[illegible]

Jianwei Li[1, 2, #, *], Xiaoyu Ma[1, #], Xichuan Li[3], Junhua Gu[1, *]

[1]Institute of Computational Medicine, School of Artificial Intelligence, Hebei University of Technology, Tianjin, China

[2]Tianjin Key Laboratory of Bioelectromagnetic Technology and Intelligent Health, Hebei University of Technology, China

[3]Tianjin Key Laboratory of Animal and Plant Resistance, College of Life Sciences, Tianjin Normal University, Tianjin, China

#These authors contributed equally to this work

* Corresponding author

E-mail: lijianwei@hebut.edu.cn, jhgu@hebut.edu.cn

Abstract

Background: The interactions between proteins and aptamers are prevalent in organisms and play an important role in various life activities. Thanks to the rapid accumulation of protein-aptamer interaction data, it is necessary and feasible to construct an accurate and effective computational model to predict aptamers binding to certain interested proteins and protein-aptamer interactions, which is beneficial for understanding mechanisms of protein-aptamer interactions and improving aptamer-based therapies.

Results: In this study, a novel web server named PPAI is developed to predict aptamers and protein-aptamer interactions with key sequence features of proteins/aptamers and a machine learning framework integrated adaboost and random forest. A new method for extracting several key sequence features of both proteins and aptamers is presented, where the features for proteins are extracted from amino acid composition, pseudo-amino acid composition, grouped amino acid composition, C/T/D composition and sequence-order-coupling number, while the features for aptamers are extracted from nucleotide composition, pseudo-nucleotide composition (PseKNC) and Moreau-Broto autocorrelation coefficient. On the basis of these feature sets and balanced the samples with SMOTE algorithm, we validate the performance of PPAI by the independent test set. The results demonstrate that the Area Under Curve (AUC) is 0.907 for prediction of aptamer, while the AUC reaches 0.871 for prediction of protein-aptamer interactions.

Conclusion: These results indicate that PPAI can query aptamers, predict aptamers

and predict protein-aptamer interactions in batch mode precisely and efficiently, which would be a novel bioinformatics tool for the research of protein-aptamer interactions. PPAI web-server is freely available at http://39.96.85.9/PPAI.

Keywords: Aptamer, Protein-aptamer interaction, Sequence features, Adaboost, Random forest

Background

Nucleic acid aptamers against proteins have attracted tremendous attention since they were discovered, the interactions between proteins and aptamers are one of hotspots of biochemistry, molecular biology, bioinformatics and biophysics [1]. Due to the high affinity and specificity of nucleic acid aptamers, protein-aptamer interactions have become more significant for targeted drug therapy of complex diseases and have a perform a variety of functions [2-5]. Aptamers are typically identified *in vitro* from random libraries of DNA or RNA molecules using an iterative process of Systematic Evolution of Ligands by Exponential Enrichment (SELEX) [6], which consists of several repeated rounds of binding, partition and amplification. The aptamers have the merits of easy synthesis and good stability, their specific bindings to proteins play an important role in various life activities of the organisms. Although the experimental aptamer screening technology has been further developed recently, it still has more disadvantages such as time-consuming, expensive and labor-intensive. For this purpose, effective computational methods for predicting aptamers and aptamer-protein interactions are urgent and necessary.

In recent several years, machine learning methods have been widely used in the prediction of protein-aptamer interactions, some computational models have been developed, for example, Li et al. [7] developed a random forest-based protein-aptamer interaction prediction model, Zhang et al. [8] presented a novel model based on the ensemble method in 2016. However, there are still certain limitations in above models. In the Li's model, feature extraction of the sequences was relatively simple and it did not balance the training samples, which resulted in high prediction accuracy for large sample class and low accuracy for small sample class. The same datasets were adopted in Zhang's model and it extracted different features based on the multi-source feature extraction strategy and reconstructed training dataset. To reconstruct the training dataset, the positive and negative samples were split into three groups according to the ratio 1:1, each group was consisted of 580 positive samples and 580 negative samples. These three data sets were facilitated as training sets of three random forest models and the averaged results of the three random forest classifiers were accepted as the final prediction results. Zhang's model also balanced the accuracy of large sample class (i.e. negative samples) and that of small sample class (i.e. positive samples), but the negative samples of each random forest classifier is less due to the split of training data, which led to a decrease of the overall accuracy.

In order to predict aptamers and protein-aptamer interactions more accurately, in response to the above problems, we improved the processing of building datasets and extracting more sequence features, integrated predictive capabilities of two machine learning methods, and developed a novel web server named PPAI. In our study, there

was a prominent imbalance ratio between positive samples and negative samples which could lead to the inherent learning biases [9]. Therefore, the SMOTE [10] method was first to utilize to amplify small sample data for the unbalanced datasets, the balanced dataset could avoid biases in the machine learning. Moreover, PPAI also stores more known protein-aptamer interactions into its database for making user query at easy. In previous studies, more useful features based on structural and evolutionary information did not fully understand or used [7, 8]. Multiple useful features can preserve enough discriminative information for protein-aptamer interaction pairs [11], the combination of various features from different heterogeneous features is a good strategy for enhancing the performance and robustness of a predictor [12]. Based on the multiple feature extraction strategy, multiple key sequence features of proteins and aptamers were synthesized. After analyzing the unique secondary structure characteristics of the aptamers deeply, we screened negative samples and selected the optimal feature set. In general, an ensemble method that integrated diverse learning polices of multiple basic classifiers could outperform its component classifiers [13]. Therefore, an ensemble method combining the adaboost and random forest method was developed to predict protein-aptamer interactions in PPAI.

Results

The performance of protein-aptamer interaction prediction

Experiments were performed to show both the accuracy of our classifier and

effectiveness of feature extraction in depth. The performance comparison based on the same dataset can reflect the performance of a predictor more reliably. To better evaluate performance of PPAI objectively, we compared PPAI with Li's model on the same datasets, using various combination of features and machine learning classifier. The method and features of Li's model given in the published article were utilized to repeat the experiment and reproduce the model. The Receiver-operating characteristics (ROC) curve was drawn through the reproduced model. The detailed prediction results are shown in Table 1 and Figure 1. According to Table 1, PPAI has better predictive performance in terms of area under ROC curve (AUC) by comparing Li's features versus ours features. After balancing training set with the SMOTE method, more balanced sensitivity and specificity were obtained. Moreover, better AUCs could be achieved after introducing SMOTE (Figure 1). We have also adopted the statistical method to test whether such improvement of AUC should be significant by using the pROC package of R software. The result of statistical test has shown that while the improvement of Li's method is not significant (DeLong's test. P=0.471), applying SMOTE indeed significantly enhance the prediction performance of our method (DeLong's test, P=0.0401). The ROC curve was drawn by plotting *Sn* versus *Sp* at different thresholds, it can more intuitively compare the performance of the above models [14]. Where both *F1* value and AUC (Area under ROC curve) are improved, the overall prediction performance is also improved. Furthermore, As the Zhang's ensemble model has no tool or open source code, its threshold is fixed leading to a single point that corresponds to its performance in the ROC curve.

Accuracy and AUC of our model reaches 0.806 and 0.871 respectively, both higher than two previous models. Moreover, our model obtains a more balanced performance with Sn (0.796) and Sp (0.810). To further explore the important features that contribute to the prediction performance, we also extracted the feature importance scores from the model. The top protein features and top nucleotide features are shown in Supplementary Tables S1 and S2, respectively. Notably, features from various encodings constitute the top feature set, emphasizing the importance of using multiple sequence feature encodings.

The performance of aptamer prediction

Besides predicting aptamer-protein interactions, PPAI offered function of predicting aptamers for nucleotide sequences that user input. In order to verify the performance of the aptamer prediction model, the experiment was conducted on an independent test set. To our best knowledge, we cannot find the similar method to be compared with it, so only the performance of PPAI was reported (see Figure 2).

As shown in Figure 2, the accuracy on the independent test set reached 0.847, and the *F1* value reaches 0.849, which suggests the accuracy and practicability of the decision model constructed in this experiment. In order to more intuitively reflect the predictive performance of this model, ROC curve was drawn and the AUC was further calculated to evaluate the performance of aptamer prediction in the independent test (Figure 3), where PPAI has achieved an AUC of 0.907.

Overview of PPAI web server

In order to facilitate the community, we developed a web server named PPAI. Depends on user inputs, PPAI provides three major functions. First, if users input a name or a sequence of a protein/aptamer in the ‘Query’ page of PPAI, the query function is provided. After data collection, we all gathered 704 aptamers and 156 proteins data in PPAI for users to query (see Supplementary Tables S3-S4). If users submit an aptamer name/sequence, the tabular result is firstly provided, which includes the number of protein-aptamer interactions and the associated proteins. For each associated protein in the tabular results, user can also click the interesting protein name to view its details from Uniprot database. User can also input a protein name/sequence, PPAI provides its detail information which involved organism, recommended name, entry id of Uniprot, the number of aptamers and the corresponding aptamer names. For example, the user can enter the name or sequence of the aptamer ‘9572845-PKR protein-3’, and click the ‘submit’ button to get information about the aptamer. The result is shown in Supplementary Figure S1. The user can further click on the name of the protein of interest in the ‘Protein’ column of the result table to get information about that protein. Second, for aptamer prediction, users can choose ‘Predict Aptamer’ page which supports prediction of aptamers from a sequence list of DNA/RNA. The sequences of DNA and RNA should be submitted separately because the threshold values for DNA and RNA feature calculation are different. For example, if the user submits nucleic acid sequences in FASTA format (here we use the sequences of several known aptamer and non-aptamers as the examples), PPAI will give a prediction result for each sequence. It should be noted

that the submitted file name should not contain special characters such as "( )", "." and so on, so as not to cause the program to encounter the file error. The prediction result contains three columns, which are the sequence name, the prediction result (‘yes’ represents that it is an aptamer, and ‘no’ represents that it is not an aptamer), and prediction score (the probability that the sequence is an aptamer). The example result is shown in Supplementary Figure S2. Third, if users have both interesting protein sets and aptamer sets, aptamer-protein interaction prediction is enabled in the ‘Predict Pairs’ page. Users should submit the aptamer file and protein file separately, and PPAI will give a prediction for each possible interaction pairs between each aptamer and each protein. Here we used the aptamers 17030508-Bovinefactor-IX-1 and 9452437-oligoadenylatesynthetase-4, together with 9 proteins as the example input. The result is shown in Supplementary Figure S3. Each prediction result includes aptamer name, protein name, prediction result and predicted score. It should also be noted that the predictions of DNA-protein pairs and RNA-protein pairs should be done separately, because the physicochemical properties of DNA and RNA are not the same during feature extraction. Furthermore, the submitted file name should not contain special characters such as "( )", "." and so on, so as not to cause the program to encounter the file error. One key point for the successful determination of protein-aptamer pair was the selecting of an appropriate threshold. In the simulation experiment with 3000 random sample pairs, we found 0.44 is an optimum threshold to predict the true protein-aptamer pairs, users can also reset the threshold according to their own needs. In general, higher threshold will increase specificity but will also

miss more true positives.

PPAI used MySQL database to store the datasets, and its interface was implemented by HTML and CSS. EasyUI framework was adopted to enhance the page load and response faster. The asynchronous submission and partial refresh mode of PPAI were realized with Jquery+AJAX. The scripting language was C#, both extraction of sequence feature and calculation of predicted score were performed by Python.

Discussion

Through the comparison experiments of different machine learning algorithms and features, the effectiveness of the algorithm and feature space mentioned in this paper is proved. The model constructed in this study for protein-aptamer interaction prediction has higher accuracy and more balanced sensitivity and specificity compared with two previous models, suggesting that PPAI has a fairly good prediction performance in predicting protein-aptamer interactions. The ROC curves of each model in Figure 1 more intuitively reflect that the prediction performance of the PPAI model is superior to other models. Besides, above results effectively demonstrated its potential ability of predicting aptamers, it was beneficial for understanding the functions of aptamers and improving aptamer-based therapies. In addition, based on the current situation of lack of tools for protein-aptamer prediction, a user-friendly PPAI system was developed that provides query functions, protein-aptamer interaction prediction functions, and aptamer judgment functions.

Methods

Datasets

In line with previous studies of predicting aptamer-protein interactions, we also downloaded the datasets constructed by Li [7] which adopted the data from Aptamer Base database [15] (see Supplementary Table S5). It is the largest data set currently available, and it was adopted by most existing methods. Aptamer Base was a collaborative database including protein-aptamer interactions, detailed experimental conditions and reference literatures. The dataset was divided into a training dataset and an independent testing dataset in advance. We first discarded problematic data whose sequence contained B, N, or a mixture of U and T. For easy to compare, the same datasets were adopted in our study, the training set was composed by 561 positive samples, 1682 negative samples, and the test set contained 143 positive samples and 421 negative samples. The positive samples are the protein-aptamer pairs with interaction, and the negative samples are the protein-aptamer pairs without interaction. There was an extremely imbalance between the number of positive samples and the number of negative samples, which would cause biases in the machine learning [16]. Therefore, the SMOTE algorithm [17, 18] was employed to balance the samples in our study. In SMOTE algorithm, the oversampling of the small sample was not done by simply copying the known samples, but by synthesizing new samples according to the feature space which could solve the overfitting problem resulting by simple copy effectively. In order to ensure the validity of the prediction,

the SMOTE method was only utilized to balance the training set, and the independent test set was solely consisted of real samples. After amplifying the small class samples, the training set for predicting the protein-aptamer interactions included 1681 positive samples and 1682 negative samples. In addition, because our models are based on the sequence information of aptamers and proteins, if there is sequence redundancy in the data set, it may cause biases in prediction performance. To check this, we have also used CD-HIT to remove redundant sequences (50% identity threshold for proteins and 80% identity threshold for nucleotides) in the dataset and re-analyzed the performance. The results suggest that the prediction performance is acceptable either before or after removing sequence redundancy, while the better performance of our method can be still observed (Supplementary Table S6).

As for the prediction of aptamers, 704 positive samples and 700 negative samples (350 DNAs and 350 RNAs) were chosen as training dataset and independent testing dataset for predicting the aptamers (Supplementary Table S7). The positive samples refer to known aptamers, and the negative samples refer to randomly generated nucleotide sequences that show highly distinct secondary structure characteristics compared with known aptamers. The most important difference between aptamers and common RNAs/DNAs was that aptamers were easily folded into a pseudoknot, and the stem-ring structures were mostly convex rings and circle rings. Aptamers often had a large contact area to specifically bind to the target molecule with high specificity [19]. In our study, the RNAFold [20] was utilized to predict secondary structure with randomly generated sequences, those sequences that did not conform to

the secondary structure pattern of the aptamers were assigned as the negative samples. Because the accuracy of RNAfold's prediction of secondary structure is about 70%~80%, there might be false negative samples. In order to reduce false negatives as much as possible, the negative samples were screened based on both the minimum free energy and the secondary structure, which could effectively reduce the occurrence of false negative samples. The aptamer has a more stable structure, and its minimum free energy is smaller. The characteristics of the secondary structure of aptamers were fully analyzed, and negative samples were selected from various aspects such as stem-loop structure and number of unpaired bases. These distinctions of secondary structures were obvious and not likely to be confused between aptamers and non-aptamers. The distribution of the lengths of aptamers (DNA or RNA) was shown in Figure 4, about 80% of the sequence lengths were between 30nt and 80nt, the most common aptamer lengths are 40nt, 30nt, 50nt and 80nt. Based on the above, the lengths of the aptamers in the generating 700 negative samples were according with the length assignments of the aptamers in the 704 positive samples.

Feature extraction

Converting an input sample sequence into a set of numerical features is a crucial problem in designing a predictor. Previous studies [7, 8] had shown that pseudo-amino acids and pseudo-nucleotides were effective features for predicting protein-aptamer interaction pairs. The specific binding between aptamers and proteins is closely related to their respective physicochemical properties, which are crucial

factors for their secondary structures [21]. The secondary structures of nucleic acid strands are the main and effective features for distinguishing the aptamers from the common nucleic acid strands. In general, a single feature extraction strategy can only represent partial samples' characteristics, multiple feature extraction strategies can enhance the prediction accuracy. Based on above description, this study combined several key physicochemical features of proteins and aptamers in both the aptamer prediction and the protein-aptamer interaction prediction, and these features were calculated by the iFeature [22] package and the pseKNC [23] package, respectively.

Based on the large numbers of experiments, the considered features of proteins in our study included amino acid composition [24], pseudo-amino acid composition [25], grouped amino acid composition [26], C/T/D composition [27] and sequence-order-coupling number. Amino acid composition means the frequency that is the number of times that each amino acid occurred in the sequences composed by 20 kinds of amino acids. The pseudo-amino acid composition is originally proposed by Chou to predict protein properties [25]. Pseudo amino acid composition has been proved to be an effective feature for many biological problems [7, 28, 29]. Twenty kinds of amino acids are divided into 5 groups in grouped amino acid composition according to their physicochemical properties such as hydrophobicity, charge and molecular size. The groups are non-polar fatty acid amino acids, aromatic amino acids, R-base positively charged amino acids, R and A negatively charged amino acid and an R based uncharged amino acid. Each group is defined as

$$f(g) = \frac{N(g)}{N},\ g \in \{g1, g2, g3, g4, g5\} \quad (1)$$

where $g1$, $g2$, $g3$, $g4$, $g5$ represent aliphatic group, aromatic group, positive charge group, negative charged group and uncharged group and *N(g)* is the number of amino acid in each group.

C/T/D is a pattern of amino acid distributions of specific structural or physicochemical properties in a protein or peptide sequence. C/T/D composition means the ratio of amino acids of specific nature to the total number of amino acids. The physicochemical properties of seven amino acids were used in this study, which were hydrophobic, standardized van der Waals volume, polarity, polarizable, secondary structure, positive/negative charge and solubility. Each attribute is further divided into 3 groups according to its property. The calculation of the attributes is defined as

$$f(r) = \frac{N(r)}{N},\ r \in \{r1, r2, r3\} \quad (2)$$

where $r1$, $r2$, $r3$ represent polar, neutral and hydrophobic and *N(r)* is the number of amino acid type r in the encoded sequence.

Sequence-order-coupling number is the distance between two amino acids calculated by the physicochemical distance matrix of amino acids. The distance matrix used the physicochemical matrix of Schneider-Wrede [30] and the chemical matrix of Grantham [31]. The calculation of each distance matrix is defined as

$$f_d = \sum_i^{N-d}\left(d_{i,i+d}\right),\ d = 1,2,3 \ldots \lambda \quad (3)$$

where $d_{i,i+d}$ is the distance between two amino acids in a given distance matrix, and λ (default is 30) is the maximum distance of the amino acids.

On the other side, the features of aptamers were extracted from nucleotide

composition, pseudo-nucleotide composition (PseKNC) and normalized Moreau-Broto autocorrelation coefficient [32]. The nucleotide composition is the frequency at which each nucleotide (A, C, G, T/U) appears in the sequence. The pseudo-nucleotide composition is a feature proposed based on the pseudo-amino acid composition. The DNA/RNA sequence is converted into a set of discrete values. The calculating method of the pseudo-nucleotide is described in reference [33]. The normalized Moreau–Broto autocorrelation (NMBAC) was proposed by Feng et al. [32] to predict membrane protein types. We used NMBAC to extract features from 11 physical and chemical properties (shift, slide, rise, tilt, roll, twist, stacking- energy, twist, entropy, free energy, hydrophilicity) for protein-aptamer interaction prediction.

PPAI Model based on integrated framework of adaboost and random forest

A novel model for PPAI was developed to predict aptamers and protein-aptamer interactions with a machine learning framework integrated adaboost [34] and random forest [35]. Adaboost combines multiple weak classifiers into the final strong classifier. It would update the sample weights according to each training sample while training. For the misclassified samples, the weights of them are increased, the training set will be trained iteratively. The weight of each weak classifier will be calculated according to the error rate, the higher the error rate, the smaller the weight. Finally, all weak classifiers are weighted and summed to obtain final classification results. Normally, the classification result of each sample is often determined by a classifier with a larger weight. The final model can be calculated from h_t and a_t using follow

formula:

$$H(x) = sign(\sum_{t=1}^{T} a_t h_t(x)) \quad (4)$$

where h_t is the basic classifier and a_t is the weight of it. Furthermore, a_t is calculated by ϵ_t which is the deviation of h_t:

$$a_t = \frac{1}{2} ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \quad (5)$$

The adaboost classifier can often omit some unnecessary training data features and focus on key features [36]. Besides that, the other advantage of adaboost method is that the feature selection process can be omitted. The default basic weak classifier of adaboost algorithm is decision tree which has the shortcomings of the low accuracy and classification efficiency for multi sequence features of proteins/aptamers. However, random forest is a strong classifier that integrates multiple decision trees. In order to improve the performance of predictive model, we adopted adaboost in combination with random forest. The random forest was modified as the basic classifier of adaboost. Furthermore, it was found through experiments that the prediction performance of the protein aptamer interaction was not satisfactory when adopting the adaboost method alone, while the prediction performance was greatly improved by utilizing the combination of adaboost and random forest, and was better than the current methods (Supplementary Table S8). Moreover, the prediction performance of different machine learning algorithms (Bayes, SVM, decision tree, random forest, PPAI) were also compared with the same set of features and the same datasets. The results show that the prediction method proposed in this study (adaboost combined with random forest) has the best prediction performance (see Supplementary

Table S9). Adaboost algorithm was implemented with Python's sklearn package. After parameter optimization, in terms of the parameters of the random forest, the number of trees adopted the value of 10, and the parameter of ‘max_depth’ was set to 150. The adaboost method here mainly has three parameters, namely base_estimator, n_estimators and learning_rate (which determines the end condition of iteration). These three parameters are set to ‘random forest’, 300 and 0.75 respectively.

The extracted feature vectors were used as input to train the model in PPAI, and the obtained model was tested with independent test dataset. The flowchart of protein-aptamer interaction prediction in PPAI is shown in Figure 5. The predict_proba is as return value of the model, which is a real and presents the possibility of positive sample. Furthermore, we could adjust the threshold to optimize the prediction performance of the model. Setting threshold of predicting protein-aptamer interaction was to determine whether a protein and an aptamer interacted with each other. It was judged as ‘yes’ (interaction) when the score was greater than or equal to the threshold, and it was judged as ‘no’ (no interaction) while it was less than the threshold. The smaller the threshold, the higher the sensitivity (*Sn*) and the lower the specificity (*Sp*). When the threshold was 0.44, the *Sn* and *Sp* achieve the best balance, so 0.44 was set to the default threshold of predicting protein-aptamer interaction. In essence, aptamer prediction is a problem of binary classification like the prediction of protein-aptamer interactions. Therefore, the similar machine learning method was also adopted in the model. The threshold was set to determine whether the submitted sequence was an aptamer of one protein. By repetitious experiments,

the best threshold was 0.48, and it was as the default threshold of predicting aptamer in PPAI.

Performance evaluation criterion

The evaluation criteria of prediction performance adopted in this study were sensitivity (*Sn*), specificity (*Sp*), accuracy (*Acc*) and Matthews correlation coefficient (*M*CC). They are the most commonly utilized and basic evaluation index, which can show the prediction accuracy of positive and negative samples and the prediction accuracy of all samples, which can be defined as

$$Sn = \frac{TP}{TP+FN} \tag{6}$$

$$Sp = \frac{TN}{TN+FP} \tag{7}$$

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \tag{8}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{9}$$

where *TP*, *FP*, *TN* and *FN* represent true positive rate, false positive rate, true negative rate and false negative rate, respectively.

Moreover, because the testing dataset was unbalanced which would lead to a biased estimate of the accuracy, accuracy rate did not objectively evaluate the performance of the PPAI. Therefore, the weighted average of the precision rate and the recall rate (*F1*) is introduced as another criterion for performance evaluation, which is currently a widely used and effective evaluation standard for unbalanced data. It is defined as

$$F1 = \frac{2 \times P \times R}{P+R} \tag{10}$$

where *P* and *R* are called Precision and Recall, respectively. While the *R* is equal to

Sn, the Precision is defined as

$$P = \frac{TP}{TP+FP} \quad (11)$$

In the case of imbalanced data sets, the prediction accuracy is often biased toward the accuracy of the larger sample class (i.e. the negative samples for our cases), and cannot objectively reflect the prediction performance of the model. Therefore, ROC curve and AUC metric were introduced as the more appropriate evaluation criteria for such imbalanced dataset. The ROC curve can more intuitively compare the prediction performance of the models. The larger the area under the ROC curve (AUC), the better the prediction performance.

Conclusions

It is important for biology research and drug design to accurately predict aptamers and protein-aptamer interactions by using various kinds of key sequence features of proteins and aptamers. In this paper, a novel ensemble method which is integrated with adaboost and random forest has been developed with a combination of various sequence features extracted from amino acid composition, pseudo-amino acid composition, grouped amino acid composition, C/T/D composition, sequence-order-coupling number, nucleotide composition, pseudo-nucleotide composition (PseKNC) and normalized Moreau-Broto autocorrelation coefficient to predict aptamers and protein-aptamer interactions. In order to solve the imbalance problem effectively, the SMOTE method was adopted to obtain balanced training datasets. To facilitate the community, a web server named PPAI was built with the abstracted sequence features and the machine learning framework mentioned above.

PPAI has a user-friendly interface and step-by-step guide. The reliable performance of PPAI has been demonstrated in verification experiments with independent test datasets, we can draw a conclusion that PPAI is an efficient tool to predict protein-aptamer interactions which is better than the existing mainstream models. Comparing with other models, PPAI has two advantages: (1) More sequence features were introduced, which acquired more discriminative information for the predicitons; (2) The integration of adaboost and random forest, which results in a better performance. However, there exist some limitations, one major limitation is that the process of extracting sequence features is complex and time consuming, which is mainly caused by relative complex extracting algorithm. One solution is to improving things algorithmically and fixing inefficient code. Although limitations exist, we believe the PPAI provides aptamer researchers a valuable and efficient tool to predict protein-aptamer interactions.

Abbreviations

pseudo-nucleotide composition, PseKnc

Area Under Curve, AUC

Systematic Evolution of Ligands by Exponential Enrichment, SELEX

Receiver-operating characteristics, ROC

normalized Moreau–Broto autocorrelation, NMBAC

Support Vector Machine , SVM

sensitivity, Sn

specificity, Sp

accuracy, Acc

Matthews correlation coefficient, MCC

Additional files

Additional file 1: Supplementary Tables 1-7. (XLSX 422 KB)

Supplementary Table S1. The feature importance scores of the top protein features of the model.

Supplementary Table S2. The feature importance scores of the top aptamer features of the model.

Supplementary Table S3. The information of aptamers.

Supplementary Table S4. The information of proteins.

Supplementary Table S5. The datasets of protein-aptamer interactions prediction.

Supplementary Table S6. The performance comparison before and after removing redundancy of the dataset.

Supplementary Table S7. The datasets for predicting aptamers.

Supplementary Table S8. Performance comparison between using adaboost alone and using adaboost and random forest in combination.

Supplementary Table S9. Comparison of prediction performance of different machine learning algorithms for predicting protein-aptamer interactions.

Additional file 2: Supplementary Figures 1-3. (DOCX 1493 KB)

Supplementary Figure S1. Example diagram of query module result of PPAI website.

Supplementary Figure S2. Example diagram of predict aptamer module result of PPAI website.

Supplementary Figure S3. Example diagram of predict protein-aptamer pairs module result of PPAI website.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are included with the article and its additional files.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by National Natural Science Foundation of China [81672113 to J.L.], Natural Science Foundation of Hebei Province [C2018202083 to J.L.].

Authors' contributions

X.M. collected the data and implemented the experiments. J.L. and J.G. conceived and led the project, designed the experiments and wrote the paper. J.L. and X.L. evaluated the methods, suggested improvements and analyzed the results. All the authors have read and approved the final manuscript.

Acknowledgements

We thank Dr. Yuan Zhou from Department of Biomedical Informatics, Peking University for his helpful suggestions. We also appreciate the researchers who shared their protein-aptamer interaction profiles and analysis work.

References

[1] Nimjee SM, White RR, Becker RC, Sullenger BA. Aptamers as Therapeutics. Annu Rev Pharmacol Toxicol. 2017;57:61-79.

[2] Nabavinia M S, Gholoobi A, Charbgoo F, et al. Anti-MUC1 aptamer: A potential opportunity for cancer treatment. Med Res Rev, 2017, 37(6): 1518-1539.

[3] De Franciscis V. Challenging cancer targets for aptamer delivery. Biochimie, 2018, 145: 45-52.

[4] Tan K X, Danquah M K, Sidhu A, et al. Towards targeted cancer therapy: Aptamer or oncolytic virus?. Eur J Pharm Sci, 2017, 96: 8-19.

[5] Liu W, Zhang K, Zhuang L, et al. Aptamer/photosensitizer hybridized mesoporous MnO2 based tumor cell activated ROS regulator for precise photodynamic therapy of breast cancer. Colloids Surf B Biointerfaces, 2019, 184: 110536.

[6] Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science, 1990, 249(4968): 505-10.

[7] Li BQ, Zhang YC, Huang GH, Cui WR, Zhang N, Cai YD. Prediction of aptamer-target interacting pairs with pseudo-amino acid composition. PLoS One 2014; 9: e86729.

[8] Zhang L, Zhang C, Gao R, et al. Prediction of aptamer-protein interacting pairs using an ensemble classifier in combination with various protein sequence attributes. BMC Bioinformatics, 2016, 17(1): 225.

[9] Sanders WS, Johnston CI, Bridges SM, Burgess SC, Willeford KO. Prediction of cell penetrating peptides by support vector machines. PLoS Comput Biol. 2011, 7(7):e1002101.

[10] Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013;14:106.

[11] Zhang YN, Yu DJ, Li SS, Fan YX, Huang Y, Shen HB. Predicting protein-ATP binding sites from primary sequence through fusing bi-profile sampling of multi-view features. BMC Bioinformatics. 2012, 13:118.

[12] Hayat M, Tahir M, Khan SA. Prediction of protein structure classes using hybrid space of multi-profile Bayes and bi-gram probability feature spaces. J Theor Biol. 2014, 346:8-15.

[13] Xie HL, Fu L, Nie XD. Using ensemble SVM to identify human GPCRs N-linked glycosylation sites based on the general form of Chou's PseAAC. Protein Eng Des Sel. 2013, 26(11):735-42.

[14] Chen X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. Sci Rep, 2015, 5:13186.

[15] Cruz-Toledo J, Mckeague M, Zhang X, et al. Aptamer Base: a collaborative knowledge base to describe aptamers and SELEX experiments. Database (Oxford), 2012, 2012: bas006.

[16] Gautam A, Chaudhary K, Kumar R, Sharma A, Kapoor P, Tyagi A, et al. In silico approaches for designing highly effective cell penetrating peptides. J Transl Med. 2013,11:74

[17] Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics, 2013, 14:106.

[18] Tan KX, Danquah MK, Sidhu A, Ongkudon CM, Lau SY. Towards targeted cancer therapy: Aptamer or oncolytic virus? Eur J Pharm Sci. 2017, 96:8-19.

[19] Chen LL, Li J, Zhang XQ, Song L, Qian C, Ge JW. Screening and structure analysis of the aptamer target to Escherichia coli tolC protein. Beijing Da Xue Xue Bao Yi Xue Ban. 2014, 46(5):698-702.

[20] Hofacker I L , Fontana W , Stadler P F , et al. Fast Folding and Comparison of RNA Secondary Structures. Monatsh Chem, 1994, 125(2):167-188.

[21] Delisi C, Crothers D M. Prediction of RNA secondary structure. Proc Natl Acad Sci U S A, 1971, 68(11): 2682-5.

[22] Chen Z, Zhao P, Li F, et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics, 2018, 34(14): 2499-2502.

[23] Chen W, Zhang X, Brooker J, et al. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics, 2015, 31(1): 119-20.

[24] Bhasin M, Raghava G P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem, 2004, 279(22): 23262-6.

[25] Chou K C. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins, 2001, 43(3): 246-55.

[26] Lee T Y, Lin Z Q, Hsieh S J, et al. Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. Bioinformatics, 2011, 27(13): 1780-7.

[27] Dubchak I, Muchnik I, Holbrook S R, et al. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci U S A, 1995, 92(19): 8700-4.

[28] Limongelli I, Marini S, Bellazzi R. PaPI: pseudo amino acid composition to score human protein-coding variants. BMC Bioinformatics, 2015, 16: 123.

[29] Ehsan A, Mahmood M K, Khan Y D, et al. iHyd-PseAAC (EPSV): Identifying Hydroxylation Sites in Proteins by Extracting Enhanced Position and Sequence Variant Feature via Chou's 5-Step Rule and General Pseudo Amino Acid Composition. Curr Genomics, 2019, 20(2): 124-133.

[30] Schneider G, Wrede P. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. Biophys J, 1994, 66(2 Pt 1): 335-44.

[31] Grantham R. Amino acid difference formula to help explain protein evolution. Science, 1974, 185(4154): 862-4.

[32] Feng Z P, Zhang C T. Prediction of membrane protein types based on the hydrophobic index of amino acids. J Protein Chem, 2000, 19(4): 269-75.

[33] Chen W, Feng P M, Deng E Z, et al. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem, 2014, 462: 76-83.

[34] Liu Y, Aleksandrov M, Zlatanova S, et al. Classification of Power Facility Point Clouds from Unmanned Aerial Vehicles Based on Adaboost and Topological Constraints. Sensors (Basel), 2019, 19(21).

[35] Breiman L, Random forests, Mach. Learn, 2001,45(1):5-32.
[36] Wang Y, Zheng B, Xu M, et al. Prediction and analysis of Hub Genes in Renal Cell Carcinoma based on CFS Gene selection method combined with Adaboost algorithm. Med Chem, 2019.

Figure and Table Legends

Table 1. Performance comparison using different combination of machine classifier and features.

Method	*Sn*	*Sp*	*MCC*	*Acc*	*F1*	AUC
Li'RF (Li's features)	0.483	0.871	0.372	0.774	0.517	0.759
PPAI (Li's features)	0.572	0.924	0.538	0.836	0.636	0.827
Li'RF (ours features)	0.458	0.915	0.422	0.800	0.535	0.783
PPAI (ours features)	0.641	0.903	0.557	0.842	0.664	0.849
Li'RF (Smote) (ours features)	0.648	0.818	0.441	0.775	0.592	0.801
PPAI (Smote) (ours features)	0.796	0.810	0.555	0.806	0.675	0.871

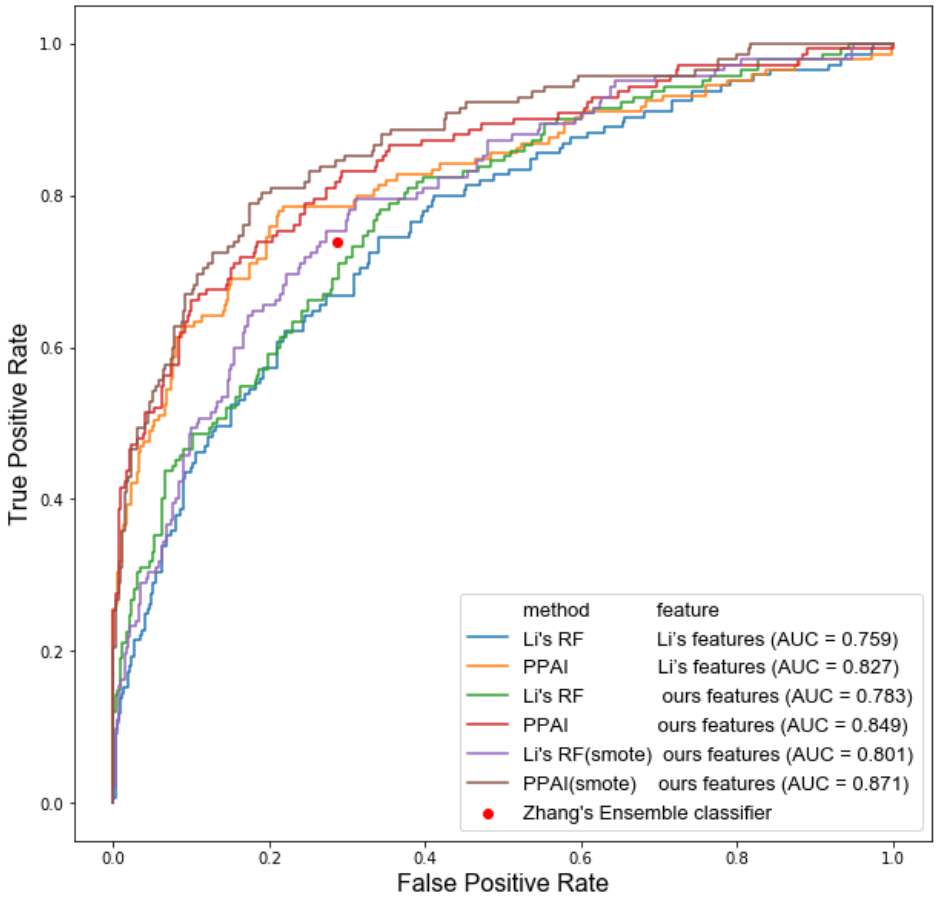

Figure 1. ROC curves illustrating the overall performance comparison results using different combination of machine classifier and features.

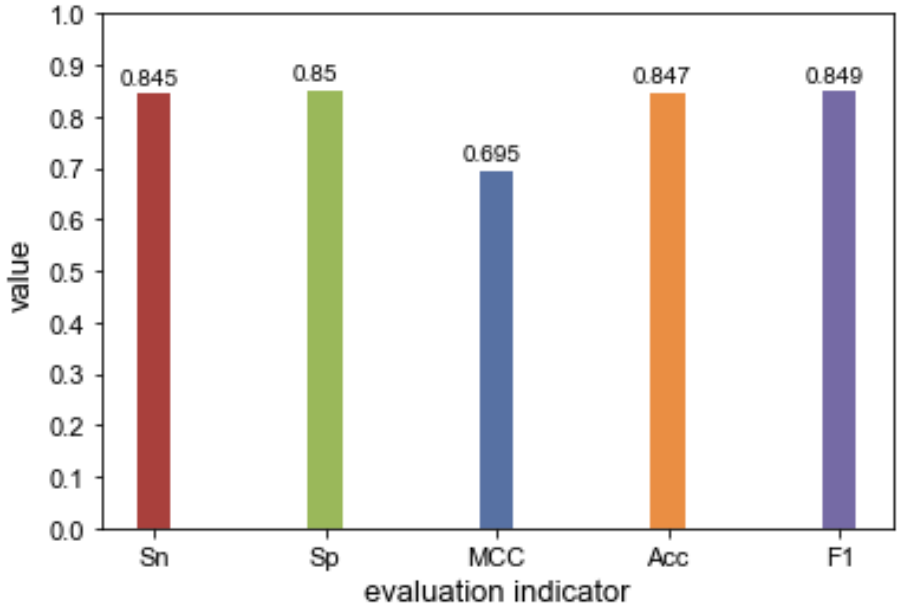

Figure 2. The performance of the aptamer prediction model.

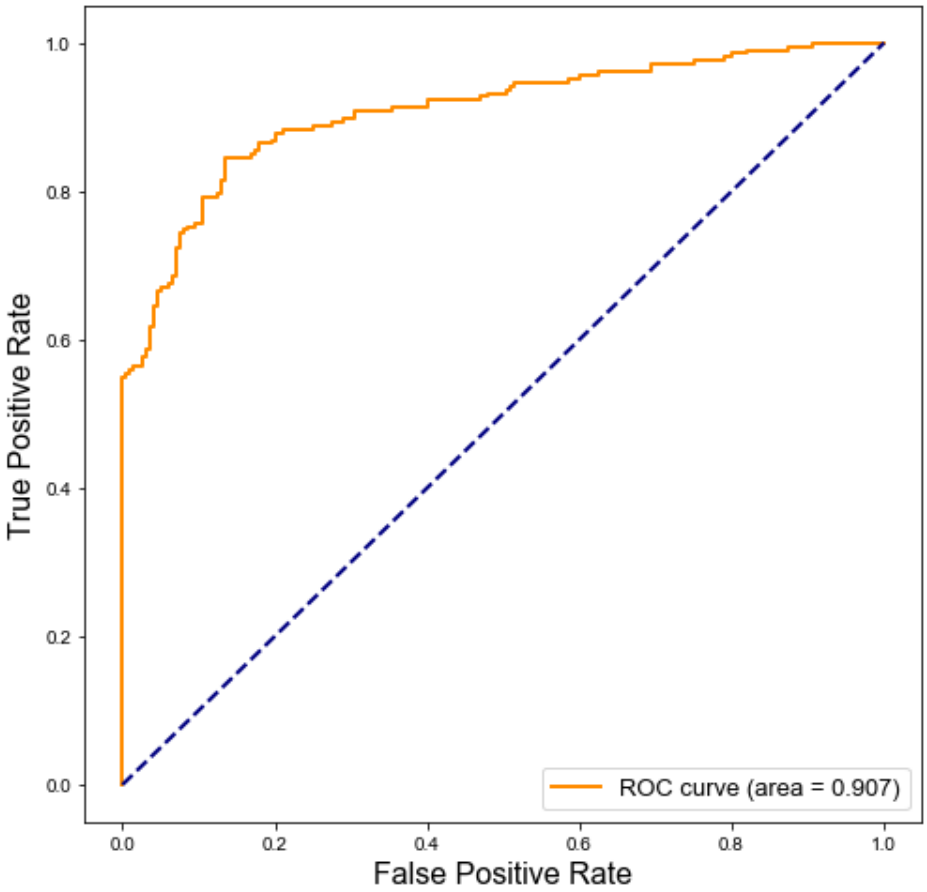

Figure 3. ROC curve of the model of aptamer prediction.

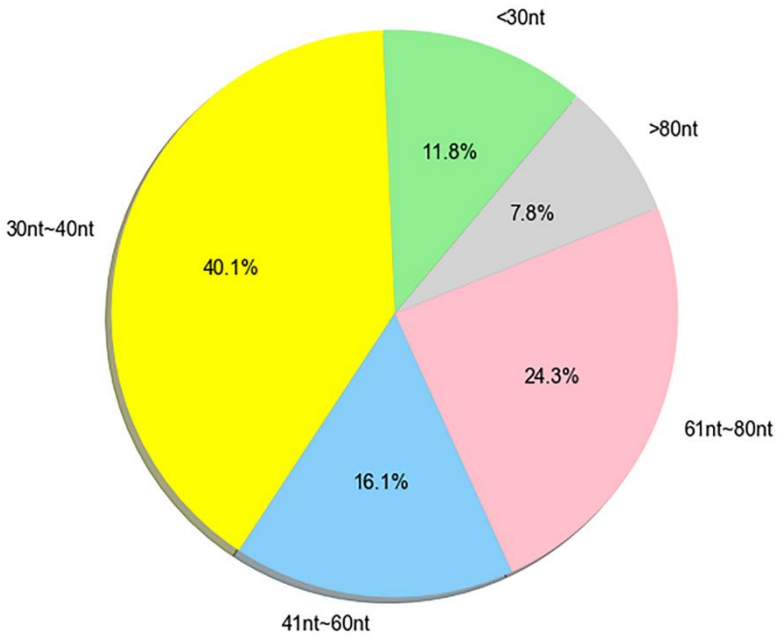

Figure 4. Length distribution of positive sample aptamer sequences.

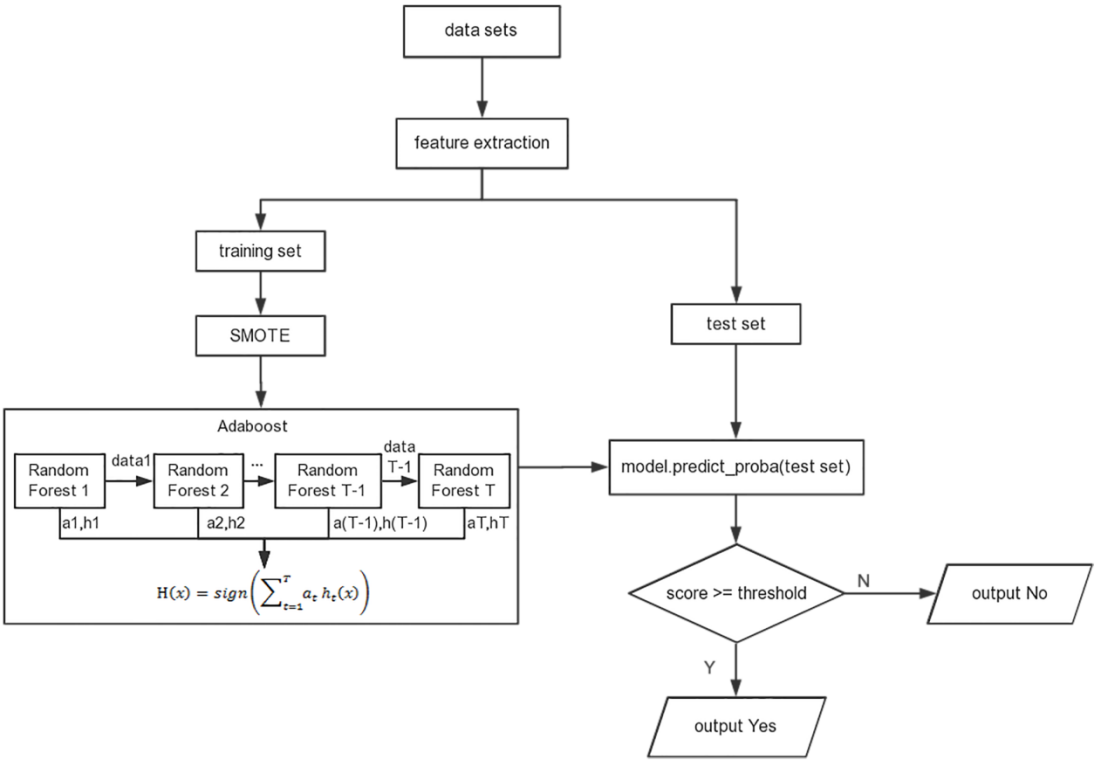

Figure 5. Flowchart of the prediction of protein-aptamer interactions in PPAI.