iterb-PPse: Identification of transcriptional terminators in bacterial by incorporating nucleotide properties into PseKNC

Projects that provide data and codes in the experiment

Description: Terminator is a DNA sequence that give the RNA polymerase the transcriptional termination signal, which plays a vital role in gene transcription, so it is crucial to identify terminators. However, accurately prediction methods are deficient and in urgent need. Especially there is no comprehensive study on the comparison and analysis of different features and machine learning models. Therefore, we used five feature extraction methods (Base-content, K-pwm, nucleotidepro, PseKNC - Ⅰ, PseKNC - Ⅱ) and utilized two-step method to select features. In addition, we compared five single models (Support Vector Machine, Naive Bayes, Logistic Regression, Decision Tree, Multi-layer Perceptron, K-Nearest Neighbor) as well as four ensemble models (AdaBoost, Bagging, Extreme Gradient Boosting (XGBoost), Gradient Boosting Method (GBM)).



|  |  |  |
| --- | --- | --- |
| Function | File | Description |
| data\_process | dataprocess.m | Import DNA sequence from csv file into matlab and remove redundant information |
| feature extraction | PSSM.m | K-pwm |
| physical properties.m | nucleotidepro |
| Pesknc1.m | PsekNC-Ⅰ |
| Pseknc2.m | PsekNC-Ⅱ |
| GC.m | Base-content |
| feature selection | write4libsvm.m | Transform the data into the format needed for feature sorting  Input format is labeled feature matrix (label in last column)  Output is a txt file in a specific format |
| fselect.py | Feature ranking using F-score  Run the following statement on the command line: ‘ptyoh fselect.py data.txt’, data.txt represents the file obtained from write4libsvm.m |
| txTOm.m | Extract the permuted feature order from the output file “\*.fscore” |
| BD.m | Feature sorting using binomial distribution |
| ifs.m | Incremental feature selection |
| featureSelect | feature selection |
| model | SVM.m | Support Vector Machine |
| XGBoost.py | XGBoost |
| LR.py | Logistic Regression |
| NB.py | Naïve Bayes |
| DT.py | Decision tree |
| KNN.py | K-Nearest Neighbor |
| MLP.py | Multi-layer Perceptron |
| Bagging.py |  |
| AdaBoost.py |  |
| GBM.py | Gradient Boosting Method |
| evaluation | crossvalid.m | 100 times 5-fold cross-validation |
| plot\_roc.m | plot ROC curve |