ProDec-BLSTM

Background: Protein remote homology detection plays a vital role in studies of protein structures and functions. Almost all of the traditional machine leaning methods require fixed length features to represent the protein sequences. However, it is never an easy task to extract the discriminative features with limited knowledge of proteins. On the other hand, deep learning technique has demonstrated its advantage in automatic learning representations. It is worthwhile to explore the applications of deep learning techniques to the protein remote homology detection.

Results: In this study, we employ the Bidirectional Long Short-Term Memory (BLSTM) to learn effective features from protein sequences, also propose a predictor called ProDec-BLSTM: it includes input layer, bidirectional LSTM, time distributed dense layer and output layer. This neural network can automatically extract the discriminative features by using bidirectional LSTM and the time distributed dense layer.

Conclusion: Experimental results on a widely-used benchmark dataset show that ProDec-BLSTM outperforms other related methods in terms of both the mean ROC and mean ROC50 scores. This promising result shows that ProDec-BLSTM is a useful tool for protein remote homology detection. Furthermore, the hidden patterns learnt by ProDec-BLSTM can be interpreted and visualized, and therefore, additional useful information can be obtained.

Dependency

Keras 2.0.6
Theano 0.9.0
Numpy 1.11.2
Biopython 1.68

Content

data/: the training dataset and testing dataset in FASTA format.
results/: the prediction results of ProDec-BLSTM.

Usage

```
python ProDec-BLSTM.py [-e] [<0.0005>] [-family_index <family index>] [-train FALSE] [-test FALSE] [-pos_train_dir] [<dir>] [-neg_train_dir] [<dir>] [-pos_test_dir] [<dir>] [-neg_test_dir] [<dir>] [-model_dir] [<dir>] [-weights_dir] [<dir>]
```

-family_index: family index
-train: train a ProDec-BLSTM model
-test: load the trained ProDec-BLSTM model
-model_dir: the directory of the trained model json file of ProDec-BLSTM. If test is false, this argument can be empty.
-weights_dir: the directory of the trained model weight file of ProDec-BLSTM. If test is false, this argument can be empty.
-pos_train_dir: the directory of positive training dataset
-neg_train_dir: the directory of negative training dataset
-pos_test_dir: the directory of positive testing dataset
-neg_test_dir: the directory of negative testing dataset
-h: show this usage

The parameters of ProDec-BLSTM can be set in Paramters.py.

Example

1. An example is provided for detecting the proteins in the target family a.138.1.3 (SCOP ID) by using only one command line "./run_and_train.sh". The trained model of ProDec-BLSTM in the GPU environment is also provided. This example using the trained model can be implemented by using "./run_and_test.sh".