Machine Learning Based Disease Gene Identification and Protein-peptide Binding Prediction

Zhonghao Liu

Abstract

For many tasks from Bioinformatics such as protein structure simulation and modeling, sequence analysis, protein localization, gene expression and drug discovery etc., machine learning methods have been widely adopted and have been verified as an effective technique. In this dissertation proposal, we focus on two important and challenging topics: mislocation-related cancer gene identification and protein-peptide binding prediction.

For mislocation-related cancer gene identification, we proposed a pipeline to help identify potential mislocation-related genes among known cancer genes. The pipeline we proposed only needs protein-protein interaction network and gene expression data as input, which is very lightweight comparing with other methods. And the experiments showed that our pipeline has a good capability to identify the mislocation-related cancer genes.

For second topic, protein-peptide binding prediction, we first addressed a more specific task: peptide-MHC I binding prediction. We presented our allele-specific convolutional neural network (CNN) for peptide-MHC I binding prediction. The performance of our model on benchmark dataset are better than all existing prediction models. To improve the performance and overcome the limitations of allele-specific models, we proposed a pan-specific model for peptide-MHC I binding prediction problem.