m6Aminer: Predicting the m6Am Sites on mRNA by Fusing Multiple Sequence-Derived Features into a CatBoost-Based Classifier

m6Aminer:基于多序列衍生特征和CatBoost分类器的mRNA上m6Am位点预测器

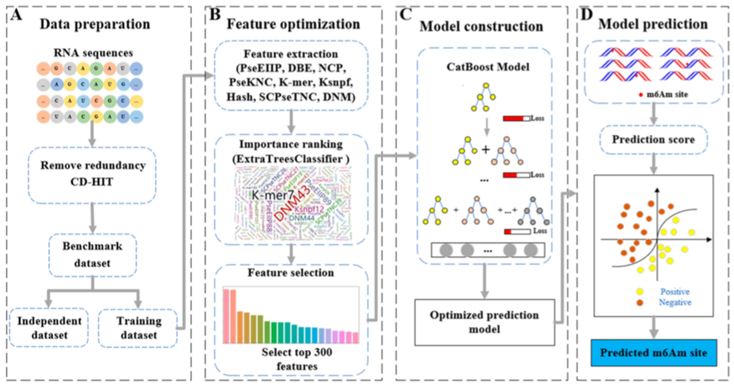
水利与建筑工程学院(水利水电科学研究院)
College of Water Resources and Architectural Engineering

小组成员：杨文慧、彭诗哲、王业强、兰鹏飞　指导老师：刘泽

1.Abstract

As one of the most important post-transcriptional modifications, m6Am plays a fairly important role in m6Am stability and cancers progression. The accurate identification of m6Am sites is critical for explaining its biological significance and developing its medical application. However, conventional experimental approaches are time-consuming and expensive, making them unsuitable for the large-scale identification of the m6Am sites. We exploit m6Aminer (based on CatBoost) to identify the m6Am sites on mRNA. For feature extraction, nine different feature-encoding schemes were utilized to form the initial feature space. The ExtraTreesClassifier algorithm was adopted to obtain optimal subset with top-300 feature by importance ranking. Comprehensive comparison results give m6Aminer leading edge over the state-of-the-art predictors m6AmPred and DLm6Am. The prediction model in this study can lay a foundation for the functional research of m6Am.

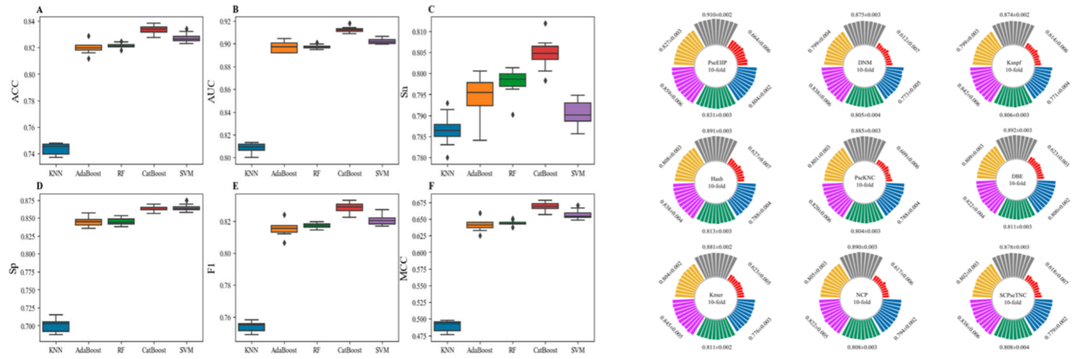
2. Model Construction



- Utilize CD-HIT to remove the redundant sequences and build the benchmark datasets.
- Form an initial feature space and select the top 300 features according to their importance.
- Construct and optimize the CatBoost-based model.
- Predict the m6Am sites using m6Aminer.

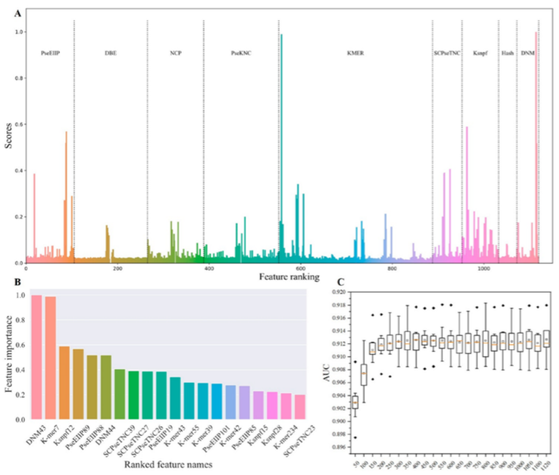
3.Results

- **Model Performance with Different Machine Learning Algorithms and Feature Subsets**



➢ Select appropriate classifiers and feature subsets to construct m6Aminer.

- **Feature Ranking and Selection**

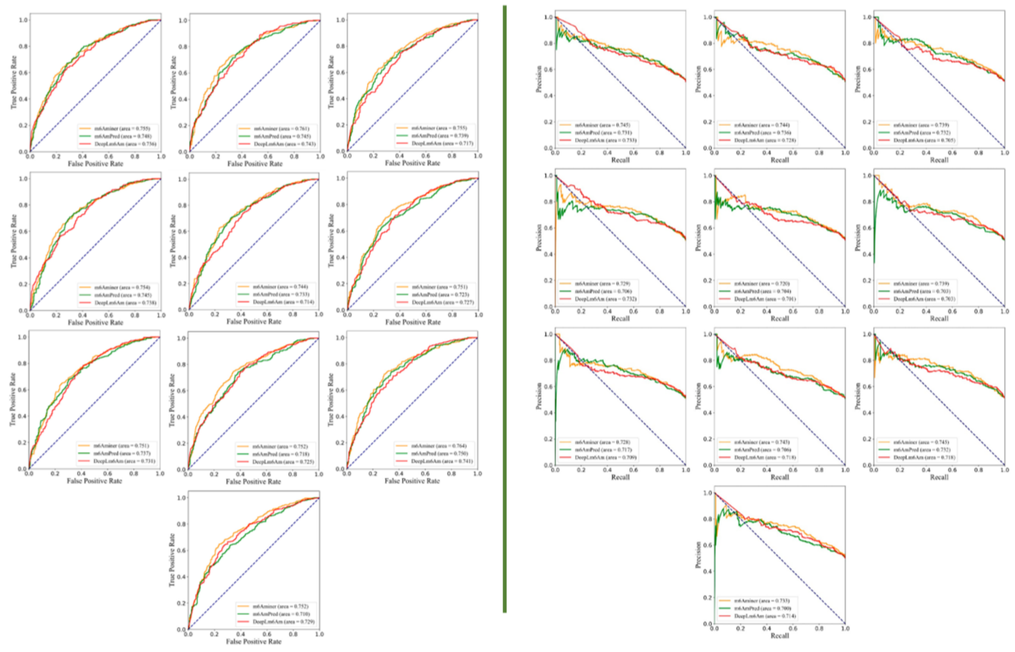


- The importance of 1120 features on a sub-training dataset
- The top 20 features ranked using the ExtraTreesClassifier algorithm
- The performance of the CatBoost-based model using different combinations of feature subsets on the 10 sub-training datasets.

➢ Mine effective features utilizing the ExtraTreesClassifier algorithm.

3.Results

- **Comparison with the State-of-the-Art Predictors**



➢ Compare m6Aminer with the state-of-the-art predictors m6AmPred and DLm6Am on a completely new dataset.

- **Web Server**

➢ Design a web server based on the flask framework to predict the m6Am sites.

4.Conclusions

- A new computational model——m6Aminer, was developed to improve the prediction of m6Am.
- The electronic, physical, and chemical properties of nucleotides have an important impact on the classification task of the m6Am sites.
- m6Aminer achieves competitive performance when compared with m6AmPred and DLm6Am.
- m6Aminer can be used as an efficient tool for identifying m6Am or can at least become an auxiliary means of identifying m6Am in the future.

5. Achievements