Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2022**; 27(3): 084
http://doi.org/10.31083/j.fbl2703084

Original Research

Comprehensive Prediction of Lipocalin Proteins Using Artificial Intelligence Strategy

Hasan Zulfiqar[1], Zahoor Ahmed[1], Cai-Yi Ma[1], Rida Sarwar Khan[1], Bakanina Kissanga Grace-Mercure[1], Xiao-Long Yu[2,*], Zhao-Yue Zhang[1,*]

[1]School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, 610054 Chengdu, Sichuan, China
[2]School of Materials Science and Engineering, Hainan University, 570228 Haikou, Hainan, China
*Correspondence: yuxiaolong@hainanu.edu.cn (Xiao-Long Yu); zyzhang@uestc.edu.cn (Zhao-Yue Zhang)
Academic Editor: Graham Pawelec
Submitted: 2 December 2021 Revised: 17 January 2022 Accepted: 20 January 2022 Published: 5 March 2022

Abstract

Background: Lipocalin belongs to the calcyin family, and its sequence length is generally between 165 and 200 residues. They are mainly stable and multifunctional extracellular proteins. Lipocalin plays an important role in several stress responses and allergic inflammations. Because the accurate identification of lipocalins could provide significant evidences for the study of their function, it is necessary to develop a machine learning-based model to recognize lipocalin. **Methods**: In this study, we constructed a prediction model to identify lipocalin. Their sequences were encoded by six types of features, namely amino acid composition (AAC), composition of *k*-spaced amino acid pairs (CKSAAP), pseudo amino acid composition (PseAAC), Geary correlation (GD), normalized Moreau-Broto autocorrelation (NMBroto) and composition/transition/distribution (CTD). Subsequently, these features were optimized by using feature selection techniques. A classifier based on random forest was trained according to the optimal features. **Results**: The results of 10-fold cross-validation showed that our computational model would classify lipocalins with accuracy of 95.03% and area under the curve of 0.987. On the independent dataset, our computational model could produce the accuracy of 89.90% which was 4.17% higher than the existing model. **Conclusions**: In this work, we developed an advanced computational model to discriminate lipocalin proteins from non-lipocalin proteins. In the proposed model, protein sequences were encoded by six descriptors. Then, feature selection was performed to pick out the best features which could produce the maximum accuracy. On the basis of the best feature subset, the RF-based classifier can obtained the best prediction results.

Keywords: lipocalins; bioinformatics; feature extraction; optimization; validation

1. Introduction

Lipocalin belongs to the calcyin family and is usually composed of 165–200 residues. It is mainly a stable and multifunctional extracellular protein. Lipocalin proteins can carry aquaphobic molecules such as lipids, steroids and retinoids [1–3]. Lipocalins have important applications in several stress responses, homeostasis, candidate markers of renal functions and allergic inflammations [4–6]. The biological role of lipocalin in human body is shown in Fig. 1.

After the Human Genome Project (HGP), biological sequence data increased significantly [7–9]. The traditional research technology based on biochemistry is time-consuming, expensive and inefficient. Therefore, it is necessary to develop computational methods that can accurately recognize biomolecular functions in a short time [10–12]. Existing computating tools, such as FASTA [13], HAlign [14,15] and BLAST [16], can search sequences with the help of known protein databases. However, these tools cannot correctly distinguish lipocalins when there is no homologous sequence in benchmark dataset. Therefore, it is urgent to establish a machine learning-based model to identify lipocalins. In previous methods, a model called lipocalin-pred [17] was established to recognize lipocalins by using amino acid composition (AAC), reduced AAC [18,19], di-peptide composition (DPC), secondary structure composition (SSC) and position-specific scoring method (PSSM). It could yield an accuracy of 90.72%. Pugalenthi *et al*. [20] proposed a predictor based on support vector machine (SVM). Several features, such as AAC, SSC and physiochemical properties, were used. As a result, they achieved an accuracy of 84.37%. Although the two models can produce encouraging outcomes, there is still room for further improvement.

To further improve the prediction accuracy, we proposed a random forest (RF)-based model to recognize lipocalins. The flowchart of the proposed model was shown in Fig. 2. Initially, the sequences were encoded by six types of features, namely AAC, CKSAAP, GD [21], NMBroto [22], CTD [23] and PseAAC [24]. Subsequently, these features were optimized by using analysis of variance (ANOVA) [25], Maximum Relevance Maximum Distance (MRMD) [26] and minimum Redundancy Maximum Relevance (mRMR) [27] with incremental feature selection

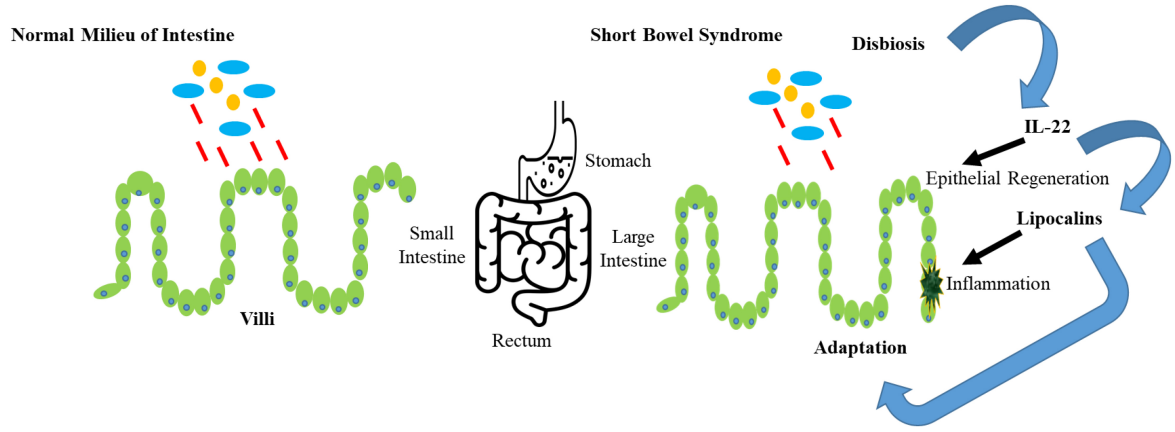

Fig. 1. The biological role of lipocalin in human body.

(IFS) [28]. The RF-based model was trained using the optimal features. The performance of the proposed model was evaluated by 10-fold cross-validation.

2. Materials and Methods

Reliable and accurate dataset is essential for the establishment of prediction model [29–34]. Therefore, we firstly collected 307 positive and 307 negative samples from the open-source database UniProt, and then excluded highly similar sequences using CD-HIT with the cutoff of 40% [35]. Finally, we obtained 211 lipocalins and 211 non-lipocalin proteins. In addition, we also built an independent dataset, including 42 lipocalin protein sequences and 42 non-lipocalin protein sequences, to test the prediction model.

2.1 Feature Descriptors

Choosing informative and autonomous feature is a significant step in generating machine learning-based models [36–41]. Formulating sequence with mathematical expression is also a crucial step in protein function prediction [42–49]. Hence, six types of features were utilized to describe the residues sequences of proteins.

2.1.1 Amino Acid Composition Descriptor

AAC is the frequency of amino acid residues in a protein sequence [50–54]. The frequencies $f(x)$of 20 residues can be calculated by

$$f(x) = \frac{N(x)}{N} \; x \in \{ACDEFGHIKLMNPQRSTVWY\} \quad (1)$$

where $N(x)$ is the x-th residue in a protein sequence with N residues.

2.1.2 Composition of k-spaced Amino Acid Pairs Descriptor

CKSAAP describes the occurrence of amino acid pairs disengaged by any K amino acid (K = 0, 1, 2, 3, 4, 5). It [50] is demarcated as k-spaced residual pairs Q_{xy} which is formulated as

$$Q_{xy} = \frac{N_{xy}}{N-k} \quad (k = 0, 1, 2, 3, 4, 5 \text{ and } xy = \text{ type of } AA) \quad (2)$$

where N_{xy} is the number of residue pairs and 'k' denotes the number of residues. In this work, for saving calculation time, the value of 'k' was set to 3 and the dimension of the features is 1600.

2.1.3 Pseudo Amino Acid Composition Descriptor

It contains the occurrence frequency of amino acids and the correlation of physiochemical properties between two amino acid residues [55]. It comprises of Ac_i and $Ac_{\partial i}$ which can be formulated as

$$Ac_i = \frac{N_i}{1 + \omega \times \sum_{i=1}^{20} \theta_i} \left(\text{here} \theta_i = \frac{\sum_{i=1}^{N-d} (Q_i - Q_{i+d})^2}{N_Q}, (i = 1, 2, 3 \ldots, 20) \right) \quad (3)$$

$$Ac_{\partial i} = \frac{\omega \times \theta_i}{1 + \omega \times \sum_{i=1}^{20} \theta_i}, (\text{ here }, \omega = 0.05) \quad (4)$$

where N_Q is properties number and N_i is the i-th frequency of amino acid. Q_i is the i-th physicochemical property value and θ_i is order factor of protein sequence.

2.1.4 Composition, Transition and Distribution Descriptor

CTD describes the composition, transition and distribution of AAC in a protein sequence [56]. Amino acids are separated into three different classes on the basis of their physiochemical properties [18,57,58]. It is calculated as follows:

$$C_a = \frac{N_a}{N} \quad (a = 1, 2, 3 \ldots) \quad (5)$$

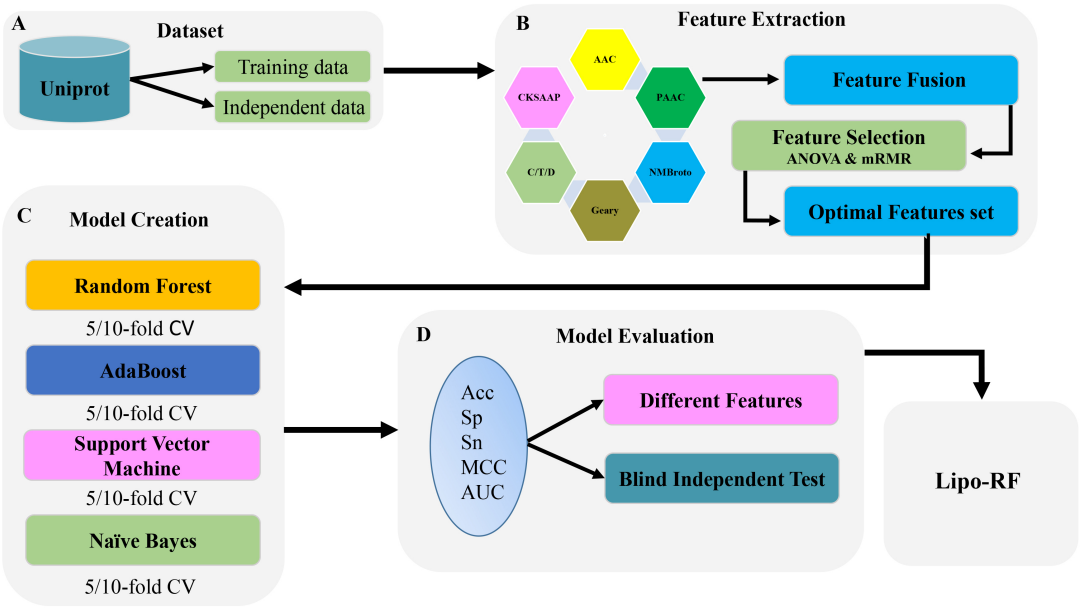

Fig. 2. The flowchart of the whole study. (A) Dataset was constructed and then divided it in to training and testing data. (B) Extracted features by utilizing six types of feature descriptors and then optimized the features by using ANOVA and mRMR. (C) Constructed the models by utilizing different classifiers on 5/10-fold CV. (D) Evaluated the models on independent dataset on the basis of accuracy, specificity, sensitivity, MCC and AUC.

$$T_b = \frac{N_{b,c} + N_{c,b}}{N - 1} \quad (b = 1, 2, 3 \ldots, c \neq b) \qquad (6)$$

$$D_{b,z} = \frac{N_{b,z}}{N} \quad (b = 1, 2 \ldots, \text{ and } z = 1, 0.15N \ldots, N) \qquad (7)$$

where N_a is number of classes, $N_{b,c}$ is the contiguous number of classes b and c and $N_{b,z}$ is the amino acids number which is in the z-th of b-th class.

2.1.5 Geary Descriptor

It is a type of association descriptor and has an extreme resemblance with M-descriptor [59]. The mathematical manifestation is shown as $Q(m)$:

$$Q(m) = \frac{N-1}{2 \times (N-m)} \times \frac{\sum_{i=1}^{N-m} (P_i - P_{i+m})^2}{\sum_{i=1}^{N} (P_i - m)^2} \quad (m = 1, 2, \ldots, 20) \qquad (8)$$

where P_i is i-th amino acids property value in the amino-acid index.

2.1.6 Normalized Moreau-broto Autocorrelation Descriptor

It is a kind of autocorrelation function [22] and has a resemblance with Moran-descriptor. The mathematical manifestation is shown as

$$Q(m) = \frac{\sum_{i=1}^{N-m} (P_i \times P_{i+m})}{N - m} \quad (m = 1, 2, 3, \ldots, 20) \qquad (9)$$

where P_i is i-th amino acids property value in the amino-acid index.

2.2 Feature Selection

Redundancy and noise in feature set may lead to disappointing performance of prediction model [60–62]. Thus, feature selection is a key step to eliminate unimportant features and improve efficiency of prediction model [63–66]. There are several feature selection techniques, such as ANOVA [25], F-score [67], mRMR [27], chi-square [68] and LGBM [69,70]. A feature set with high dimension may produce redundancy, overfitting and yield low accuracy in cross-validation prediction. Hence, ANOVA is a good choice to deal with these problems, because it takes less time and produce effective outcomes. The fusion of feature does not mean that good results can be achieved. These features may be highly correlated, which will lead to the emergence of redundant information in the feature set. Therefore, mRMR is an ideal choice to overcome these problems, because it is able to find the correlation between features. In this work, ANOVA and mRMR [27] were used to rank features. By combining with IFS [67], the optimal feature subset could be obtained. The details about ANOVA, mRMR and IFS can be found in our previous study [24]. The comparison with other different kind of feature selection techniques and the contribution of feature descriptors have been shown in Fig. 3A–D.

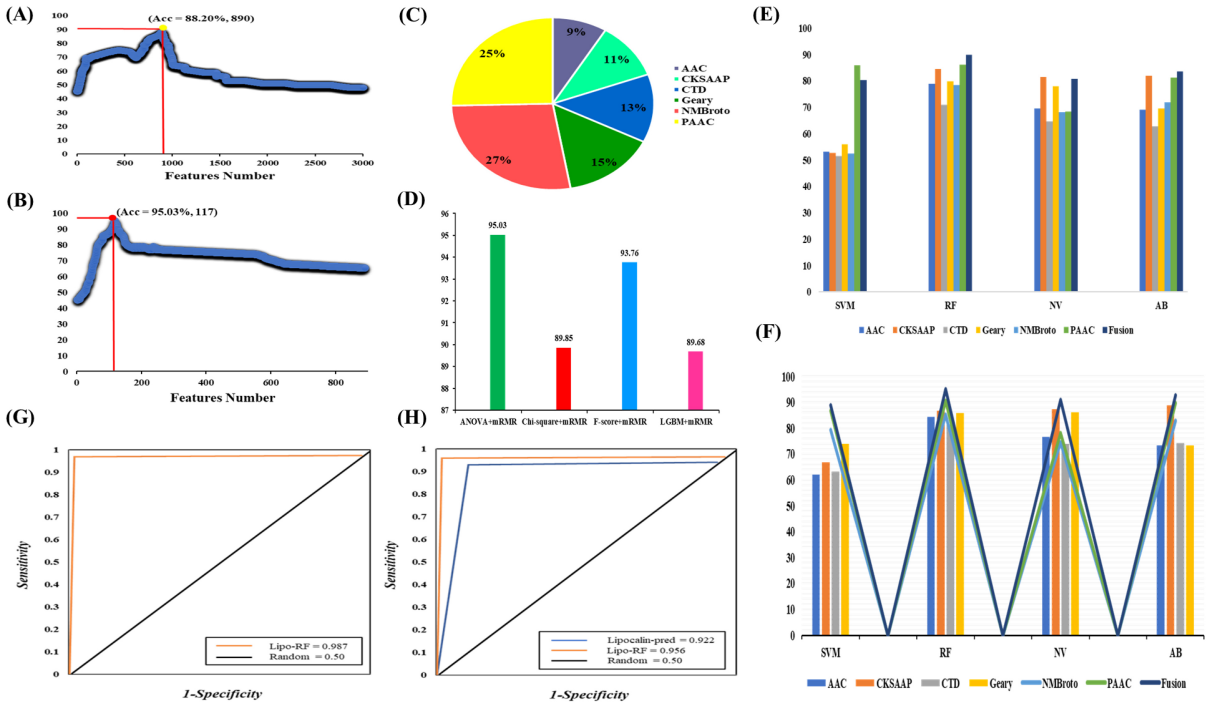

Fig. 3. Plot showing the IFS procedure for identifying lipocalins in 10-fold cross-validation. (A) Firstly, 890 features were selected from a total of 2990 features by ANOVA. (B) 117 optimal features were further obtained from 890 features by mRMR. The accuracy increases from 88.20% to 95.03%. (C) The contribution of different feature descriptors in the model based on fusion features. (D) Compare different feature selection algorithms. (E) The performance of single-encoded features and their fusion on different classifiers before feature selection. (F) The performance of optimal single-encoded features and their fusion on different classifiers. (G) *AUC* value of Lipo-RF on 10-fold cross validation. (H) Comparison of proposed model with Lipocalin-pred on independent dataset.

2.3 Machine Learning Classifiers

Classification is a form of supervised learning and plays an important role in decision making [60,71–80]. In this work, we chose RF to establish a model for recognizing lipocalin. Three machine learning methods namely Naïve Bayes (NB), support vector machine (SVM) [81,82], and Ada boost (AB) [83,84] were compared. RF is a comprehensive knowledge technology, which has been widely used in bioinformatics [85–87]. The principle is to combine multiple weak classifiers and get the results through the voting process, so that the results of the prediction model have the greatest improvement and generalization. The complete procedure has been clearly described in reference [88]. Weka version 3.8.4 (University of Waikato, Hamilton, New Zealand) [89] was utilized to implement the RF-based classifiers. The best parameters were shown in Table 1.

Table 1. Best parameters of the proposed model.

Best Parameters	
'N-estimators'	100
'Learning-rate'	0.001
'Mean absolute error'	0.143
'Kappa statistics'	0.900
'Mean square error'	0.220

2.4 Evaluation Metrics

Accuracy (*Acc*), sensitivity (*Sn*), Matthews correlation coefficient (*MCC*) and specificity (*Sp*) [90,91] were utilized to measure the performance of proposed model, which is expressed by the following formula:

$$\begin{aligned} Sn &= \frac{TP}{TP+FN} \\ Sp &= \frac{TN}{TN+FP} \\ ACC &= \frac{TP+TN}{TP+FP+TN+FN} \\ MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)\times(TN+FN)\times(TP+FP)\times(TN+FP)}} \end{aligned} \tag{10}$$

where true positive indicates lipocalin, and false positive indicates the non-lipocalin classified as lipocalin. On the other hand, true negative represents non-lipocalin, and false negative represents lipocalin classified as non-lipocalin. The receiver operating characteristic (*ROC*) curve and the area under the curve (AUC) were used to measure the effectiveness of the prediction model. AUC of perfect classifier is 1 and AUC of random behavior is 0.5.

Table 2. Performance of optimized single-encodings and fusion models on different machine learning classifiers.

Classifiers	Support Vector Machine					Random Forest					Naïve Bayes					Ada Boost				
Method	Acc	Sp	Sn	MCC	*AUC*	Acc	Sp	Sn	MCC	*AUC*	Acc	Sp	Sn	MCC	*AUC*	Acc	Sp	Sn	MCC	*AUC*
AAC	62.11	74.70	79.60	0.520	0.827	84.25	84.68	87.36	0.703	0.928	76.54	77.00	74.10	0.681	0.894	73.43	74.70	79.60	0.672	0.871
CKSAAP	66.77	65.56	84.55	0.245	0.743	86.79	84.82	90.77	0.723	0.938	87.46	88.11	89.55	0.743	0.922	88.73	89.20	90.30	0.867	0.948
CTD	63.13	78.10	74.55	0.541	0.778	81.32	77.55	88.78	0.712	0.886	73.87	68.00	83.80	0.587	0.843	74.14	73.50	79.50	0.611	0.873
Geary	74.02	77.30	81.40	0.576	0.854	85.93	83.14	88.28	0.745	0.889	86.09	84.52	89.33	0.756	0.889	73.49	70.10	68.90	0.573	0.791
NMBroto	79.38	80.60	82.00	0.623	0.890	85.38	80.50	76.60	0.735	0.876	74.48	81.60	77.20	0.600	0.877	82.81	80.20	79.88	0.705	0.862
PseAAC	86.81	83.70	88.88	0.725	0.862	88.90	83.10	91.80	0.741	0.939	78.12	78.20	88.00	0.621	0.884	89.64	84.55	87.11	0.762	0.905
Fusion	88.71	94.00	92.00	0.862	0.971	95.03	94.00	96.20	0.901	0.987	90.85	88.90	89.70	0.772	0.936	92.78	90.10	90.50	0.873	0.968

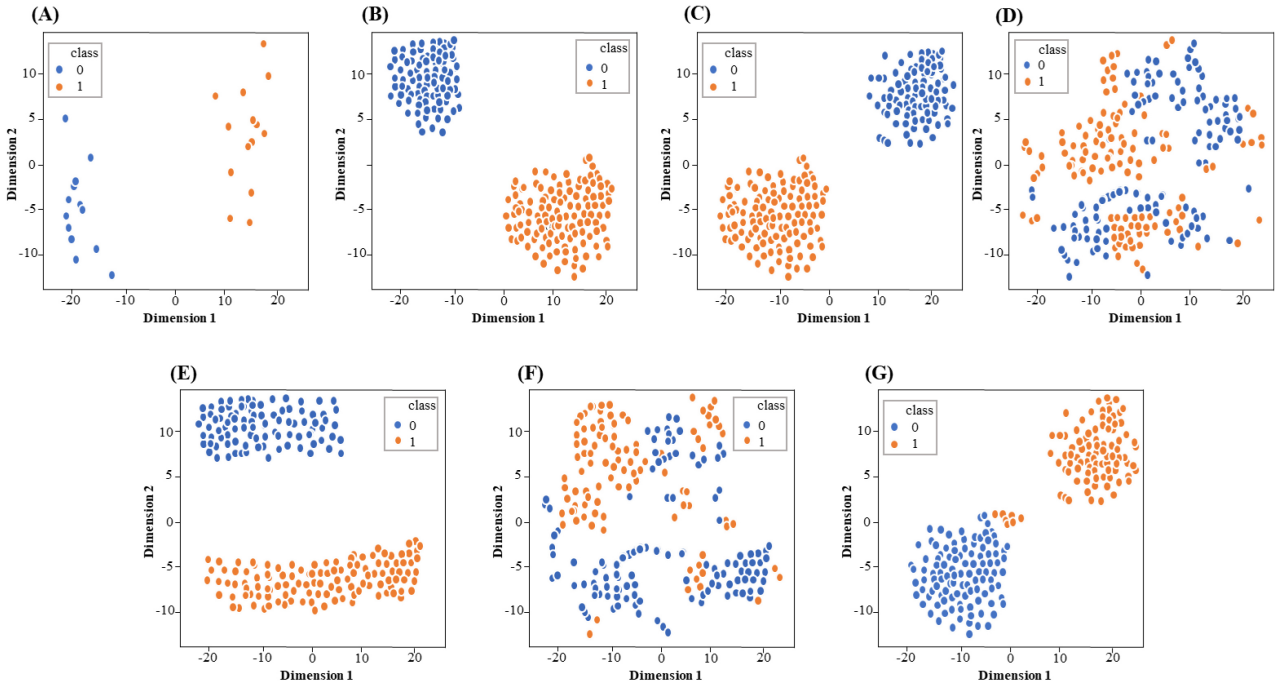

Fig. 4. The visualization of single encoding features and fusion feature through *t*-SNE. (A) AAC, (B) CKSAAP, (C) PseAAC, (D) CTD, (E) GD, (F) NMBroto and (G) fusion feature. Orange and blue represent lipocalins and non-lipocalins, respectively.

3. Results and Discussion

3.1 Performance Evaluation

Firstly, six different feature descriptors were used to transform the training data into feature vectors. Then, through 10-fold cross-validation, RF-based classifier was used to evaluate each feature descriptor. Subsequently, in order to improve the prediction accuracy, ANOVA and mRMR combined with IFS were used to select the optimal feature subset. Fig. 3A,B show the incremental feature selection curves. Fig. 3E,F show the AUC difference of single-encodings and their fusion on different ML-based classifiers before and after the feature selection. Table 2 shows the effectiveness of the improved prediction models based on single-encoding and feature fusion on several ML-based methods. The results of the model based on single-encoding and their fusion on several ML-based classifiers before feature selection have been shown in **Supplementary Table 1** in Supplementary file 1. **Supplementary Fig. 1** in Supplementary file 1 and Fig. 4 show the feature distribution of single-encoding features and fusion features before and after feature screening using *t*-SNE (*t*-distributed stochastic neighbor embedding) technique. The *AUCs* of single-encoding models are 0.928, 0.938, 0.886, 0.889, 0.876 and 0.939 for AAC, CKSAAP, CTD, GD, NMBroto and PseAAC, respectively. The *AUC* of PseAAC is 0.1%–6.3% higher than that of other encoding schemes. On the other hand, the *Acc*, *Sp*, *Sn*, *MCC*, and *AUC* of the feature fusion-based model are 95.03%, 94.00%, 96.20%, 0.901% and 0.987, respectively. The *Acc*, *Sp*, *Sn*, *MCC*, and *AUC* on independent data set are 89.90%, 92.66%, 91.73%, 0.868% and 0.956. The *AUCs* of the feature fusion-based model on training and independent datasets have been shown in Fig. 3G,H.

3.2 Performance Evaluation of Different ML Algorithms

In order to compare a variety of machine learning models, we input single-encoding features and their fusion into other machine learning methods, such as AB, NB and SVM. The 10-fold cross-validation test was used to estimate the efficiency of these models. The comparison results have been shown in Table 2. We noticed that the accuracies of feature-fusion models were higher than those of single-encoding models, demonstrating that a large amount of information can achieve better results. Fig. 3C shows the contribution of the feature descriptors in RF-based fusion model. The model based on the optimal fusion features consists of 117 features from six descriptors, AAC, CKSAAP, CTD, GD, NMBroto and PseAAC, contributed 9 %, 11%, 13%, 15% and 27% in the final optimized-fusion model, respectively. Fig. 3F displays that the RF-based prediction model performs best among all classifiers. The *AUC* of the RF-based model is 1.6%–5.1% higher than that of other classifiers, demonstrating that the RF-based model is suitable for lipocalin proteins prediction.

3.3 Comparison with Existing Model on Independent Dataset

We also compared our model with the existing model on independent dataset to examine the efficiency and per-

Table 3. Comparison between proposed model and the existing method on independent dataset.

Method	Acc	MCC	Sn	Sp	*AUC*	Reference
Lipocalin-Pred	85.73	0.776	88.41	90.11	0.922	[17]
Lipo-RF	89.90	0.868	91.73	92.66	0.956	Our Study

formance of the models. The results on independent dataset show that our model outperformed the existing model by 4.17%. The comparison between our model and the existed model has been shown in Table 3 (Ref. [17]).

4. Conclusions

Lipocalin are responsible for transporting small hydrophobic molecules such as steroids, retinoids, and lipids. They have sequence homology region and common tertiary structure [92,93]. Lipocalins have been applied in several fields, such as stress responses, homeostasis, candidate markers for kidney functions and allergic infections. So for, some models have been developed to identify lipocalins [17,20]. In this work, we developed an advanced computational model to discriminate lipocalin proteins from non-lipocalin proteins. In the proposed model, protein sequences were encoded by six descriptors. Then, feature selection was performed to pick out the best features which could produce the maximum accuracy. On the basis of the best feature subset, the RF-based classifier can obtained the best prediction results. Further studies will focus on creating a user-friendly web server for the prediction model, and will adopt additional feature selection methods and algorithms to further improve the efficiency of lipocalin recognition.

Data Availability Statement

The codes and data can be found in https://zenodo.org/record/5844993#.YeAL7fgRVPZ.

Author Contributions

HZ—Conceptualization, Methodology, Coding, Data curation, Visualization, Writing-Original draft preparation. RSK—Data curation, Methodology. CYM—Data curation, Methodology. ZA—Data curation. BKGM—Data curation. XLY—Reviewing and Editing. ZYZ—Reviewing, Editing and Supervision.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

We are very thankful to Hao Lin Center for Informational Biology, University of Electronic Science and Technology of China for their constructive suggestions and support on this work.

Funding

This work has been supported by the grant from National Natural Science Foundation of China (62102067).

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://www.imrpress.com/journal/FBL/27/3/10.31083/j.fbl2703084.

References

[1] Schiefner A, Skerra A. The Menagerie of Human Lipocalins: a Natural Protein Scaffold for Molecular Recognition of Physiological Compounds. Accounts of Chemical Research. 2015; 48: 976–985.

[2] Romana S, Denisa H, Juraj K, Daniel V, Pavel S. Multiple roles of secretory lipocalins (MUP, OBP) in mice. Folia Zoologica. 2009; 58: 29–40.

[3] Dittrich AM, Meyer HA, Hamelmann E. The role of lipocalins in airway disease. Clinical and Experimental Allergy. 2012; 43: 503–511.

[4] Li C, Chan YR. Lipocalin 2 regulation and its complex role in inflammation and cancer. Cytokine. 2011; 56: 435–441.

[5] Lögdberg L, Wester L. Immunocalins: a lipocalin subfamily that modulates immune and inflammatory responses. Biochimica Et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology. 2000; 1482: 284–297.

[6] Rodvold JJ, Mahadevan NR, Zanetti M. Lipocalin 2 in cancer: when good immunity goes bad. Cancer Letters. 2012; 316: 132–138.

[7] Lee TF. The Human Genome Project: Cracking the genetic code of life. Springer: New York. 2013.

[8] Qi C, Wang C, Zhao L, Zhu Z, Wang P, Zhang S, *et al*. SCovid: single-cell atlases for exposing molecular characteristics of COVID-19 across 10 human tissues. Nucleic Acids Research. 2022; 50: D867–D874.

[9] Liu Y, Zhang X, Zou Q, Zeng X. Minirmd: accurate and fast duplicate removal tool for short reads via multiple minimizers. Bioinformatics. 2021; 37: 1604–1606.

[10] Cheng Y, Gong Y, Liu Y, Song B, Zou Q. Molecular design in drug discovery: a comprehensive review of deep generative models. Briefings in Bioinformatics. 2021; 22: bbab344.

[11] Dong J, Zhao M, Liu Y, Su Y, Zeng X. Deep learning in retrosynthesis planning: datasets, models and tools. Briefings in Bioinformatics. 2022; 23: bbab391.

[12] Song B, Li F, Liu Y, Zeng X. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. Briefings in Bioinformatics. 2021; 22: bbab282.

[13] Pearson WR. Finding Protein and Nucleotide Similarities with FASTA. Current Protocols in Bioinformatics. 2016; 53: 3.9.1–3.9.25.

[14] Zou Q, Hu Q, Guo M, Wang G. HAlign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy. Bioinformatics. 2015; 31: 2475–2481.

[15] Wan S, Zou Q. HAlign-II: efficient ultra-large multiple sequence alignment and phylogenetic tree reconstruction with distributed and parallel computing. Algorithms for Molecular Biology. 2017; 12: 25.

[16] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, *et al*. Gapped BLAST and PSI-BLAST: a new generation

of protein database search programs. Nucleic Acids Research. 1997; 25: 3389–3402.
[17] Ramana J, Gupta D. LipocalinPred: a SVM-based method for prediction of lipocalins. BMC Bioinformatics. 2009; 10: 445.
[18] Zuo Y, Li Y, Chen Y, Li G, Yan Z, Yang L. PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. Bioinformatics. 2017; 33: 122–124.
[19] Pugalenthi G, Kandaswamy KK, Suganthan PN, Archunan G, Sowdhamini R. Identification of functionally diverse lipocalin proteins from sequence information using support vector machine. Amino Acids. 2010; 39: 777–783.
[20] Sokal RR, Thomson BA. Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. American Journal of Physical Anthropology. 2005; 129: 121–131.
[21] Horne DS. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. Biopolymers. 1988; 27: 451–477.
[22] Zhang D, Chen H, Zulfiqar H, Yuan S, Huang Q, Zhang Z, *et al*. IBLP: an XGBoost-Based Predictor for Identifying Bioluminescent Proteins. Computational and Mathematical Methods in Medicine. 2021; 2021: 6664362.
[23] Zulfiqar H, Yuan S, Huang Q, Sun Z, Dao F, Yu X, *et al*. Identification of cyclin protein using gradient boost decision tree algorithm. Computational and Structural Biotechnology Journal. 2021; 19: 4123–4131.
[24] Tang H, Zhao Y, Zou P, Zhang C, Chen R, Huang P, *et al*. HBPred: a tool to identify growth hormone-binding proteins. International Journal of Biological Sciences. 2018; 14: 957–964.
[25] He S, Guo F, Zou Q, HuiDing. MRMD2.0: a Python Tool for Machine Learning with Feature Ranking and Reduction. Current Bioinformatics. 2020; 15: 1213–1221.
[26] De Jay N, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B. MRMRe: an R package for parallelized mRMR ensemble feature selection. Bioinformatics. 2013; 29: 2365–2368.
[27] Yang W, Zhu X, Huang J, Ding H, Lin H. A Brief Survey of Machine Learning Methods in Protein Sub-Golgi Localization. Current Bioinformatics. 2019; 14: 234–240.
[28] Su W, Liu M, Yang Y, Wang J, Li S, Lv H, *et al*. PPD: a Manually Curated Database for Experimentally Verified Prokaryotic Promoters. Journal of Molecular Biology. 2021; 433: 166860.
[29] Ning L, Cui T, Zheng B, Wang N, Luo J, Yang B, *et al*. MNDR v3.0: mammal ncRNA-disease repository with increased coverage and annotation. Nucleic Acids Research. 2021; 49: D160–D164.
[30] Liang ZY, Lai HY, Yang H, Zhang CJ, Yang H, Wei HH, *et al*. Pro54DB: a database for experimentally verified sigma-54 promoters. Bioinformatics. 2017; 33: 467–469.
[31] Zulfiqar H, Masoud MS, Yang H, Han S, Wu C, Lin H. Screening of Prospective Plant Compounds as H1R and CL1R Inhibitors and its Antiallergic Efficacy through Molecular Docking Approach. Computational and Mathematical Methods in Medicine. 2021; 2021: 6683407.
[32] Cheng L, Qi C, Yang H, Lu M, Cai Y, Fu T, *et al*. GutMGene: a comprehensive database for target genes of gut microbes and microbial metabolites. Nucleic Acids Research. 2022; 50: D795–D800.
[33] Mo F, Luo Y, Fan D, Zeng H, Zhao Y, Luo M, *et al*. Integrated Analysis of mRNA-seq and miRNA-seq to Identify c-MYC, YAP1 and miR-3960 as Major Players in the Anticancer Effects of Caffeic Acid Phenethyl Ester in Human Small Cell Lung Cancer Cell Line. Current Gene Therapy. 2020; 20: 15–24.
[34] Zou Q, Lin G, Jiang X, Liu X, Zeng, X. Sequence clustering in bioinformatics: an empirical study. Brief Bioinform. 2020; 21: 1–10.
[35] Zulfiqar H, Sun Z, Huang Q, Yuan S, Lv H, Dao F, *et al*. Deep-4mCW2V: a sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli. Methods. 2021; S1046-2023(21)00198-5.
[36] Zulfiqar H, Dao F, Lv H, Yang H, Zhou P, Chen W, *et al*. Identification of Potential Inhibitors against SARS-CoV-2 Using Computational Drug Repurposing Study. Current Bioinformatics. 2021; 16: 1320–1327.
[37] Guo Z, Wang P, Liu Z, Zhao Y. Discrimination of Thermophilic Proteins and Non-thermophilic Proteins Using Feature Dimension Reduction. Frontiers in Bioengineering and Biotechnology. 2020; 8: 584807.
[38] Tao Z, Li Y, Teng Z, Zhao Y. A Method for Identifying Vesicle Transport Proteins Based on LibSVM and MRMD. Computational and Mathematical Methods in Medicine. 2020; 2020: 8926750.
[39] Cheng L, Hu Y, Sun J, Zhou M, Jiang Q. DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. Bioinformatics. 2018; 34: 1953–1956.
[40] Riaz F, Li D. Non-coding RNA Associated Competitive Endogenous RNA Regulatory Network: Novel Therapeutic Approach in Liver Fibrosis. Current Gene Therapy. 2019; 19: 305–317.
[41] Zhang D, Xu Z, Su W, Yang Y, Lv H, Yang H, *et al*. ICarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. Bioinformatics. 2020; 37: 171–177.
[42] Dao F, Lv H, Zulfiqar H, Yang H, Su W, Gao H, *et al*. A computational platform to identify origins of replication sites in eukaryotes. Briefings in Bioinformatics. 2021; 22: 1940–1950.
[43] Hasan MM, Basith S, Khatun MS, Lee G, Manavalan B, Kurata H. Meta-i6mA: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. Briefings in Bioinformatics. 2021; 22: bbaa202.
[44] Zulfiqar H, Huang QL, Lv H, Sun ZJ, Dao FY, Lin H. Deep-4mCGP: A deep learning approach to predict 4mC sites in Geobacter pickeringii by using correlation-based feature selection technique. International Journal of Molecular Sciences. 2022; 23: 1251.
[45] Basith S, Lee G, Manavalan B. STALLION: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction. Briefings in Bioinformatics. 2022; 23: bbab376.
[46] Basith S, Hasan MM, Lee G, Wei L, Manavalan B. Integrative machine learning framework for the identification of cell-specific enhancers from the human genome. Briefings in Bioinformatics. 2021; 22: bbab252.
[47] Zhai Y, Chen Y, Teng Z, Zhao Y. Identifying Antioxidant Proteins by Using Amino Acid Composition and Protein-Protein Interactions. Frontiers in Cell and Developmental Biology. 2020; 8: 591487.
[48] Hu Y, Qiu S, Cheng L. Integration of Multiple-Omics Data to Analyze the Population-Specific Differences for Coronary Artery Disease. Computational and Mathematical Methods in Medicine. 2021; 2021: 7036592.
[49] Lv Z, Jin S, Ding H, Zou Q. A Random Forest Sub-Golgi Protein Classifier Optimized via Dipeptide and Amino Acid Composition Features. Frontiers in Bioengineering and Biotechnology. 2019; 7: 215.
[50] Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W. ACPred: a Computational Tool for the Prediction and Analysis of Anticancer Peptides. Molecules. 2019; 24: 1973.
[51] Win TS, Malik AA, Prachayasittikul V, S Wikberg JE, Nantasenamat C, Shoombuatong W. HemoPred: a web server for

predicting the hemolytic activity of peptides. Future Medicinal Chemistry. 2017; 9: 275–291.
[52] Win TS, Schaduangrat N, Prachayasittikul V, Nantasenamat C, Shoombuatong W. PAAP: a web server for predicting antihypertensive activity of peptides. Future Medicinal Chemistry. 2018; 10: 1749–1767.
[53] Shoombuatong W, Schaduangrat N, Nantasenamat C. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. EXCLI Journal. 2018; 17: 734–752.
[54] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics. 2001; 43: 246–255.
[55] Dubchak I, Muchnik I, Holbrook SR, Kim SH. Prediction of protein folding class using global description of amino acid sequence. Proceedings of the National Academy of Sciences of the United States of America. 1995; 92: 8700–8704.
[56] Zheng L, Liu D, Yang W, Yang L, Zuo Y. RaacLogo: a new sequence logo generator by using reduced amino acid clusters. Briefings in Bioinformatics. 2021; 22: bbaa096.
[57] Zheng L, Huang S, Mu N, Zhang H, Zhang J, Chang Y, *et al*. RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule. Database. 2019; 2019: baz131.
[58] Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Research. 2007; 36: D202–D205.
[59] Yang H, Luo Y, Ren X, Wu M, He X, Peng B, *et al*. Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators. Information Fusion. 2021; 75: 140–149.
[60] Liu L, Zhang L, Dao F, Yang Y, Lin H. A computational framework for identifying the transcription factors involved in enhancer-promoter loop formation. Molecular Therapy - Nucleic Acids. 2021; 23: 347–354.
[61] Zeng X, Zhu S, Hou Y, Zhang P, Li L, Li J, *et al*. Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. Bioinformatics. 2020; 36: 2805–2812.
[62] Dao F, Lv H, Yang Y, Zulfiqar H, Gao H, Lin H. Computational identification of N6-methyladenosine sites in multiple tissues of mammals. Computational and Structural Biotechnology Journal. 2020; 18: 1084–1091.
[63] Long J, Yang H, Yang Z, Jia Q, Liu L, Kong L, *et al*. Integrated biomarker profiling of the metabolome associated with impaired fasting glucose and type 2 diabetes mellitus in large-scale Chinese patients. Clinical and Translational Medicine. 2021; 11: e432.
[64] Zhao X, Wang H, Li H, Wu Y, Wang G. Identifying Plant Pentatricopeptide Repeat Proteins Using a Variable Selection Method. Frontiers in Plant Science. 2021; 12: 506681.
[65] Yu L, Su Y, Liu Y, Zeng X. Review of unsupervised pretraining strategies for molecules representation. Briefings in Functional Genomics. 2021; 20: 323–332.
[66] Dao F, Lv H, Wang F, Feng C, Ding H, Chen W, *et al*. Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. Bioinformatics. 2019; 35: 2075–2083.
[67] Rachburee N, Punlumjeak W. A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining. 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE). 2015; 420–424.
[68] Lv Z, Wang D, Ding H, Zhong B, Xu L. Escherichia Coli DNA N-4-Methycytosine Site Prediction Accuracy Improved by Light Gradient Boosting Machine Feature Selection Technology. IEEE Access. 2020; 8: 14851–14859.
[69] Lv Z, Zhang J, Ding H, Zou Q. RF-PseU: A Random Forest Predictor for RNA Pseudouridine Sites. Frontiers In Bioengineering And Biotechnology. 2020; 8: 134.
[70] Charoenkwan P, Nantasenamat C, Hasan MM, Manavalan B, Shoombuatong W. BERT4Bitter: a bidirectional encoder representations from transformers (BERT)-based model for improving the prediction of bitter peptides. Bioinformatics. 2021; 37: 2556–2562.
[71] Zulfiqar H, Khan RS, Hassan F, Hippe K, Hunt C, Ding H, *et al*. Computational identification of N4-methylcytosine sites in the mouse genome with machine-learning method. Mathematical Biosciences and Engineering. 2021; 18: 3348–3363.
[72] Govindaraj RG, Subramaniyam S, Manavalan B. Extremely-randomized-tree-based Prediction of N6-methyladenosine Sites in Saccharomyces cerevisiae. Current Genomics. 2020; 21: 26–33.
[73] Manavalan B, Basith S, Shin TH, Wei L, Lee G. MAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics. 2019; 35: 2757–2765.
[74] Manavalan B, Basith S, Shin TH, Lee DY, Wei L, Lee G. 4mCpred-EL: An Ensemble Learning Framework for Identification of DNA N(4)-methylcytosine Sites in the Mouse Genome. Cells. 2019; 8:1332.
[75] Jiang Q, Wang G, Jin S, Li Y, Wang Y. Predicting human microRNA-disease associations based on support vector machine. International Journal of Data Mining and Bioinformatics. 2013; 8: 282–293.
[76] Zhao X, Jiao Q, Li H, Wu Y, Wang H, Huang S, *et al*. ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles. BMC Bioinformatics. 2020; 21: 43.
[77] Wang Y. Delivery Systems for RNA Interference Therapy: Current Technologies and Limitations. Current Gene Therapy. 2020; 20: 356–372.
[78] Lv H, Dao FY, Zulfiqar H, Lin H. DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. Briefings in Bioinformatics. 2021; 22: bbab244.
[79] Lv H, Dao FY, Zulfiqar H, Su W, Ding H, Liu L, *et al.* A sequence-based deep learning approach to predict CTCF-mediated chromatin loop. Briefings in Bioinformatics. 2021; 22: bbab031.
[80] Zhang ZM, Wang JS, Zulfiqar H, Lv H, Dao FY, Lin H. Early diagnosis of pancreatic ductal adenocarcinoma by combining relative expression orderings with machine-learning method. Frontiers in Cell and Developmental Biology. 2020; 8: 582864.
[81] Kuo J, Chang C, Chen C, Liang H, Chang C, Chu Y. Sequence-based Structural B-cell Epitope Prediction by Using Two Layer SVM Model and Association Rule Features. Current Bioinformatics. 2020; 15: 246–252.
[82] Schapire RE. Explaining AdaBoost. In Empirical Inference. 37–52. Springer: Berlin Heidelberg. 2013.
[83] Yu X, Zhou J, Zhao M, Yi C, Duan Q, Zhou W, *et al*. Exploiting XG Boost for Predicting Enhancer-promoter Interactions. Current Bioinformatics. 2020; 15: 1036–1045.
[84] Lv H, Shi L, Berkenpas JW, Dao F, Zulfiqar H, Ding H, *et al*. Application of artificial intelligence and machine learning for COVID-19 drug discovery and vaccine design. Briefings in Bioinformatics. 2021; 22: bbab320.
[85] Zeng X, Zhong Y, Lin W, Zou Q. Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. Briefings in Bioinformatics. 2020; 21: 1425–1436.
[86] Wang H, Liang P, Zheng L, Long C, Li H, Zuo Y. EHSCPr

discriminating the cell identity involved in endothelial to hematopoietic transition. Bioinformatics. 2021; 37: 2157–2164.
[87] Breiman L. Random Forests. Machine Learning. 2001; 45: 5–32.
[88] Janošcová R. Mining Big Data in WEKA. 11th IWKM, Bratislava. Slovakia. 2016; 29–39.
[89] Cao R, Freitas C, Chan L, Sun M, Jiang H, Chen Z. ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. Molecules. 2017; 22: 1732.
[90] Wang X, Yang Y, Liu J, Wang G. The stacking strategy-based hybrid framework for identifying non-coding RNAs. Briefings in Bioinformatics. 2021; 22: bbab023.
[91] Xu B, Liu D, Wang Z, Tian R, Zuo Y. Multi-substrate selectivity based on key loops and non-homologous domains: new insight into ALKBH family. Cellular and Molecular Life Sciences. 2021; 78: 129–141.
[92] Liu D, Li G, Zuo Y. Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. Briefings in Bioinformatics. 2019; 20: 1826–1835.
[93] Wang Z, Liu D, Xu B, Tian R, Zuo Y. Modular arrangements of sequence motifs determine the functional diversity of KDM proteins. Briefings in Bioinformatics. 2021; 22: bbaa215.