Advanced statistical modeling of biological sequences

Mots clés : Array

- Directeur de thèse : Martin Weigt
- Co-encadrant(s) :
- Unité de recherche : Laboratoire de Biologie Computationnelle et Quantitative
- École doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal : Sciences et technologies de l’information et de la communication

Résumé du projet de recherche (Langue 1)

Biology is undergoing a deep transformation toward a data-rich science. Approaches based on statistical modeling and learning are key to use such data to make sense of raw data coming out of numerous high-throughput experiments, to extract information hidden in such data, and to increase our understanding of complex biological processes.

Within this doctoral project, we will explore modern statistical and machine learning approaches, including in particular restricted Boltzmann machines and adversarial networks, for an unsupervised description of families of proteins, i.e. proteins coming from a common ancestor in evolution, which have conserved important structural and functional characteristics based on strongly diverged sequences. Our aim is to construct parsimonious, interpretable and generative models. Interpretability is important to improve our understanding of the selective forces underlying natural evolution of biological molecules. The generative character of these models is important in the context of the development of novel, data-driven approaches to de novo protein design.

Informations complémentaires (Langue 1)

The project will be realized in the "Statistical Genomics and Biological Physics" team of the LCQB. The PhD student will fully benefit from the national and international network of collaborations of this team.