早稲田大学大学院情報生産システム研究科

博士論文概要

論文題目

Study on Deep Transfer Learning Methods for the Predictions of Protein Functions

申請者

Xin YUAN

情報生産システム工学専攻
ニューロコンピューティング研究

2022 年 4 月

Carbohydrates, proteins and lipids are the three essential substances of life. Protein as one of them participates in almost all life activities. Protein function is the sum of all the behaviors of proteins and the behaviors that happen through proteins. Therefore, protein function prediction is of great significance for many biological studies. However, the entire experimental prediction methods have taken a lot to identify the function information but with little success. On the other hand, structural genomics projects exponentially increase the number of protein sequences by the high-throughput method in genome-wide strategies. Thus, more and more researches focus on computational prediction methods using protein sequences.

Protein function prediction can be carried out in many ways. This dissertation mainly discusses three cases: protein GO (gene ontology) annotation prediction, subcellular localization prediction, and PPI (protein-protein interaction) prediction. GO annotation is a description category of the whole functions. GO annotation prediction is a complicated multilabel classification task because there are thousands of GO classes and proteins usually have more than one annotation. Subcellular localization describes the proteins in different locations corresponding to the cellular functions. Finally, the protein interaction, as the critical point of forming protein macromolecular, includes PPI prediction and protein complex detection. They play an essential role in studying molecular functions.

Applying machine learning methods for protein function prediction has become popular in recent years. The superiority of machine learning has been proved many times. However, a powerful deep learning method is not yet successful in protein function prediction tasks. On one hand, since the prediction tasks are usually very complicated multilabel classification problems, it is crucial to implement deep neural networks for the tasks. However, experimentally annotated proteins are not available for most species. As a result, it requires to perform a transfer learning based on the experimentally annotated proteins available for some few type species. On the other hand, it has been known that it is difficult to implement a transfer learning in bioinformatics tasks because of the overfitting problems due to limited training samples.

To address the above problems, we develop a novel deep convolutional neural network (CNN) model with multi-head and multi-end (MHME) to implement a class of classifiers in one model. The deep MHME CNN model shares a deep CNN feature extractor in a class of related different prediction tasks so as to extract common feature, which makes the transfer learning possible. The proposed model is then applied for three protein function prediction tasks: GO annotation prediction, subcellular localization prediction and PPI prediction based on a transfer learning from different tasks and different species.

The dissertation contains five chapters as follows:

Chapter 1 first introduces the background of proteins and protein functions. Then we discuss the three prediction issues and the related researches and summarize the challenges. At last, the goal of deep modeling for protein function predictions and our proposal are listed.

Chapter 2 develops a deep hierarchical model for GO annotation prediction. GO annotation prediction is first formulated as a very complicated task of hierarchical multilabel classification, consisting of a set of hierarchically organized local classifiers. A deep MHME CNN model is then proposed to implement the whole set of hierarchically related local classifiers in one model. The proposed model consists of three parts: the body part of a deep CNN model shared by different local classifiers for feature extracting and mapping; the multi-end part of a set of autoencoders performing feature fusion transforming the input vectors of different local classifiers to feature vectors with the same length to share the feature mapping part; and the multi-head part of a set of linear multi-label classifiers. In this way, by sharing a deep CNN with multiple local classifiers, we can extract common feature and construct more powerful local classifiers for each level with limited training samples and achieve better classification performance. Experiment results on various benchmark datasets from Uniprot show that the proposed deep CNN based model has better performance than the state-of-the-art traditional models. Moreover, it gives rather good performance even under transfer learning of same tasks, but different species.

Chapter 3 introduces a deep protein subcellular localization predictor enhanced with transfer learning of GO annotation. GO annotations have been known to be useful for the subcellular localization prediction. However, experimentally annotated proteins are not always available. It is motivated to perform deep learning of GO annotations on the available experimentally annotated proteins for some type species and transfer it to subcellular localization prediction on other species. The proposed deep protein subcellular localization predictor consists of a linear classifier and a deep CNN feature extractor. By using the deep MHME CNN model, a deep CNN feature extractor is first shared and pretrained in a deep GO annotation predictor, and then is transferred to the subcellular localization predictor with fine-tuning using protein localization samples. In this way, we have a deep protein subcellular localization predictor enhanced with transfer learning of GO annotation. The proposed method has good performances on the Swiss-Prot datasets when transfer learning using the protein samples both within and out species. Moreover, it outperforms the state-of-the-art traditional methods on benchmark datasets.

Chapter 4 proposes a deep PPI predictor and reconstructs a PPI network based on deep transfer learning for protein complex detection. The completeness of a PPI network is crucial for the detection of protein complexes. However, complete PPI networks are not available for most species because experimentally identified PPIs are usually very limited. To solve the problem, a deep learning based PPI predictor is proposed to estimate the unknown PPIs, and construct a complete PPI network, from which protein complexes are detected using a spectral clustering method. Considering the facts that the similarities of GO annotations contribute to protein interactions, and the differences of subcellular localizations contribute to negative interactions, the deep MHME CNN model is used to pretrain a deep CNN feature extractor in a class of deep GO annotation and subcellular localization predictors using datasets from the type species, then transfer it to the PPI prediction model for fine-tuning, so as to have a deep PPI detector enhanced with transfer learning of GO annotation and subcellular localization prediction. Experimental results on benchmark datasets CYC2008 and MIPS show that the proposed method outperforms the state-of-the-art methods.

Chapter 5 concludes the contributions of this dissertation and provides future works. In summary, this dissertation proposes a deep MHME CNN model and applies it to the predictions of protein GO annotation, subcellular localization and PPI based on a transfer learning from different tasks and different species.