# Data

The data included a sample set of seven species used in the experiment

# matlab

|  |  |
| --- | --- |
| File name | Summarize |
| allsequence1-6 | This data shows all the nucleotide fragments that are needed to calculate the frequency information for different values of k |
| Base-content | Base-content- feature extraction method |
| dataprocess.m | This file contains our pre-processing of the sample file |
| featureSelect.m | According to the sorting of features, a specific number of features are selected |
| fselect.py | F-score toolkit |
| ifs.m | The second step feature selection：IFS |
| J.m | PaeKKNC-ⅡThe physicochemical properties of nucleotide in feature extraction method are calculated |
| myproject1.m | SVM Parameter adjustment and training |
| phy.mat | The standard value corresponding to the physical and chemical properties |
| pinci1-6 | Extract the frequency of nucleotide fragments at k values 1 to 6 |
| Pseknc2.m | PseKNC-Ⅱfeature extraction method |
| textDtonum.m | TF-IDF scores of all nucleotide fragments were calculated according to the sequence information |
| textD.m | Extract the characteristics of each sample according to TF-IDF |
| txtTOm.m | According to the running results of F-score, the sorting of features is extracted |
| write4libsvm.m | Data preprocessing before F-score is used requires the use of this file to process standard intervals and features into a format acceptable to F-score |

# pyhon

|  |  |
| --- | --- |
| File name | Summarize |
| AdaBoost | AdaBoost Model tuning and training |
| Bagging | Bagging Model tuning and training |
| DT | Decision Tree Model tuning and training |
| GBM | GBM Model tuning and training |
| KNN | KNN Model tuning and training |
| LR | Logistic Regression Model tuning and training |
| MLP | MLP Model tuning and training |
| NB | Naïve Bayes Model tuning and training |
| XGBoost | XGBoost Model tuning and training |