An *in silico* approach to identification, categorization and prediction of nucleic acid binding proteins

Lei Xu[1], Shanshan Jiang [2], Quan Zou[3,*]

[1]School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen 518000, China

[2] School of Software & Microelectronics, Peking University, Beijing 102600, China

[3]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 651004, China

*To whom correspondence should be addressed.

Abstract

The interaction between proteins and nucleic acid plays an important role in many processes, such as transcription, translation and DNA repair. The mechanisms of related biological events can be understood by exploring the function of proteins in these interactions. The number of known protein sequences has increased rapidly in recent years, but the databases for describing the structure and function of protein have unfortunately grown quite slowly. Thus, improving such databases is meaningful for predicting protein-nucleic acid interactions. Furthermore, the mechanism of related biological events, such as viral infection or designing novel drug targets, can be further understood by understanding the function of proteins in these interactions. The information for each sequence, including its function and interaction sites, were collected and identified, and a database called PNIDB was built. The proteins in PNIDB were grouped into 27 classes, such as transcription, immune system, and structural protein, etc. The function of each protein was then predicted using a machine learning method. Using our method, the predictor was trained on labeled sequences, and then the function of a protein was predicted based on the trained classifier. The prediction accuracy achieved a score of 77.43% by 10-fold cross validation.

Availability and Implementation: PNIDB is now fully working and can be freely accessed at: http://server.malab.cn/PNIDB/index.html. All the data are publicly available for non-commercial use, distribution, and reproduction in any medium.

Contact: zouquan@nclab.net

1 Introduction

As the chief actors within the cells, proteins are involved in many essential activities in the cell, and the interactions between proteins with nucleic acid are extremely important for many cellular processes, such as transcription, translation, and DNA repair. Therefore, the study of protein-nucleic acid binding activities can help with understanding protein interaction networks or even the mechanism of related cellular processes. With the development of sequencing technology, the amount of protein sequence information has increased rapidly over recent years, but the growth rate of databases describing the structure and function of proteins has been very slow, and cannot match the growth rate of protein sequence databases. Thus, it is essential to narrow the gap between sequence database size and functional database size by characterizing nucleic acid binding proteins and their functional groups. Therefore, a database called PNIDB (Protein-Nucleic acid Interactions Database, PNIDB for short), in which DNA-binding proteins and RNA-binding proteins are denoted by gene ontology, was built. Sequence information was extracted, and an efficient classification method for predicting DNA-binding and RNA-binding proteins is proposed in this article. The process of our work is summarized in Fig. 1.

As shown in Fig. 1, the data set was built including DNA-binding and RNA-binding proteins. Each sample was represented by a 473-dimensional vector describing the sequence information and secondary structural information. The parameters of a random forest model were trained using the training samples. The testing samples were then classified by the random forest model with the previously learned parameters.

There are studies focusing on identifying nucleic acid binding proteins and non-nucleic acid binding proteins, and the accuracy of existing methods has been improved over time (Qu & Zou, 2018). However, these methods cannot distinguish the type of binding proteins, such as DNA binding proteins or RNA binding proteins, so it is meaningful to be able to predict the functions of these proteins. Based on existing research, our work focused on identifying and distinguishing DNA-binding proteins and RNA-binding proteins. Since the secondary structure of RNA is diverse, it is difficult to identify common characteristics of RNA binding proteins. As a result, the problem of identifying RNA binding proteins has been rarely considered in previous studies. Thus, using current methods, RNA binding proteins are likely to be recognized as DNA binding proteins. In fact, the function of DNA binding proteins is quite different from the function of RNA binding proteins; thus, the study of distinguishing DNA binding proteins and RNA binding proteins should be considered. In our current study, RNA binding proteins were divided into different classes depending on their characteristics, and the identification of DNA binding proteins was discussed. Furthermore, for the purpose of revealing the biological functions of binding proteins in cellular activities, a protein-nucleic acid binding database called PNIDB was created and can be accessed online. The information, such as

functional classification of protein chain and binding events at the sequence level, are all described in PNIDB. Moreover, the database can predict protein-nucleic acid binding events with the information given by PNIDB.

In contrast to previous work, the nucleic-acid binding proteins were further divided into RNA binding proteins and DNA binding proteins in the proposed method, and both of them were identified by gene ontology in PNIDB. The rationale was that the functions of proteins can be learned thoroughly when the proteins are identified precisely. The contributions of our work include:

(1) A database describing protein-nucleic acid interactions, known as PNIDB, which can be accessed by researchers at the website http://server.malab.cn/PNIDB/index.html. The information provided by PNIDB can help reveal the biological functions of binding proteins in cellular activities. The proteins in PNIDB are labeled by gene ontology identifiers.

(2) An efficient classifier for predicting DNA binding proteins and RNA binding proteins was proposed based on sequence information and secondary structural information, and the accuracy of the proposed method achieved a correlation of 77.43%, which outperforms other methods. The experimental results demonstrated that the combined information could improve the prediction accuracy of our method.

(3) A web server for the prediction of protein-nucleic acid was also developed. The web server was used to predict the function of proteins, which can help researchers studying protein-nucleic acid interactions.

In the rest of the paper, the usage manual PNIDB is introduced in Section 2. Section 3 introduces the methodology for the classification of functional proteins in detail. The experimental results are reported in Section 4. Finally, conclusion are made in Section 5.

2 Usage of PNIDB

There are other databases that provide information regarding protein-DNA/RNA interactions. For example, the Protein-DNA Interface database (PDIdb) (Norambuena, 2010) is a repository containing structural information for 922 protein-DNA complexes with a resolution of 2.5 Å or more (while in fact there are 2396 this kind of complexes in the databaset). The Nucleic acid-Protein Interaction database (NPIDB) (Kirsanov, Zanegina, Aksianov, Spirin, & Alexeevski, 2012) (Olga et al., 2015) contains structural classifications and detailed information on both DNA-protein and RNA-protein complexes extracted from PDB. The current version of NPIDB contains 5046 structures,

while PNIDB contains 6228 PDB structures overall. In contrast to the above databases, proteins are denoted by gene ontology in PNIDB.

PNIDB provides detailed atom-based interaction information. The significance of PNIDB is that it specifically focuses on sequence-level annotations and provides functional clustering, which should be of benefit for sequence-based research and functional prediction of protein-nucleic acid interactions.

PNIDB is a repository of protein-nucleic acid interaction information derived from 6,798 nucleic acid containing structures collected from the Protein Data Bank (Burley et al., 2015). The Bio3d (Skjaerven, Yao, Scarabelli, & Grant, 2014) package was used to read, analyze and manipulate PDB structures in R (http://www.R-project.org/). For each PDB file, we identified the proteins, and the nucleic acid chains were then extracted. Moreover, possible binding residues of the protein chain, which were defined as the residues with at least one atom within 5 Å from any nucleic acid atom, and corresponding binding nucleotides of the nucleic acid chain were also calculated in PNIDB.

Protein chains in interaction pairs were classified according to their mmCIF keywords, interaction type and Gene Ontology (Ashburner, Ball, Blake, Botstein, & Cherry, 2000) terms. There were a total of 84,753 chains extracted from those structures, in which 20,927 chains contained nucleic acid binding residues. All the protein chains were clustered into 27 functional groups, with 17 kinds of DNA binding proteins and 10 kinds of RNA binding proteins. Moreover, each protein chain in these interactions was linked to their respective accession numbers from UniProt as well as the corresponding InterPro identifiers (Finn, 2017) and GO identifiers (Ashburner et al., 2000; Carbon, 2017) mapped from the SIFTS project (Velankar, 2013).

For convenience, the residues and nucleotides are cited by their relative position in the sequence of their separate chains. In addition, the 2D and 3D visualization interfaces are provided online. Fig. 2A and Fig. 2B show the interfaces of the 2D and 3D visualization, respectively, in PNIDB.. In Fig. 2B and 2D, the visualization interface focuses on nucleic acid sequences. The binding protein residue and the position is highlighted in the figure. The residues exceeding 3.9 Å are considered binding residues.

A search page is also provided online, and users can search by keywords. A quick search and advanced search were implemented. In quick search mode, users can start a search by specifying a keyword, such as a PDB ID (Burley et al., 2015) organism, interaction type, classification or Uniprot accession number (Rolf et al., 2004). In the advanced

search mode, for the convenience of other researchers, a web server for binding protein prediction was developed. The web server can handle up to 10 fasta sequences at the same time. Then, the results are returned by email. The web server is shown in Fig. 2C. The results predicted by a learned classifier are shown in Fig. 2D.

The search page of PNIDB is provided, which is shown in Fig. 2E. There are three options, "search by interaction type", "search by organism" and "search by classification". Users can search proteins by describing the requirements. Moreover, proteins can be searched by combining several parameters simultaneously. The matching results will be retrieved when the requirements are submitted. Furthermore, more information can be referred using PNIDB, such as the molecule name of the protein chain, the sequence of the protein chain, the binding residues of the protein chain, the sequence of the nucleic acid chain, the binding nucleotides of the nucleic acid chain, the corresponding InterPro IDs (Finn, 2017) of the protein chain, the GO identifiers of the protein chain, and the 3D visualization with labels of contacting residues and nucleotides based on 3Dmol.js (Rego & Koes, 2015). The schematic diagrams of protein-nucleic acid interactions based on NUCPLOT (Luscombe & N., 1997) can be obtained by clicking the tab on the webpage.

For the convenience of related study, all binding residues/nucleotides were renumbered according to their corresponding chain sequence. In addition, users can also browse the interactions in the browse page by selecting specific classifications of the protein chains in the menu on left side.

3 Methodology

The method for predicting the function of proteins is described in this section. First, the benchmark dataset used is introduced in Section 3.1. Then, the method of feature extraction is described in Section 3.2. Lastly, the classification of the binding proteins is illustrated in Section 3.3.

3.1 Benchmark Data Set

The data used in this work was selected from the SwissProt data set (https://web.expasy.org/docs/swiss-prot_guideline.html). The data in SwissProt contained GO protein sequences, which were selected from the UniProt data set (https://www.uniprot.org/) with high confidence. The SwissProt dataset was composed of DNA binding proteins and RNA binding proteins. The DNA binding proteins had non-IEA source gene ontologies. The sequences that were more similar were removed using CD-HIT(Li & Godzik, 2006). The similarity degree between the

sequences used in our experiments was less than 30%. The benchmark dataset contained five classes. The benchmark dataset used in these experiments is summarized in Table 1.

Tab. 1 Sample detail of benchmark dataset

Class	No. of samples
DNA binding transcription factor	200
Helicase activity	256
Nuclease	200
RBP spliceosomal complex	199
RBP structural constituent of ribosome	213

3.2 Feature Extraction

In the literature, residue sequences are usually represented by a vector $\boldsymbol{v}$ before the process of prediction. An efficient feature set is expected to distinguish positive samples and negative samples with high accuracy. The quality of a feature set is critical to the performance of any predictor. In our method, the sequential evolution information, as well as the local and global secondary structural information, were combined for representing a protein sequence $\boldsymbol{S}$. The features used in our method were extracted from sequence $\boldsymbol{S}$. The features include PSI-BLAST features (Altschul et al., 1997) and PSI-PRED features (Jones, 1999). PSI-BLAST describes the evolutional information, and the secondary structural information is shown by PSI-PRED. The combination of these features has been previously used for protein fold prediction (Wei, Liao, Gao, & Zou, 2015).

A protein sequence S_L is denoted as $S_L=\{R_1,R_2,...,R_L\}$, and L is the length of the residue. PSI-BLAST is based on the protein database *nrdb90* (L. Holm, 1998) and a position-specific score matrix (PSSM). A PSSM is a matrix with L × 20, written as Eq. (1):

$$\begin{bmatrix} M_{1,1} & \cdots & M_{1,20} \\ \vdots & \ddots & \vdots \\ M_{L,1} & \cdots & M_{L,20} \end{bmatrix}_{L\times 20} \tag{1}$$

$M_{i,j}$ is denoted as the score of the residue at the ith position of S_L being mutated to the residue type j during an evolutionary process. The 20 features are computed by the average value of each column (Eq.(2)):

$$F_v = \bar{M}_j = \frac{1}{L}\sum_{i=1}^{L} M_{i,j} \quad j = 1,...,20 \tag{2}$$

The evolution information is also used for describing a sequence. During the evolutionary process, an amino acid located at the *i*th position in the residue may be mutated to type *j*, and the score of it is denoted as $\boldsymbol{M_{ij}}$ (Eq. (3)):

$$M'_{i,j} = 2^{M_{i,j} \times P_j} \tag{3}$$

$\boldsymbol{P_j}$ represents the background frequency of residue type *j*, and the background depends on the average occurrence frequency of all 20 amino acid in each sequence of the protein database *PDB25* (Sussman et al., 2010).

δ_i is represented as the residue located at the *i*th position in an original sequence, $\boldsymbol{S_L}$, replaced by the δ_ith amino acid in the amino acid alphabet. The sequence $\boldsymbol{S_L}$ is transferred into a consensus sequence $\boldsymbol{S_{con}}$ by using δ_i. The frequency of each δ_i in the sequence is denoted as a feature. Thus, there are 20 amino acids, so 20 features are extracted from each sequence.

$\boldsymbol{R_i}$ represents the *i*th residue of a peptide. There are 20 types of native amino acids, which means that there are 20 possible values for R_i. Thus, the number of two consecutive amino acids R_iR_j is 400 (20×20), meaning that the number of dimensions describing the occurrence frequency of two constitute amino acids is 400. The 400-dimensional features have been widely used in the literature of bioinformatics, such as Alzheimer's disease identification (Xu, Liang, Liao, Chen, & Chang, 2018) and detection of anticancer peptides (Xu, Liang, Wang, & Liao, 2018). Thus, sequence information can be revealed using these 400-dimensional features. In contrast to using 400-dimensional features, the features used in our work were based on the consensus sequence S_{con}. The frequency of $\delta_i\delta_j$ is denoted as a feature. The frequency of the occurrence of $\delta_i\delta_j$ is denoted as a feature in $\boldsymbol{v}$. Therefore, another 400 features are extracted from each sequence.

PSI-PRED-based features have been widely used in secondary structure prediction. The features include six structure-sequence based features and $\boldsymbol{a} \times 3^n+3^n$ structure probability matrix-based features. The value of $\boldsymbol{a}$ is set to be 8, and $\boldsymbol{n}$ is 1. Thus, there are 33 PSI-PRED features. Therefore, there are 473 (20+20+400+33) features used to represent a sequence $\boldsymbol{S_L}$ in total.

3.3 Classification

Support vector machine, naive Bayes and ensemble methods have all been widely used in bioinformatics, such as prediction of tumor detection (Tang, Wan, Yang, Teschendorff, & Zou, 2018), function prediction of proteins, and disease detection (Xu, Liang, Liao, Chen, & Chang, 2019). The performance of a predictor is also related to the classifier used. Thus, an efficient classifier is critical for the performance of a computational predictor. In our work, a random forest model was used to predict the function of proteins.

Random forest models are a type of ensemble classifier. The key idea of random forest model is that a number of decision trees are used together for prediction. The decision trees are trained by the datasets, which are built based on bagging. Each decision tree makes a decision, and the final decision is made by a voting process. A sample is then classified into the class with the most votes. The process of a random forest model is shown in Fig.3.

4 Results

To demonstrate the efficiency of our proposed model, our proposed method is compared with other methods, which have been used widely in the literature.

188D is proposed to extract features from a sequence (Cai, Han, Ji, Chen, & Chen, 2003). The amino acid composition, distribution and physicochemical property are described in 188D. This method has been used in bioinformatics, such as for the identification of antioxidant proteins (Xu, Liang, Shi, & Liao, 2018).

The Kmer (k = 2) method extracts features representing the occurrence frequency of k consecutive amino acid in a residue. In our experiments, k was set to 2. The number of dimensions of Kmer features is (n-k)+1, where n is the length of residue.

PC-PseAAC was proposed based on pseudo amino acid components (PseAAC) and has been used in protein identification (Chou, 2011). The information of location residue and global residue are mixed into PseAAC in PC-PseAAC.

An autocorrelation (AC) is the correlation between any two residues with distance lag on the same properties (Liu, Liu, Fang, Wang, & Chou, 2015).

The experiments were based on a 10-fold cross validation. In a 10-fold cross validation, a dataset is divided into 10 parts. Ninety percent of samples are then used for training parameters, and the remaining 10% are considered as testing samples.

The evaluation metric was measured by accuracy (ACC), which was denoted as the rate of correctly classified samples using method ***G***. ACC has been widely used in the literature (Chen, Peng, Han, Cai, & Cai, 2018). The data set was composed of both positive and negative samples. The result set was divided into four parts, which were true positive (TP), false positive (FP), true negative (TN) and false negative (FN). TP is the number of positive samples classified correctly. FP is the number of negative samples labeled as positive samples. TN denotes the number of negative samples that were labeled correctly. FN is the number of positive samples that were recognized as negative samples. The accuracy (ACC) of method ***G*** was computed using Eq. (4):

$$ACC_G = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

The comparison of the feature sets is shown in Fig. 2F. The experimental results show that our proposed method performed better than other feature sets. The combination of sequence information features improved the prediction accuracy. In fact, the features used in Kmer (k=2) were a part of the features in 473D. In the experimental results, the accuracy was improved by nearly 19% (((77.43–65.07)/65.07)%) when the sequences were represented by the PSI-BLAST and PSI-PRED features compared with only using the Kmer (k=2) features. The sequence information was also extracted in 188D, so the accuracy was 0.69, which outperforms other existing methods except 473D. The accuracy of AC and CC were 0.41 and 0.47, respectively. The accuracy of the combination of AC and CC was 0.4625, which was not as good as other methods, such as 188D. In the experiments, the features of AC and CC were not suitable for predicting the function of proteins. The sequence information and secondary structural information were helpful for improving the accuracy and have been used in our proposed method.

5 Conclusions

Due to the rapid growth in the number of protein sequences without identification of their functions, a database describing the protein-nucleic acid interactions (PNIDB) was provided in our work. The functions of sequences were labeled using GO identifiers in PNIDB. PNIDB provides a convenient and user-friendly interface to query and browse detailed information on protein-nucleic acid interactions. Different from existing databases, PNIDB focuses on both protein-DNA and protein-RNA interactions, and the functional classifications are considered at the sequence level. Moreover, a benchmark database is available for the prediction of protein-nucleic acid binding events at either the protein residue level or nucleotide level. PNIDB will also aid in the functional prediction of nucleic-binding proteins based on protein sequence, and may help for providing putative drug targets and novel therapy options. The problem of classification of DNA-binding proteins and RNA-binding proteins was also considered in this work. The sequences are represented by PSI-BLAST features and PSI-PRED features, and a random forest model was used to predict the type of protein, such as DNA-binding proteins and RNA-binding proteins. The accuracy of our proposed method was 0.774, which performs better than other methods. A web server for protein prediction was provided online for the convenience of other researchers. Above all, PNIDB labeled by gene ontology identifiers was built for describing the function of proteins, and a computational predictor was developed for classifying DNA-binding proteins and RNA-binding proteins.

Funding

This work was supported by the Natural Science Foundation of China (Nos. 61902259, 61771331), the Natural Science Foundation of Guangdong province (grant no. 2018A0303130084), and the Science and Technology Innovation Commission of Shenzhen (grant nos. JCYJ20170818100431895).

Conflict of Interest: None declared.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389-3402. doi: 10.1093/nar/25.17.3389

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., & Cherry, J. M. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics, 25*(1), 25-29.

Burley, S. K., Berman, H. M., Christie, C., Duarte, J., Feng, Z., Westbrook, J., . . . Zardecki, C. (2015). RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Science A Publication of the Protein Society*.

Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X., & Chen, Y. Z. (2003). SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Research, 31*(13), 3692-3697.

Carbon, S. e. a. (2017). Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Res., 45*, D331-D338.

Chen, J., Peng, H., Han, G., Cai, H., & Cai, J. (2018). HOGMMNC: a higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification. *Bioinformatics, 35*(4), 602-610. doi: 10.1093/bioinformatics/bty662

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology, 273*(1), 236-247.

Finn, R. D. e. a. (2017). InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research*(45), D190-D199.

Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology, 292*(2), 195-202.

Kirsanov, D. D., Zanegina, O. N., Aksianov, E. A., Spirin, S. A., & Alexeevski, A. V. (2012). NPIDB: Nucleic acid - Protein interaction database. *Nucleic Acids Research, 41*(Database issue).

L. Holm, C. S. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics, 14*, 423-429.

Li, W., & Godzik, A. J. B. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *22*(13), 1658.

Liu, B., Liu, F., Fang, L., Wang, X., & Chou, K. C. (2015). repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics, 31*(8), 1307.

Luscombe, & N. (1997). NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. *Nucleic Acids Research, 25*(24), 4940-4945.

Norambuena, T. a. M., F. (2010). The Protein-DNA Interface database. *Bmc Bioinformatics, 11*, 262.

Olga, Z., Dmitriy, K., Eugene, B., Anna, K., Andrei, A., & Sergey, S. (2015). An updated version of NPIDB includes new classifications of DNA–protein complexes and their families. *Nucleic Acids Research*(D1), D1.

Qu, K., & Zou, Q. (2018). A Review of DNA-binding Proteins Prediction Methods. *Current Bioinformatics, 14*. doi: 10.2174/1574893614666181212102030

Rego, N., & Koes, D. (2015). 3Dmol.js: molecular visualization with WebGL. *Bioinformatics, 31*(8), 1322-1324.

Rolf, A., Amos, B., Wu, C. H., Barker, W. C., Brigitte, B., Serenella, F., . . . Michele, M. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*(suppl_1), suppl_1.

Skjaerven, L., Yao, X.-Q., Scarabelli, G., & Grant, B. J. (2014). Integrating protein structural dynamics and evolutionary analysis with Bio3D. *Bmc Bioinformatics, 15*(1), 1-11.

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., & Abola, E. E. (2010). Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica, 54*(6-1), 1078-1084.

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., & Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinform., 34*(3), 398-406. doi: 10.1093/bioinformatics/btx622

Velankar, S. e. a. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res., 41*, 483-489.

Wei, L., Liao, M., Gao, X., & Zou, Q. (2015). Enhanced Protein Fold Prediction Method Through a Novel Feature Extraction Technique. *IEEE Transactions on Nanobioscience, 14*(6), 649-659.

Xu, L., Liang, G., Liao, C., Chen, G.-D., & Chang, C.-C. (2018). An Efficient Classifier for Alzheimer's Disease Genes Identification. *Molecules, 23*(12), 3140.

Xu, L., Liang, G., Liao, C., Chen, G.-D., & Chang, C.-C. (2019). k-Skip-n-Gram-RF: A Random Forest Based Method for Alzheimer's Disease Protein Identification. *Frontiers in Genetics, 10*(33). doi: 10.3389/fgene.2019.00033

Xu, L., Liang, G., Shi, S., & Liao, C. (2018). SeqSVM: A Sequence-Based Support Vector Machine Method for Identifying Antioxidant Proteins. *International Journal of Molecular Sciences, 19*(6).

Xu, L., Liang, G., Wang, L., & Liao, C. (2018). A Novel Hybrid Sequence-Based Model for Identifying Anticancer Peptides. *Genes, 9*(3), 158.

Fig. 1 The process of protein prediction.

Fig. 2 (A) 3D visualization interface of the interaction between a protein and a nucleic acid residue. (B) 2D visualization interface of an interaction (solid line for hydrogen bond). (C) The web server for binding protein prediction. (D) Results provided by the web-based predictor. (E) Search page on the web server. (F) Comparison with other methods by ACC. (Nucleic acids are labeled in red, while protein residues are in grey, and in the 2D visualization, the dash line denotes residues within 3.9 Å.)

Fig. 3 Process of random forest model generation.

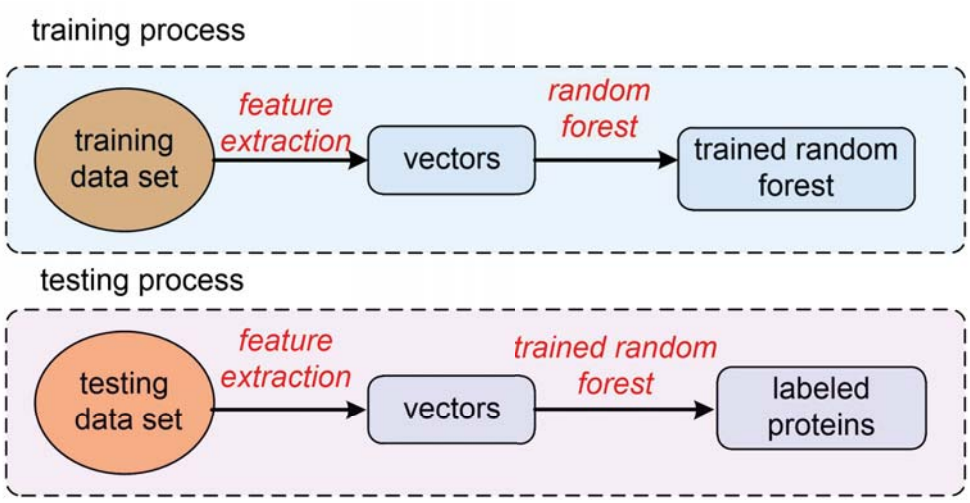

Fig.1

Fig.2

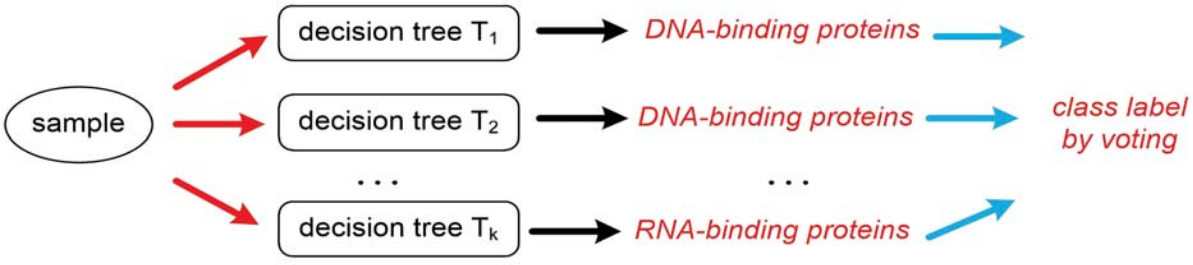

Fig.3