Learning representations of protein from sequence, structure, and network

Understanding of protein structure and protein-protein interaction is crucial for studying molecular pathways and gaining insights into various biochemical processes. Data-driven approaches for predicting protein structure and interaction have been recently improved, partially due to the advances in machine learning. The success of machine learning algorithms often depends on data representation, which encodes explanatory factors of variation behind the data. Although our prior knowledge in protein science can help design good representations of proteins, powerful techniques capable of identifying protein patterns and sharing insights across diverse datasets are needed. In this talk, I will discuss three recent work on learning protein representations from sequence, structure and network data. First, I will introduce DeepContact, a deep convolutional neural-network (CNN) based approach that identifies conserved structural motifs, automatically and effectively leveraging patterns of residue-residue contacts to enable accurate inference of contact probabilities. Second, I will discuss DeepFold, another CNN-based approach to extract structural motifs within protein structure to enable accurate and efficient alignment-free structure search. Lastly, I will present Mashup, a feature learning algorithm to integrate protein-protein interaction networks for functional inference. In addition to the state-of-the-art performance, we expect these representation learning algorithms to provide biologically meaningful and deep insights into the organizational structure of protein folds and interaction networks.