids that have been discovered in several protein families through statistical analyses of sequence alignments. I will show that selection acting on any functional property of a protein, represented by an additive trait, can give rise to such a sector.