

Abstract

It is widely believed that the protein structures play key roles in determining the functions, activity, stability and subcellular localization of the proteins, and the mechanisms of protein-protein interactions in cells. However, it is extremely labor-expensive and sometimes even impossible to experimentally determine the structures for hundreds of thousands of protein sequences. In this thesis, we aim at designing computational methods to predict the protein structures from sequences. Since the protein structures involve many aspects, we focus on predicting the general protein structural topologies (as opposed to specific 3-D coordinates) of different levels in the structure hierarchy, including secondary structures, tertiary structures and quaternary folds for homogeneous multimers. Specifically, given a protein sequence, our goal is to predict what are the secondary structure elements, how they arrange themselves in three-dimensional space, and how multiple chains associate into complexes.

Traditional approaches for protein structure prediction are sequence-based, i.e. searching the database using PSI-BLAST or matching against a hidden Markov model (HMM) profile built from sequences with similar structures. These methods work well for simple conserved structures with strong sequence similarities, but fail when the similarity across proteins is poor and/or there exist long-range interactions, such as those containing β -sheets or α -helical couples. These cases necessitate a more expressive model, which is able to capture the structured features in protein structures (e.g. the long range interactions).

In this thesis, a framework of conditional graphical models is developed to predict protein structures. Specifically, we define a special kind of undirected graph, i.e. a protein structure graph, whose nodes represent the state of the concerned structure elements (either a residue or a secondary structure element) and whose edges indicate local interactions between adjacent nodes in the linear chain or the long-range interactions between neighboring nodes in three-dimensional space. Following the idea of discriminative model, the conditional probability of the labels given the observed sequences is defined directly as an exponential function of the features over the graph, without any assumptions about the data generating process. In this way, our framework is able to handle the long-range interactions directly and incorporate any overlapping or long-range interaction features easily. Within this framework, we develop conditional random fields and kernel conditional random fields for protein secondary structure prediction; segmentation conditional random fields and chain graph model for tertiary fold (or motif) recognition and alignment prediction; and the linked segmentation conditional random fields for quaternary fold prediction and alignment prediction.

The thesis work makes contributions from two perspectives: computationally, it enriches current graphical models for the prediction problem with structured-outputs, in particular to handle the long-range interaction problem common in various applications, such as information extraction, machine translation and so on. Furthermore it relaxes the independent and identically distributed (ii) assumptions for data with inherent structures theoretically; biologically, it significantly improves the protein structure predictions, provides a better understanding of the mapping from protein sequences to structures, and hopefully our prediction results will shed light on the functions of some protein folds and aid the studies on drugs designs.