**Feature-Analysis: a Python pipeline tool for feature extraction, feature selection, machine learning and deep learning based on DNA, RNA and protein sequence data**

**ALLFEATURE: an integrated Python pipeline tool for extracting, selecting, modeling and evaluating features of DNA, RNA and Proteins**

**sequencing data**

**Authors…**

**ABSTRACT**

The faster and cheaper modern sequencing technologies revolutionize genomic research and trigger explosive growth of DNA, RNA and proteins sequencing data. Using bioinformatics tool to analyze and process the DNA, RNA and protein sequence data efficiently and accurately has become a popular and challenging task in genomics and proteomics. With the development of experimental methods and statistical learning, an integrated and user-friendly tool that containing the state-of-the-art data mining methods are needed. Here, we propose **ALLFEATURE**, a comprehensive Python pipeline tool that integrating multiple steps to analyze DNA, RNA, protein sequencing data: feature extraction, dimensionality reduction, feature selection, machine learning and deep learning models construction. We also use DNA N6-methyladenosine sites, RNA N 6-methyladenosine sites and protein-protein interactions datasets to evaluate the validation of the package. It is anticipated that

**ALLFEATURE** can effectively perform biological sequence analysis and predictor construction. The Python pipeline tool can be download **…**

Key words: python package; web server; feature extraction; feature selection; machine learning; deep learning

# INTRODUCTION

The appearance of next generation sequencing (NGS) technology have significantly improved the qualities and quantities of biological sequences (DNA, RNA and protein). With NGS, the duration for sequencing an entire human genome is reduced from a decade to a single day [1] and its cost dropped from $300000 to less than $1000 [2]. This faster and cheaper modern sequencing technology revolutionizes genomic research field and trigger explosive growth of DNA, RNA and protein sequencing data. The most recent released version, 232 of GenBank in NCBI contains 213,387,758 sequences and WGS in NCBI includes 1,022,913,321 sequences [3]. The sequencing data contains valuable information for current biological researchers to explore. How to study the structure and function from the perspective of biological sequences is significant for biological research, disease diagnosis, biotechnology and many other areas [4]. However, analyzing sequencing data with human inspection alone is hardly possible. Thus, indispensable computational tools are widely used in sequencing data analysis.

Recent research shows that machine learning methods and algorithms have been heavily applied to explore structural and functional properties of DNA, RNA and protein sequencing data [5]. The analysis of biological sequence can be mainly regarded as binary and multi-class prediction tasks [6-10], including DNA N6-methyladenosine site, DNA N4-methylcytosine site, RNA N6-methyladenosine site, RNA-binding protein identification, protein function site [10-11], protein fold recognition [12], protein-protein interaction [13-14] prediction and many other bioinformatics research purposes.

For the process to construct predictors, the first step is feature extraction, which transform the character sequence data into numeric vectors of the same length. The feature information in sequencing data can be classified as sequence-based, physicochemical property-based, evolution-based and structure-based biological information. After extracting feature vectors, selecting a suitable classifier algorithm is another step for the prediction model construction. There are many popular models can be used for researchers to analyze the structure and function of sequencing data, such as XGBoost, support vector machine (SVM) and k nearest neighbors (KNN) and etc. Inspired by this, several computational tools or web servers have been released, including PROFEAT, Pse-in-one, PyFeat, POSSUM, iFeature, BioSeq-Analysis, iLearn and BioSeq-Analysis2.0.

However, some challenges still remain in the biological sequence analysis. First, overfitting and time-consuming issues often exist during performing accurate prediction in machine learning. Based on our observation, extracted feature vectors from feature extraction step often displays high dimensionality. They often contain a lot of noisy features which can cause poor prediction and time-consuming issues while modeling [15]. Secondary, most computational tools or web servers only focus on one individual step instead of integrating all functionalities for sequencing data analysis. For example, repDNA is a python package to only generate various types of feature vectors for DNA sequences [16]; Pse-in-one 2.0 and PyFeat are web server and generation tools only for extracting different feature models from DNA, RNA, and protein sequences [17]. To our knowledge, there are two computational tools: iLeran and BioSeq-Analysis that integrating multiple steps for sequencing data analysis, but the integrated classifier algorithms are not sufficient. Finally, deep learning is very powerful computational tools for classification tasks via layer by layer learning. The popular deep learning methods such as autoencoder network, deep neural network, convolutional neural network, and recurrent neural network show convincing performances for prediction. However, the packages above lack the usages of deep learning theories.

Feature selection is the step to overcome these challenges by only selecting those features that contribute most to the prediction and removing other noisy features. In addition, for some feature vectors with extremely high dimensionalities, dimension reduction methods are often involved in the sequencing data analysis to pre-filter vectors to reduce the time cost and improve performance. Inspired by above discussion, we propose a new python pipeline, **ALLFEATURE**, that integrating 20 feature selection methods, total of 16 dimensionality reduction methods and 13 prediction/classification models. In addition, our pipeline also contains the step of feature extraction to generate total of 60 different models of features from DNA, RNA and protein sequencing data. And the flowchart of python pipeline tool are shown in Figure 1. Compared with other software packages, the proposed ALLFEATURE has following advantages: (i) sufficient feature selection methods and dimension reduction methods, including regularization-based, statistic-based, information-based, tree-based and recursive feature elimination-based approaches. (ii) Ten machine learning methods and three deep learning approaches, deep neural network (DNN), convolutional neural network (CNN), and recurrent neural network (RNN). (iii) More abundant graphical display results, including box figure, ROC curves, PR curves and etc.

G:\feature_selection_package\正文\flowchart\flowchart (第二版).tif

**Figure 1.** The overflow of ALLFEATURE. The python package contains feature extraction, feature selection, machine learning, deep learning, and performance evaluation. The input are the DNA, RNA and protein sequences with FASTA format, or the high-dimension matrix with csv format. The outputs of pipeline will provide generated feature vectors, prediction accuracy comparison and suggestion of best model for researchers.

# MATERIALS AND METHODS

The DNA, RNA and protein biological sequence  with  residues can be regarded as:

 (1)

where  represents the  residue. And the biological sequence contains the important and effective information of structure and function. How to analyze the attributes of sequence and construct the useful predictor based on machine learning is necessary. The commonly steps for modeling the predictor based on DNA, RNA and protein sequence: (i) Feature extraction, which obtain the composition information, physiochemical information and evolutionary information and etc. (ii) Feature selection (dimension reduction), which remove redundancy and noise from the extracted feature vectors and retain clean, effective and understandable feature information. (iii) Machine learning and deep learning, which could predict the structure and function via predictor construction based on sequence data. (iv) Model evaluation, which generate the prediction results of tables and figures.

## Feature extraction

**ALLFEATURE** directly extracts features from DNA, RNA or protein sequences based on the total of 60 different types of feature methods. The step of feature extraction consists 16 feature extraction methods for DNA; 12 feature extraction methods for RNA; 32 feature extraction methods for protein sequences, which can be shown in Table 1 and Table 2, respectively.

Table 1. List of 16 DNA feature extraction methods and 12 RNA feature extraction methods

|  |  |  |  |
| --- | --- | --- | --- |
| **DNA Feature Extraction Methods** | **RNA Feature Extraction Methods** | **Extraction Method Description** | |
| Kmer | Kmer | | DNA or RNA sequence are represented as the occurrence frequencies of k neighboring nucleic acids [21, 22] |
| Reverse Compliment Kmer (RCKmer) | Reverse Compliment Kmer (RCKmer) | | A variant of Kmer descriptor by removing the reverse compliment Kmer [21, 23] |
| Pseudo Dinucleotide Composition (PseDNC) | Pseudo Dinucleotide Composition (PseDNC) | | Incorporating the contiguous local sequence-order and global sequence-order information [24] |
| Pseudo k-tuple Nucleotide Composition (PseKNC) | - | | Extending the PseDNC by incorporating k-tuple nucleotide composition [25] |
| Dinucleotide Based Auto Covariance (DAC) | Dinucleotide Based Auto Covariance (DAC) | | Measuring the correlation of the same physicochemical index between two dinucleotides separated by lag along the sequence [26, 27] |
| Dinucleotide Based Cross Covariance (DCC) | Dinucleotide Based Cross Covariance (DCC) | | Measuring the correlation of two different physicochemical indices between two dinucleotides separated by lag nucleic acids [26, 27] |
| Dinucleotide Based Auto-cross Covariance (DACC) | Dinucleotide Based Auto-cross Covariance (DACC) | | Combining of DAC and DCC [26-28] |
| Trinucleotide Based Auto Covariance (TAC) | - | | Measuring the correlation of the same physicochemical index between trinucleotides separated by lag nucleic acids [28] |
| Trinucleotide Based Cross Covariance (TCC) | - | | Measuring the correlation of two different physicochemical indices between two trinucleotides separated by lag nucleic acids [28] |
| Trinucleotide Based Auto-Cross Covariance (TACC) | - | | Combining of TCC and TACC [28] |
| Nucleic Acid Composition (NAC) | Nucleic Acid Composition (NAC) | | Calculating the frequency of each nucleic acid type in nucleotide sequence [29] |
| Di-Nucleotide Composition (DNC) | Di-Nucleotide Composition (DNC) | | Containing 16 NAC descriptors [29] |
| Tri-Nucleotide Composition (TNC) | Tri-Nucleotide Composition (TNC) | | Containing 64 NAC descriptors [29] |
| zCurve Mathematical Formula (zCurve) | zCurve Mathematical Formula (zCurve) | | Calculating three components in three axes in genomic sequence analysis [30] |
| monoMonoKGap Theoretical Description (MonoKGap) | monoMonoKGap Theoretical Description (MonoKGap) | | - |
| monoDiKGap Theoretical Description (MonoDiKGap) | monoDiKGap Theoretical Description (MonoDiKGap) | | - |

Table *2*. List of 32 Protein feature extraction methods and their description

|  |  |  |
| --- | --- | --- |
| **Protein Feature Extraction** | **Extraction Method Description** | |
| Amino Acid Composition (AAC) | | Calculating the frequencies of 20 kinds of amino acids [31] |
| Dipeptide Composition(DC) | | transforming the variable length of proteins to fixed length feature vectors [31] |
| Composition of K-Spaced Amino Acid Pairs (CKSAAP) | | Extracting important intrinsic correlation information of protein sequences in multidimensional space [32-34] |
| Grouped Dipeptide Composition (GDC) | | A variation of the DPC descriptor which generates 25 descriptors [35] |
| Grouped Tripeptide Composition (GTC) | | Another variation of TPC descriptor which generates 125 descriptors [35] |
| Conjoint Triad (CT) | | Calculating the frequency of occurrence of each triad [36] |
| K-Spaced Conjoint Triad (KSCTriad) | | Combining CT and considers the continuous amino acid units that are separated by any *k* residues [37] |
| Composition (C)  Transition (T)  Distribution (D) | | Calculating composition descriptors  Calculating transition descriptors  Calculating distribution descriptors [38-40] |
| Encoding Based on Grouped Weight (EBGW) | | Capturing the continuity and discontinuity features based on grouped weight coding [41] |
| Auto Covariance (AC) | | Measuring the correlation of the same property between two residues separated by distance of *l* [42] |
| Moreau-Broto autocorrelation (Morean-Broto) | | Measuring the physiochemical and position information between two amino acid [43] |
| Moran Autocorrelation (Moran) | | Measuring the physiochemical information of adjacent amino acid [44] |
| Geary Autocorrelation (Geary) | | Measuring the physiochemical information and generate positive values [45, 46] |
| Quasi-Sequence-Order (QSO) | | Obtaining the sequence distribution patters for a specific physicochemical property [47] |
| Pseudo-Amino Acid Composition (PseAAC) | | Extracting the physicochemical information and sequence order information [48, 49] |
| Amphiphilic Pseudo-Amino Acid Composition (APAAC) | | Extracting the type-2 pseudo amino acid composition [48, 49] |
| Amino Acid Composition PSSM (ACC-PSSM) | | Calculating process of amino acid composition PSSM [50, 51] |
| Dipeptide Composition PSSM (DPC-PSSM) | | Extracting the sequence-order information in the PSSM [51] |
| Bi-gram PSSM (Bi-PSSM) | | Calculating the frequency of the transition between amino acids [52] |
| Auto Covariance PSSM (AC-PSSM) | | Measuring the correlation of the same property between two residues separated by lag [53] |
| Pseudo PSSM (PsePSSM) | | Calculating the PsePSSM feature vector according to the pseudo amino acid composition [54] |
| AB-PSSM | | Calculating feature vector based on averaged PSSM over blocks [55] |
| Secondary Structure Composition (SSC) | | Calculating feature based normalized count of frequency of the structural motifs present at the amino-acid residue positions [56] |
| Accessible Surface Area composition (ASA) | | Calculating feature based on normalized sum of accessible surface area [56] |
| Torsional Angles Composition (TAC) | | Calculating features based four different types of torsional angles [56] |
| Torsional Angles bigram (TA-bigram) | | Calculating feature based on the bigram of the torsional angles [56] |
| Structural Probabilities bigram (SP-bigram) | | Calculating feature based on structural probabilities for each position of amino acid residue [56] |
| Torsional Angles Auto-Covariance (TAAC) | | Calculating feature from the torsional auto-covariance [56] |
| Structural Probabilities Auto-Covariance (SPAC) | | Calculating feature from the structural probabilities [56] |

## Feature selection and dimensionality reduction

The dimension of feature vectors after feature extraction and feature fusion could be extremely high. Take composition of k-spaced amino acid pairs (CKSAAP) for example, when the parameter , the size of dimension could be 1200. At the same time, autocorrelation descriptor using physicochemical properties to extract sequence information. As we can kown, the AAindex database contains 554 physicochemical propertyies. High dimensions feature vectors usually contain some redundant and noisy features, which may increase memory storage and run time. Especially, this case may lead to overfitting and performance degradation. It is desirable to apply feature selection and dimensionality reduction to eliminate redundancy and noise information. Therefore, ALLFEATURE integrated a step of feature selection and dimensionality reduction in Table 3 and Table4. Selecting those features which contribute most to the prediction can effectively improve prediction accuracy and reduce the implementation time. These feature selection methods rank features according to their feature importance score for prediction of structure and function. Only wanted number of top-ranking features will be selected in the step of feature selection. Dimensionality reduction methods project the raw feature space to new low-dimension feature space, which could mine the linear or nonlinear relationship, eliminate redundancy and retain effective feature information.

Table 3. Feature Selection and Dimensionality Reduction Methods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature Selection Method** | | **Description** | **Dimensionality Reduction Method** | **Description** |
| Lasso | Using Lasso liner model to recursively eliminate features [58] | | K-means | Clustering data by separating samples in n groups of equal variances |
| ElasticNet | Using ElasticNet model to recursively eliminate features | | T-SNE | Visualizing high-dimensional data |
| L1-SVM | Using SVM with L1 penalty model to recursively eliminate features | | Principal Component Analysis (PCA) | Linear dimensionality reduction using singular value decomposition |
| CHI2 | Retrieving best features based on test | | Kernel PCA (KPCA) | Non-linear dimensionality reduction through use of kernels |
| Pearson Correlation (PC) | Retrieving best features based on Pearson correlation | | Locally linear embedding (LLE) | Reducing projection of data which preserves distances within local neighborhoods |
| ExtraTree | Using ExtraTree model to recursively eliminate features | | Truncated Singular Value Decomposition (TSVD) | Linear dimensionality reduction by means of truncated singular value decomposition |
| xgBosst | Using xgBoost model to recursively eliminate features | | Non-negative matrix factorization (NMF) | Reducing dimension by finding two non-negative metrices [8] |
| SVM-RFE | Using linear SVM model to recursively eliminate features | | Multi-dimensional Scaling (MDS) | Reducing dimension by modeling data as distances in a geometric space [9] |
| LOG-RFE | Using Logistic Regression model to recursively eliminate features | | Independent Component Analysis (ICA) | Reducing dimension by finding components with some sparsity |
| Mutual Information (MI) | Retrieving best features based mutual information | | Factor Analysis (FA) | Reducing dimension by performing a maximum likelihood estimate |
| Minimum Redundancy Maximum Relevance (MRMR) | Selecting features that still having high correlation to the classification variable | | Agglomerate Feature (AF) | Recursively merges feature instead of samples |
| Joint Mutual Information (JMI) | Retrieving best features based joint mutual information | | Gaussian Random Projection (GRP) | Reducing the dimension by projecting the original input space using the Gaussian distribution |
| Maximum Relevance Maximum Distance (MRMD) | Retrieving best features by measuring relevance and redundancy between features | | Sparse Random Projection (SRP) | Reducing the dimension by projecting the original input space using a sparse random matrix |
| ReliefF | Retrieving best features by calculating and ranking a feature score for each feature | | Autoencoder | Reducing the dimension using encode and decode neural network |
| Trace Ratio | Retrieving best features by calculating the corresponding score in trace ratio form | | Gaussian Noise Autoencoder (GNA) | Corrupting input before being passed to autoencoder neural network |
| Gini Index | Retrieving best features by constructing the measure function based on Gini-Index | | Variational Autoencoder (VA) | Neural network can be trained with stochastic gradient descent |
| Spectral Feature Selection (SPEC) | Retrieving best features based on structure induced | | - | - |
| Fisher Score | Retrieving best features based on scores of features under the Fisher criterion | | - | - |
| T Score | Retrieving best features based on their t-score | | - | - |
| Information Gain (IG) | Retrieving best features based on their information gain | | - | - |

## Machine learning and deep learning

A lot of analysis of biological sequence could be regarded as classification tasks in bioinformatics and computational biology, so machine learning and deep learning methods are key step for predictor construction. SVM is a kernel-based classifier through constructing the optimal hyperplane. RandomForest is a tree-based and widely used machine learning method. SVM and RandomForest are commonly used in the area of prediction tasks. Recently, gradient boosting decision tree, XGBoost and LightGBM are demonstrated to be excellent classifiers via gradient boosting algorithm. **ALLFEATURE** integrated 10 popular machine learning methods SVM, KNN, RandomForest, LightGBM, XGBoost, Adaboost, ExtraTree, gaussian Naïve Bayes, gradient boosting. In recent years, deep learning frameworks provide effective solutions for the analysis of DNA, RNA and protein sequence. Deep learning can mine essential sequence represent information via hierarchical structure, which can accurately, effectively and better perform the sequence prediction tasks. **ALLFEATURE** integrated three deep learning methods, including deep neural network (DNN), convolutional neural network (CNN), recurrent neural network (RNN).

## Model evaluation

In this paper, in order to measure the designed statistical prediction model, sensitivity is also the recall rate (SE), specificity (SP), accuracy (ACC), Matthew's correlation coefficient (MCC), precision (PRE), and F-measure (F1) are sued to evaluate the model performance.

 (2)

 (3)

 (4)

 (5)

 (6)

 (7)

where TP, TN, FP and FN in the above equations represent true positive, true negative, false positive and false negative, respectively. The Receiver Operating Characteristic (ROC) curve, the Precision-Recall (PR) curve and their corresponding AUC (area under the ROC curve) and AUPR (area under the PR curve) are also important indicators for evaluating the predictor.

# RESULTS AND DISCUSSION

## Web server

For convenience, ALLFEATURE python pipeline is composed with three independent sections, including DNA-Analysis, RNA-Analysis, and protein Analysis. The input sequences must be FASTA format. At the same time,

## Application of ALLFEATURE

For testing the usage of our pipeline tool, three prediction tasks were performed for DNA, RNA and protein sequences, respectively. In this paper, we use DNA N6-methyladenosine sites, RNA N 6-methyladenosine sites and protein-protein interactions datasets to evaluate the validation of **ALLFEATURE**. The prediction results can be automatically and easily generated. And the model performance could be comparable and even higher than the state-of-the-art approaches, which indicate our proposed python packages are useful and powerful when analyzing DNA, RNA and protein biological sequences.

## Task 1: DNA N6-methyladenine sites prediction

N6-methyladenine (6mA) is one kind of post-replication modification occurring in a wide range of DNA sequences. In prokaryotes, 6mA has been found to be associated with a wide range of biological processes such as DNA replication, repair, transcriptio, and cellular defense. 6mA site-containing sequences were taken from the genome of *Mus musculus* in the MethSMRT database, including 1934 positive samples and 1934 negative samples. Firstly, we use Kmer, Psednc, binary, TNC and MonoKGap five feature extraction methods to construct initial feature vectors, multi-information fusion can represent more effective information of DNA sequences. Lasso, ElasticNet, L1-SVM, CHI2, Pearson Correlation, ExtraTree, XGBoost, SVM-RFE, LOG-RFE, Mutual Information, Minimum Redundancy Maximum Relevance, Joint Mutual Information, Maximum-Relevance-Maximum-Distance, ReliefF, Trace Ratio, Gini index, Spectral Feature Selection, Fisher Score, T Score, Information Gain are employed to fulfill dimensionality reduction. Finally, the optimized feature vectors via the process of feature selection are fed into SVM, KNN, RandomForest, LightGBM, XGBoost, Adaboost, ExtraTree, gaussian Naïve Bayes, gradient boosting, DNN, CNN and RNN.

To our observation and comparison, when fusing Kmer, Psednc, binary, TNC and MonoKGap to extract features, ExtraTree, Fisher Score, mRMR feature selection methods can achieve better prediction performance. The boxplot and ROC curves under different classifiers via the operation of MRMR feature selection can be shown in Figure 2.

G:\feature_selection_package\正文\Figure 2.emf

**Figure 2.** The boxplot and ROC cures of DNA N6-methyladenine sites using various classifiers. (A) 13 classifiers all achieve satisfactory accuracy, and SVM, RandomForest, XGBoost, DNN, RNN obtain superior performance than other classifiers. (B) The ROC curves of 13 classifier indicate ALLFEATURE can obtain accurate prediction results.

## Task 2: RNA N6-methyladenine sites prediction

N6-methyladenosine (m6A) refers to methylation of the adenosine nucleotide acid at the nitrogen-6 position. It plays an important role in a series of biological processes, such as splicing events, mRNA exporting, nascent mRNA synthesis, nuclear translocation and translation process. It is highly desirable to build up an effective predictive model to identify RNA N6-methyladenine sites. In this paper, the dataset contains 2260 sequences, where 1130 represents true methyladenosine sites, and the remaining 1130 are false methyladenosine sites. Firstly, we use Kmer, Psednc, binary, TNC and MonoKgap five feature extraction methods to construct initial feature vectors. Different feature information complements with each other, and multi-information fusion can obtain fully represent the feature information, leading to better model performance and elaborating biological feature information. In order to evaluate the effectiveness of feature selection, we using Lasso, ElasticNet, L1-SVM, CHI2, Pearson Correlation, ExtraTree, XGBoost, SVM-RFE, LOG-RFE, Mutual Information, Minimum Redundancy Maximum Relevance, Joint Mutual Information, Maximum-Relevance-Maximum-Distance, ReliefF, Trace Ratio, Gini index, Spectral Feature Selection, Fisher Score, T Score, Information Gain to determine optimal feature subset, which can eliminate the redundant and noisy features, retain valuable features and reduce run time. The predictive performance of SVM-RFE via different classifiers can be shown in Figure 3.

G:\feature_selection_package\正文\Figure3.emf

**Figure 3.** The boxplot and ROC cures under different classifiers on RNA N6-methyladenine sites dataset. (A) The boxplot show machine learning and deep learning method can better predict RNA N6-methyladenine sites via SVM-RFE feature selection. (B) The 13 classifiers all achieve the good true positive rate at corresponding false positive value, and the AUC values are also high.

## Task 3: Protein-protein interactions prediction, protein site prediction

# CONCLUSION

In this paper, we design a python package and web server for feature extraction, feature selection, machine learning, deep learning and model evaluation called ALLFEATURE. We also test the proposed package using DNA N6-methyladenine sites, RNA N6-methyladenine sites and protein-protein interactions dataset. Experimental results indicate ALLFEAURE is more useful, effective and accurate compared with other state-of-the-art approaches. We design 20 types feature selection methods and 16 types dimensionality reduction approaches to increase computational efficiency and construct better generalization model. Especially, in order to build up biological sequence predictor, ALLFEAURE integrates SVM, KNN, RandomForest, LightGBM, XGBoost, Adaboost, ExtraTree, gaussian naïve bayes, gradient boosting, DNN, CNN and RNN 13 types classifiers to construct sequence analysis model. It is anticipated that ALLFEATURE can be useful tools for bioinformatics and computational biology.

# REFERENCES