9-1-2019

FKRR-MVSF: A Fuzzy Kernel Ridge Regression Model for Identifying DNA-Binding Proteins by Multi-View Sequence Features via Chou's Five-Step Rule

Yi Zou

Yijie Ding

Jijun Tang

Fei Guo

Li Peng

International Journal of
Molecular Sciences

MDPI

Article

FKRR-MVSF: A Fuzzy Kernel Ridge Regression Model for Identifying DNA-Binding Proteins by Multi-View Sequence Features via Chou's Five-Step Rule

Yi Zou [1,2], Yijie Ding [3,*], Jijun Tang [4], Fei Guo [5] and Li Peng [1,2,*]

1 School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China
2 Engineering Research Center of Internet of Things Applied Technology, Ministry of Education, Wuxi 214122, China
3 School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China
4 Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA
5 School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
* Correspondence: wuxi_dyj@tju.edu.cn (Y.D.); pengli@jiangnan.edu.cn (L.P.)

Received: 30 July 2019; Accepted: 19 August 2019; Published: 26 August 2019

Abstract: DNA-binding proteins play an important role in cell metabolism. In biological laboratories, the detection methods of DNA-binding proteins includes yeast one-hybrid methods, bacterial singles and X-ray crystallography methods and others, but these methods involve a lot of labor, material and time. In recent years, many computation-based approachs have been proposed to detect DNA-binding proteins. In this paper, a machine learning-based method, which is called the Fuzzy Kernel Ridge Regression model based on Multi-View Sequence Features (FKRR-MVSF), is proposed to identifying DNA-binding proteins. First of all, multi-view sequence features are extracted from protein sequences. Next, a Multiple Kernel Learning (MKL) algorithm is employed to combine multiple features. Finally, a Fuzzy Kernel Ridge Regression (FKRR) model is built to detect DNA-binding proteins. Compared with other methods, our model achieves good results. Our method obtains an accuracy of 83.26% and 81.72% on two benchmark datasets (PDB1075 and compared with PDB186), respectively.

Keywords: DNA-binding proteins prediction; fuzzy kernel ridge regression; multiple kernel learning; feature extraction; protein sequence

1. Introduction

The interaction between DNA and protein exists in various tissues of the living body. For example, DNA–protein interactions during many activities such as DNA replication, DNA repair, DNA packaging, DNA modification, and viral infection. The study of DNA binding residues in DNA–protein interactions facilitates a comprehensive understanding of the mechanisms of chromatin recombination and gene-regulated expression. The methods of detecting DNA-binding proteins are mainly deployed by biochemistry and physical chemistry methods. However, wet experiment-based methods are both time and money consuming.

The protein information of 3D structures or their complexes is important for drug design. X-ray crystallography is expensive and time-consuming [1–3]. Lots of sequence-based information, such as PTM (posttranslational modification) sites in proteins [4–9], DNA-methylation sites [10], protein–drug interaction in cellular networking [11], protein–protein interactions [12] and

recombination spots [13], have been predicted by sequential tools such as Pseudo Amino Acid Composition (PseAAC) [14] and Pseudo K-tuple Nucleotide Composition (PseKNC) approach [15]. Bioinformatics has played important roles in the development of novel drugs.

Computational methods based on Machine Learning (ML) have been developed to predict DNA-binding proteins. Currently, ML technology is playing key roles in lots of biological field, including prediction of DNA methylcytosine sites [16,17], O-GlcNAcylation sites [18], potential disease-associated microRNAs [19,20], protein remote homology [21], protein subcellular localization [22], electron transport proteins [23] and analyzing microbiology [24] et al. The computational methods can be classified into two types of methods: sequence-based models and a structure-based models.

The sequence-based methods extract features from protein sequences and employ ML to build predictive models. PseAAC and Support Vector Machine (SVM) [25] were used to construct a model for identifying DNA-Binding Proteins [26]. Kumar et al. [27] used Position Specific Scoring Matrix (PSSM) of protein sequences to develop an SVM classifier called DNAbinder. The PSSM describes proetin sequences. PSI-BLAST [28] can calculate PSSM for target protein. Liu et al. [29] proposed iDNAPro-PseAAC model, which employed PseAAC and PSSM features. Wei et al. [30] used local PSSM features to represent local information of proteins. Sequence-based approachs can implement large-scale predictions.

Structure-based models employ structure features to predict DNA-binding proteins. Compared with sequence-based methods, structure-based models achieve better performance. The main reason is that 3D structure of proteins determine the shape and surface area of the protein. Nimrod et al. [31] used the average surface electrostatic potentials of the protein to build a Random Forest (RF) model to predict DNA-binding proteins. Due to the known structures being less than sequences, the structure-based models can not predict all proteins.

In recent publications [32–35] and two review papers [36,37], researchers developed useful predictors for bioinformatics. Many methods obeyed a rule, called Chou's five-step rule. This rule contains five steps: (1) a benchmark dataset is constructed to train and test the predictive models; (2) the selected samples should truly reflect their correlation of the target; (3) the prediction problem can be solved by a powerful algorithm; (4) the cross-validation tests are performed to evaluate the performance of the methods; (5) building a web-server for the predictive model. The above rule is clear in logic, and completely transparent in operation. This rule can easily repeat the reported results by other researchers and is very convenient for the experimental scientists. Our method is also based on Chou's five-step rule.

To avoid losing the sequence–pattern information of proteins, the PseAAC [14,36,38] was proposed by Chou. Chou's general PseAAC [36] has been widely used to extract features from sequence and PSSM of protein. In addition, a useful web-server called "Pse-in-One2.0" [39,40] has been established. The server can extract feature vectors for DNA/RNA and protein/peptide sequences. We also emply Pse-in-One2.0 to extract features from protein sequences.

In this study, we propose a novel model via a Fuzzy Kernel Ridge Regression model based on Multi-View Sequence Features (FKRR-MVSF) to predict DNA-binding proteins. The multiple sequence features are extracted and constructed to multiple kernels, respectively. Next, a Multiple Kernel Learning (MKL) algorithm linearly weights these kernels. Fuzzy membership scores of each training sample are calculated by an integrated kernel. Finally, Fuzzy Kernel Ridge Regression (FKRR) is trained to predict DNA-binding proteins.

2. Results

To evaluate our proposed method (FKRR-MVSF), two benchmark datasets of DNA-binding proteins are employed in our study. First of all, we analyze the performance of different features. Then, our model is compared with other methods via a Jackknife test. Finally, an independent test set is used to test the robustness of FKRR-MVSF.

2.1. Data Sets

In our study, two benchmark datasets (PDB1075 and PDB186 datasets) are used to test our predictive model of DNA-binding proteins. PDB1075 and PDB186 were collected from the Protein Data Bank (PDB) [41]. Liu et al. [26] randomly extracted non-DNA-binding and DNA-binding proteins from the PDB database. The similarity of any two sequences does not exceed 25%. A total of 525 DNA-bind proteins and 550 non-DNA-binding proteins form the PDB1075 dataset. PDB186 dataset [42] contains 93 DNA-bind and 93 non-DNA-bind proteins. Table 1 lists the information of the two benchmark data sets.

Table 1. The detail information of two benchmark data sets.

Data Sets	PDB1075	PDB186
Positive	525	93
Negative	550	93
Total	1075	186

2.2. Measurements

Accuracy (ACC), Sensitivity (SN), Specificity (SP) and Matthew's Correlation Coefficient (MCC) are used to evaluate the performance of predictive model. These coefficients are calculated as follows:

$$ACC = 1 - \frac{N_{-}^{+} + N_{+}^{-}}{N^{+} + N^{-}} \tag{1a}$$

$$SN = 1 - \frac{N_{-}^{+}}{N^{+}} \tag{1b}$$

$$Spec = 1 - \frac{N_{+}^{-}}{N^{-}} \tag{1c}$$

$$MCC = \frac{1 - (\frac{N_{-}^{+}}{N^{+}} + \frac{N_{+}^{-}}{N^{-}})}{\sqrt{(1 + \frac{N_{+}^{-} - N^{+}}{N^{+}})(1 + \frac{N_{-}^{+} - N_{+}^{-}}{N^{-}})}} \tag{1d}$$

where N^{+} and N^{-} are the total number of positive and negative samples, respectively. N_{+}^{-} and N_{-}^{+} are the number of false positive and false negative, respectively. And Area Under ROC curve (AUC) is also an effective evaluation method for binary classification.

2.3. Performance Analysis of Different Features on the PDB1075 Data Set

The single type feature can not fully describe the properties of a protein, so we build the predictive model with multi-view sequence features to represent the protein. We test (Jackknife test evaluation) these features (kernels) on the PDB1075 dataset, as shown in Table 2. The PSSM-based features (PSSM-AB and PsePSSM feature) achieve better performance than non-PSSM (MCD and NMBAC feature) single features. The performance (MCC) of MCD, NMBAC, PSSM-AB and PsePSSM feature are 0.4139, 0.4564, 0.5113 and 0.5886, respectively. In addition, mean weighted kernels (KRR) combines the above 4 kernels (features) via average weight and obtains better performance (MCC: 0.6398) than single feature. Compared with mean weightes (KRR), MKL (KRR) achieves a higher value of MCC (0.6439). FKRR weighs training sets by fuzzy membership, which can filter outliers. So, mean weights (FKRR) (MCC: 0.6554) and MKL (FKRR) (MCC: 0.6664) are both better than KRR because of using multiple kernel information and fuzzy membership. Moreover, MKL (FKRR) achieves a better MCC of 0.6664.

In addition, we test the SVM model with different features on the PDB1075 dataset. In Table 2, the performance (MCC) of SVM (with MKL, MCC: 0.6568) is better than KRR (with MKL, MCC: 0.6439). However, the MCC (0.6568) of SVM (with MKL) is slightly lower than FKRR (with MKL, MCC: 0.6664). The reason may be the fuzzy membership for building predictor. The ROC curve also reflects the

excellent performance of MKL (FKRR) in Figure 1. Our method (FKRR-MVSF) employs MKL and FKRR to build a final predictor for DNA-binding proteins.

Table 2. The performance of different features on the PDB1075 dataset (Jackknife test).

Feature Type	Model	ACC	SN	Spec	MCC	AUC
MCD	KRR	0.7070	0.7086	0.7088	0.4139	0.7751
NMBAC	KRR	0.7284	0.7181	0.7382	0.4564	0.7857
PSSM-AB	KRR	0.7553	0.7695	0.7418	0.5113	0.8352
PsePSSM	KRR	0.7944	0.7905	0.7982	0.5886	0.8637
MW [a]	KRR	0.8195	0.8362	0.8036	0.6398	0.8998
MKL	KRR	0.8214	0.8438	0.8000	0.6439	0.9032
MCD	SVM	0.7088	0.7345	0.6819	0.4171	0.7611
NMBAC	SVM	0.7116	0.6909	0.7333	0.4244	0.7706
PSSM-AB	SVM	0.7693	0.6981	0.8438	0.5467	0.8391
PsePSSM	SVM	0.7851	0.7472	0.8247	0.5731	0.8566
MW [a]	SVM	0.8201	0.8232	0.8170	0.6421	0.9011
MKL	SVM	0.8299	0.8541	0.8057	0.6568	0.9101
MW [a]	FKRR	0.8270	0.8533	0.8018	0.6554	0.9094
MKL	FKRR	0.8326	0.8571	0.8091	0.6664	0.9115

[a] MW denotes combining kernels by the mean weights.

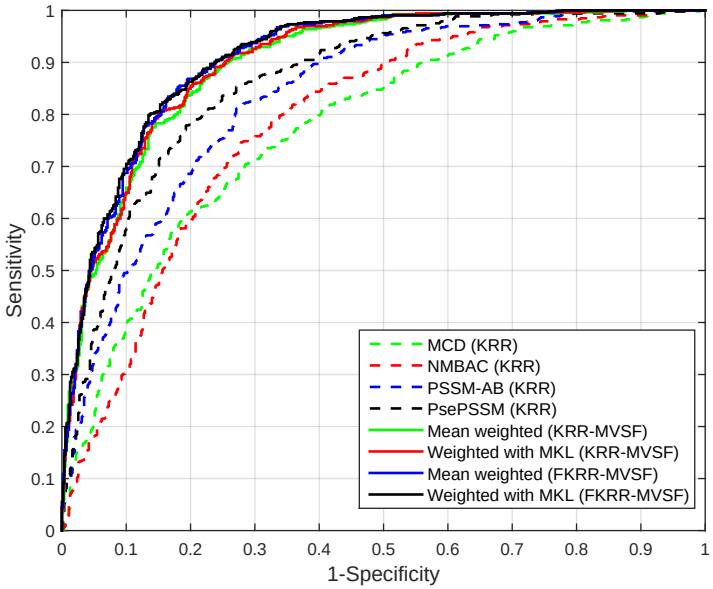

Figure 1. The ROC curve of different kernels (features) on the PDB1075 dataset (Jackknife test).

Figure 2 shows the weight of each feature. The highest weight of feature is PsePSSM, which has a similar trend of their single feature performance. To reduce bias of features, the MKL algorithm can estimate the optimal weights of features.

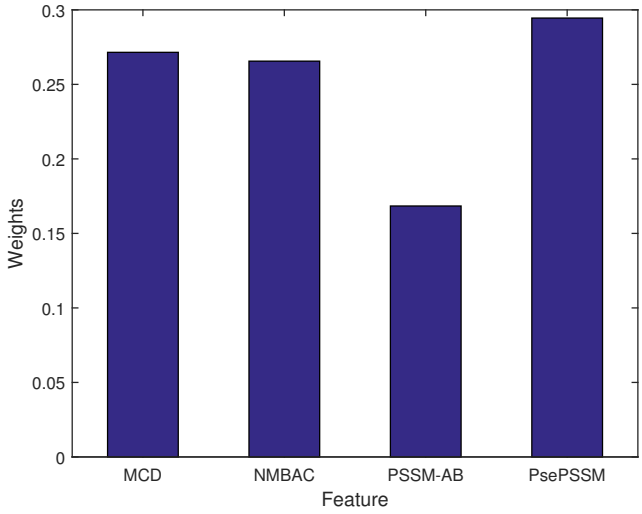

Figure 2. The weights of different kernels (features).

We test our method and other existing methods on the PDB1075 dataset. Table 3 lists the results of comparison between our method and other methods. PseDNA-Pro [26], IDNA-Prot|dis [29], IDNA-Prot [43], DNAbinder [27], DNA-Prot [44], iDNAPro-PseAAC [45], Local-DPP [30], Adilina's work [46] and Kmer1+ACC [47] are benchmark methods. And IDNA-Prot|dis (MCC: 0.54), PseDNA-Pro (MCC: 0.53) iDNAPro-PseAAC (MCC: 0.53) and Local-DPP (MCC: 0.59) obtain better performance. Our proposed model (FKRR-MVSF) obtains best MCC (0.67) on the PDB1075 data set.

Table 3. Comparison between our method and other existing methods on the PDB1075 dataset (Jackknife test).

Methods	ACC (%)	MCC	SN (%)	Spec (%)
IDNA-Prot	75.40	0.50	83.81	64.73
DNAbinder	73.95	0.48	68.57	79.09
DNA-Prot	72.55	0.44	82.67	59.76
iDNAPro-PseAAC	76.56	0.53	75.62	77.45
IDNA-Prot|dis	77.30	0.54	79.40	75.27
Kmer1+ACC	75.23	0.50	76.76	73.76
Local-DPP	79.10	0.59	84.80	73.60
PseDNA-Pro	76.55	0.53	79.61	73.63
Adilina's work	70.21	0.41	61.00	79.70
Our method (FKRR-MVSF)	83.26	0.67	85.71	80.91

2.4. Performance on an Independent DataSet of PDB186

In order to evaluate the generalization performance of predictive models, FKRR-MVSF and other methods are also tested on the independent dateset (training set is PDB1075). The results are shown in Table 4.

Our method (FKRR-MVSF) achieves 81.7% of ACC, 0.676 of MCC and 98.9% of SN. In MCC, FKRR-MVSF is better than Local-DPP (MCC: 0.625), DBPPred (MCC: 0.538), MSFBinder [48] (MCC: 0.640), Adilina's work (MCC: 0.670) and iDNAPro-PseAAC (MCC: 0.442).

Table 4. Compared with existing methods on the PDB186 dataset (Independent test).

Methods	ACC (%)	MCC	SN (%)	Spec (%)
IDNA-Prot	67.2	0.344	67.7	66.7
DNA-Prot	61.8	0.240	69.9	53.8
IDNA-Prot|dis	72.0	0.445	79.5	64.5
DNAbinder	60.8	0.216	57.0	64.5
DBPPred	76.9	0.538	79.6	74.2
Kmer1+ACC	71.0	0.431	82.8	59.1
iDNAPro-PseAAC	71.5	0.442	82.8	60.2
Local-DPP	79.0	0.625	92.5	65.6
Adilina's work	82.3	0.670	95.0	69.9
MSFBinder (SVM)	81.7	0.640	89.3	74.2
Our method (FKRR-MVSF)	81.7	0.676	98.9	64.5

3. Discussion

To improve the performance of predicting DNA-binding proteins, we employ an MKL algorithm and fuzzy-based model to integrated different features and further handle the outliers, respectively. There are many ways in machine learning to avoid overfitting and generating skewed models caused by outliers, e.g., adjustment of the cost value in SVM. For different training samples, the parameter of cost should be different. Different samples have different contributions to the model. In Table 2, the performance (MCC: 0.6664) of fuzzy-based models (FKRR with MKL) is better than non-fuzzy models (KRR with MKL, MCC: 0.6439).

Compared to other single kernels, the PsePSSM-based kernel achieves the highest weight and highest value of MCC (0.5886). MKL could integrate multiple information of sequence. Our method (KRR with MKL) also achieves better performance of MCC (0.6439) than a single kernel model on the PDB1075 dataset. In addition, the performance of KRR with MKL (MCC: 0.6439) is better than KRR with mean weights (MCC: 0.6398) under PDB1075 dataset.

On the independent test dataset, our method (FKRR with MKL) also achieves better MCC (0.676). MSFBinder (SVM) [48] is a two-layer model with SVM. MSFBinder (SVM) also employed several features to build a predictive model. The generalization performance of FKRR (withe MKL) is better than MSFBinder (MCC: 0.640) on an independent test set (PDB186). The above two models are similar. The main reason of different results is that the parameter C of FKRR is different for each train sample. Fuzzy membership may reduce the effect of some noise samples in the model.

4. Materials and Methods

The prediction of DNA-binding proteins can be regarded as a task of binary classification. The protein can be represented by some feature vectors. The DNA-binding proteins and non-DNA-binding proteins are labeled as $+1$ (positive samples) and -1 (negative samples), respectively. We construct a Fuzzy Kernel Ridge Regression model based on Multi-View Sequence Features (FKRR-MVSF) to determine whether a protein binds to DNA. We employ Normalized Moreau–Broto Auto Correlation (NMBAC) [49,50], PSSM based Average Blocks (PSSM-AB) [51], Multiple-scale Continuous and Discontinuous descriptor (MCD) [52] and PsePSSM algorithms to extract four types of PSSM-based features. Radial Basis Function (RBF) is used to build four types of kernels from the above four kinds of features. In our study, the MKL algorithm is employed to calculate the weights of kernels and to combine four kernels. Then, a membership score is estimated for each training sample. Finally, a fuzzy kernel ridge regression model for identifying DNA-binding proteins is constructed via membership scores and a combined kernel. The framework of proposed method is showed in Figure 3. In the literature [13,33], the researchers have made good use of flowcharts to describe the main framework of their methods. In our work, we employ Figure 4 to describe the flow of our model. Firstly, we extract four types of feature from a sequence. Then, Radical Basis Function

(RBF) is used to build four kernels. These kernels are conbined by MKL. Finally, combined kernel and training labels are employed to construct the FKRR model and predict new samples.

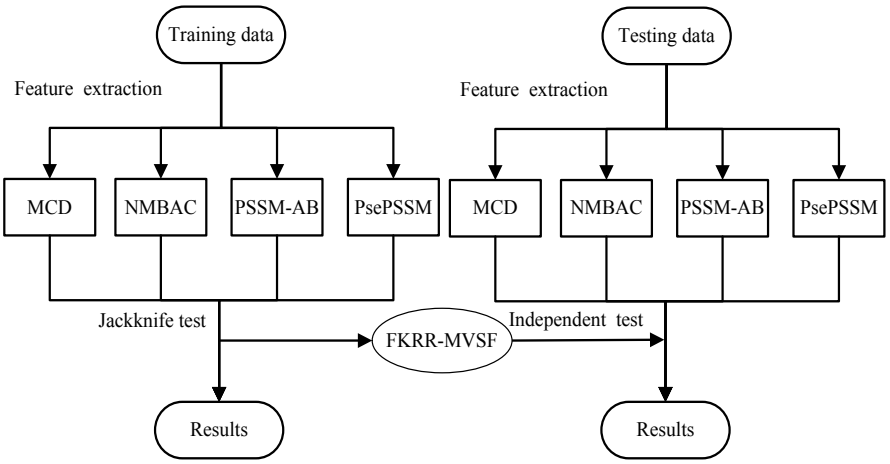

Figure 3. The process of DNA-binding protein prediction.

4.1. Feature Extraction

Extracting features from proteins is a challenge for identifying DNA-binding proteins. A suitable feature extraction algorithm can adequately represent the properties of the protein. We use four types of feature to describe a protein.

4.1.1. MCD Feature

You et al. clustered the 20 amino acids into seven groups according to dipoles and volumes of side chains. These groups are {A, G, V}, {C}, {F, I, L, P}, {D, E}, {H, N, Q, W}, {K, R} and {M, S, T, Y}. A protein sequence "AVDCALSK" can be described as "11321476" via Multi-scale Continuous and Discontinuous descriptor (MCD) [52]. Then, above sequence was split into 10 local regions, which described multiple overlapping continuous and discontinuous interaction patterns. Composition (C), Transition (T) and Distribution (D) were calculated in each local region. The detailed descriptions of MCD algorithm can refer to You's work [52]. The MCD feature was 882-dimentional vector.

4.1.2. NMBAC Feature

Normalized Moreau–Broto Auto Correlation (NMBAC) [49,50] was proposed for extracting the sequence feature of membrane proteins. A protein sequence (string) can be represented as discrete numerical sequence via six physicochemical properties of Amino Acids (AA): including Hydrophobicity (H), Net Charge Index of Side Chains (NCISC), Solvent-Accessible Surface Area (SASA), Volumes of Side Chains of amino acids (VSC), Polarity (P1) and Polarizability (P2), respectively. The six physicochemical properties of amino acids are list in Table 5. To extract the feature of a protein X with L-length, the NMBAC feature is calculated by following equation:

$$NMBAC(lag, j) = \frac{1}{(n - lag)} \sum_{i=1}^{n-lag} (X_{i,j} \times X_{i+lag,j}) \quad (2)$$

where i denote the position in the sequence, and $i = 1, 2, ..., n - lag$. j is the type of physicochemical properties, $j = 1, 2, ..., 6$. $lag \in [1, lg]$ is the gap between amino acids. lg is a parameter of maximum distance.

Table 5. The values of the 6 properties for twenty amino acids.

Amino Acid	H	VSC	P1	P2	SASA	NCISC
A	0.62	27.5	8.1	0.046	1.181	0.007187
C	0.29	44.6	5.5	0.128	1.461	−0.03661
D	−0.9	40	13	0.105	1.587	−0.02382
E	−0.74	62	12.3	0.151	1.862	0.006802
F	1.19	115.5	5.2	0.29	2.228	0.037552
G	0.48	0	9	0	0.881	0.179052
H	−0.4	79	10.4	0.23	2.025	−0.01069
I	1.38	93.5	5.2	0.186	1.81	0.021631
K	−1.5	100	11.3	0.219	2.258	0.017708
L	1.06	93.5	4.9	0.186	1.931	0.051672
M	0.64	94.1	5.7	0.221	2.034	0.002683
N	−0.78	58.7	11.6	0.134	1.655	0.005392
P	0.12	41.9	8	0.131	1.468	0.239531
Q	−0.85	80.7	10.5	0.18	1.932	0.049211
R	−2.53	105	10.5	0.291	2.56	0.043587
S	−0.18	29.3	9.2	0.062	1.298	0.004627
T	−0.05	51.3	8.6	0.108	1.525	0.003352
V	1.08	71.5	5.9	0.14	1.645	0.057004
W	0.81	145.5	5.4	0.409	2.663	0.037977
Y	0.26	117.3	6.2	0.298	2.368	0.023599

4.1.3. PSSM-AB Feature

Position Specific Scoring Matrix (PSSM) contains evolutionary information of protein sequence. The PSSM of protein sequence is generated by PSI-BLAST [28]. PSSM is a $L \times 20$ matrix (L rows and 20 columns):

$$PSSM = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,20} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,20} \\ \vdots & \ddots & \vdots & \vdots \\ P_{L,1} & P_{L,2} & \cdots & P_{L,20} \end{bmatrix}_{L \times 20} \tag{3}$$

PSSM-AB extracts local average values of PSSM:

$$PSSM - AB(k) = \frac{20}{L} \sum_{z=1}^{L/20} PSSM(z + (i-1) \times L/20, j) \tag{4}$$

where k is a linear index used to scan the cells of PSSM. $i, j = 1, 2, ..., 20, k = j + 20 \times (i-1)$. The PSSM-AB algorithm can extract the information of relationship between target residue and neighboring residues.

4.1.4. PsePSSM Feature

PsePSSM [53] is an effective feature based on PSSM. PSSM $\in L \times 20$ is standardized as following:

$$PSSM'(i,j) = \frac{PSSM(i,j) - mean(PSSM(i,*))}{STD(PSSM(i,*))} \quad i = 1, 2, ..., L; j = 1, 2, ..., 20 \tag{5}$$

where $STD(PSSM(i,*))$ denotes the standard deviation of the elements. $mean(PSSM(i,*))$ represents the mean of the elements that are located in the i-th row. $*$ denotes the all elements of the i-th row. Then, we obtain the PsePSSM feature as the following:

$$Pse(k) = \begin{cases} \frac{1}{L}\sum_{i=1}^{L} PSSM'(i,j) & k = 1,...,20 \\ \frac{1}{L-lag}\sum_{i=1}^{L-lag}\left[PSSM'(i,j) - PSSM'(i+lag,j)\right]^2 & \\ j = 1,...,20; lag = 1,...,15; & \\ k = 20 + j + 20 \times (lag-1) & \end{cases} \quad (6)$$

where k is index of feature vector and lag denotes the distance between one residue and its neighbors.

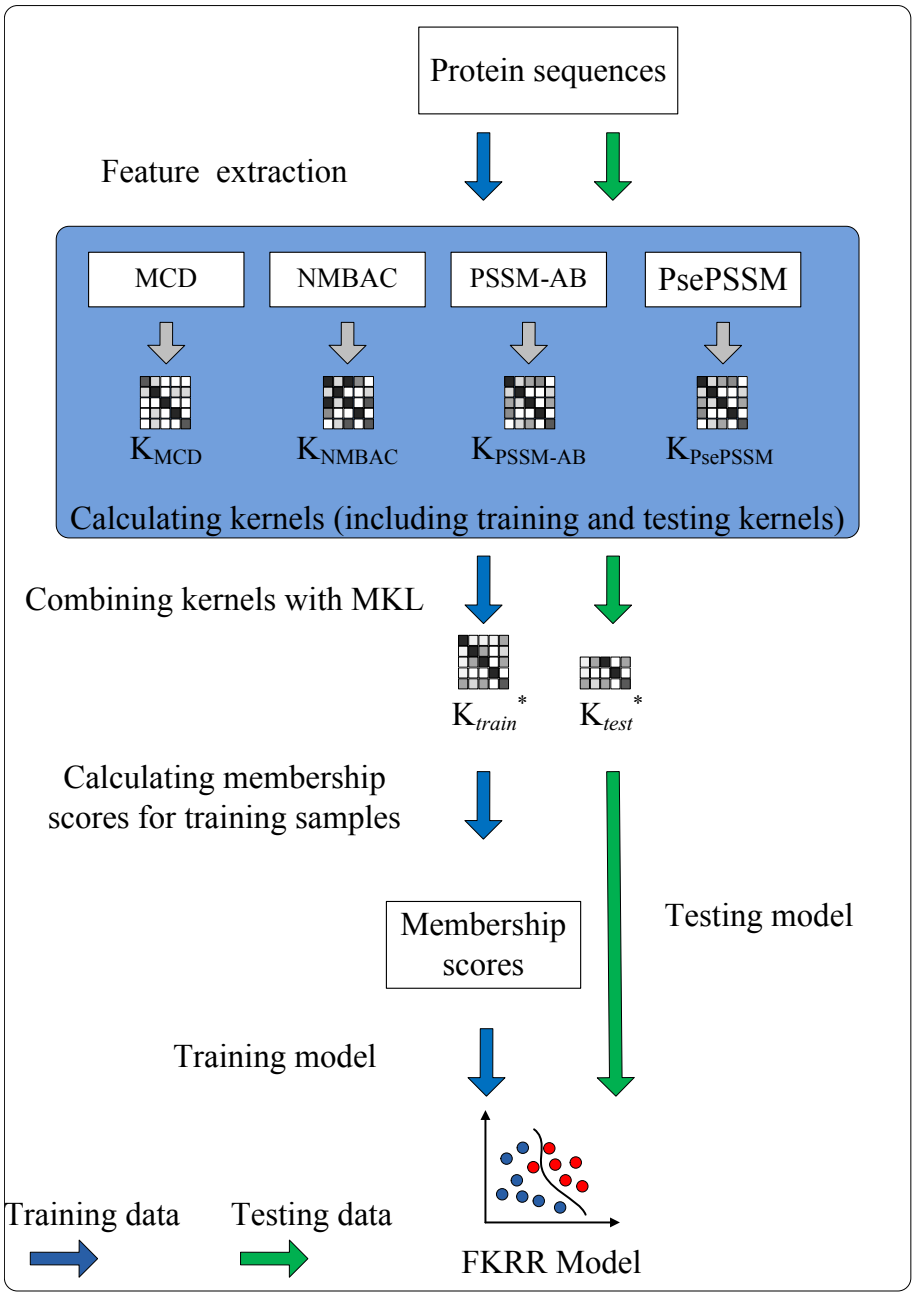

Figure 4. The process of FKRR-MVSF.

4.2. Multiple Kernel Learning

RBF is employed to construct 4 types of kernels via above features (including MCD, NMBAC, PSSM-AB and PsePSSM):

$$K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2),\ i, j = 1, 2, ..., N \tag{7}$$

where γ is the Gaussian kernel bandwidth. N is the number of samples. $\mathbf{x}_i$ and $\mathbf{x}_j$ are the feature vector of sample i and j. The 4 types of feature can be represented as a kernel set as: $\{\mathbf{K}_{MCD}, \mathbf{K}_{NMBAC}, \mathbf{K}_{PSSM-AB}, \mathbf{K}_{PsePSSM}\}$.

The MKL algorithm combines multi-view features from different sources. Some kernels may have bias in the learning process. MKL can reduce bias of kernels by low weights. The optimal kernel $\mathbf{K}^*_{train}$ is obtained as follows:

$$\mathbf{K}^*_{train} = \sum_{h=1}^{H} \omega_h \mathbf{K}_h, \quad \mathbf{K}^*, \mathbf{K}_h \in \Re^{N \times N} \tag{8}$$

where H denotes the number of basic kernels.

MKL algorithm [54] can estimate the optimal weights of kernels by minimize the distance between ideal kernel $\mathbf{K}_{ideal}$ and optimal kernel $\mathbf{K}^*_{train}$. The $\mathbf{K}_{ideal} = \mathbf{y}_{train}\mathbf{y}^T_{train} \in \Re^{N \times N}$ denote the information of label space. $\mathbf{y}_{train} \in \Re^{N \times 1}$ is the labels of training set. We hope that optimal kernel $\mathbf{K}^*_{train}$ is close to the $\mathbf{K}_{ideal}$ kernel:

$$\min_{\boldsymbol{\omega}, \mathbf{K}^*} \|\mathbf{K}^*_{train} - \mathbf{K}_{ideal}\|^2_F + \lambda \|\boldsymbol{\omega}\|^2_F \tag{9a}$$

$$subject\ to\ \mathbf{K}^*_{train} = \sum_{h=1}^{H} \omega_h \mathbf{K}_h, \tag{9b}$$

$$\omega_h \geq 0,\ h = 1, 2, ..., H, \tag{9c}$$

$$\sum_{h=1}^{H} \omega_h = 1 \tag{9d}$$

where $\|X\|^2_F = Trace(XX^T)$, λ is a regularization parameters, $\boldsymbol{\omega} = [\omega_1, \omega_2, ..., \omega_h]^T$ is the weights of kernels.

4.3. Fuzzy Kernel Ridge Regression

Kernel ridge regression is a method from statistics that implements a form of Regularized Least Squares (RLS). Given a training sample $\mathbf{x}_i, y_i$, $i = 1, 2, ..., N$. N, $\mathbf{x}_i$ and y_i is the number of samples, feature vector and label. The RLS aims to find the minimum of the following function:

$$J = \frac{C}{2} \|\mathbf{K}_{train}\boldsymbol{\alpha} - \mathbf{y}_{train}\|^2 + \frac{1}{2}\|f\|^2_K \tag{10}$$

where $\mathbf{K}_{train} \in \Re^{N \times N}$ is the training kernel, C is the non-negative regular term. The solution of KRR is:

$$\boldsymbol{\alpha} = (\mathbf{K}_{train} + \frac{1}{C}\mathbf{I})^{-1}\mathbf{y}_{train} \tag{11}$$

In this paper, we present a Fuzzy Kernel Ridge Regression (FKRR) for classification. We need to minimize the sum of errors ($\|\mathbf{K}_{train}\boldsymbol{\alpha} - \mathbf{y}_{train}\|^2$). The contribution of sample $\mathbf{x}_i$ to the decision boundary should be proportional to its fuzzy membership value. The objective function is following function:

$$J = \frac{C}{2} \|\mathbf{D}(\mathbf{K}_{train}\boldsymbol{\alpha} - \mathbf{y}_{train})\|^2 + \frac{1}{2}\|f\|^2_K \tag{12}$$

where $\mathbf{D} \in \Re^{N\times N}$ is a diagonal matrix whose element D_{ii} ($0 \leq D_{ii} \leq 1$) represents a fuzzy membership value for sample $\mathbf{x}_i$.

We set $\partial J/\partial \boldsymbol{\alpha} = 0$ and the solution of $\boldsymbol{\alpha}$ can be obtained as follows:

$$\partial(\frac{C}{2}\|\mathbf{D}(\mathbf{K}_{train}\boldsymbol{\alpha} - \mathbf{y}_{train})\|^2 + \frac{1}{2}\|f\|_K^2)/\partial\boldsymbol{\alpha} = 0 \tag{13a}$$

$$\partial(\frac{C}{2}\|\mathbf{D}(\mathbf{K}_{train}\boldsymbol{\alpha} - \mathbf{y}_{train})\|^2 + \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}_{train}\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = 0 \tag{13b}$$

$$C\mathbf{K}_{train}^T\mathbf{D}^T(\mathbf{D}\mathbf{K}_{train}\boldsymbol{\alpha} - \mathbf{D}\mathbf{y}_{train}) + \mathbf{K}_{train}\boldsymbol{\alpha} = 0 \tag{13c}$$

$$C\mathbf{D}^2(\mathbf{K}_{train}\boldsymbol{\alpha} - \mathbf{y}_{train}) + \boldsymbol{\alpha} = 0 \tag{13d}$$

$$\boldsymbol{\alpha} = (\mathbf{K}_{train} + \frac{1}{C}\mathbf{D}^{-2}\mathbf{I})^{-1}\mathbf{y}_{train} \tag{13e}$$

where $\mathbf{I} \in \Re^{N\times N}$. So, the decision function is following:

$$\mathbf{y}_{test} = sign[\mathbf{K}_{test}\boldsymbol{\alpha}] \tag{14a}$$

$$= sign[\mathbf{K}_{test}(\mathbf{K}_{train} + \frac{1}{C}\mathbf{D}^{-2}\mathbf{I})^{-1}\mathbf{y}_{train}] \tag{14b}$$

where $\mathbf{y}_{test} \in \Re^{M\times 1}$ is predictive labels. $\mathbf{K}_{test} \in \Re^{M\times N}$ denotes the kernel of testing samples, M is the number of testing samples.

To compute fuzzy membership values of train samples, we employ the optimal kernels $\mathbf{K}_{train}^*$ (training kernel) as following function:

$$score_t = \frac{1}{N^2}(\sum_{y_t=y_i} K_{train}^*(\mathbf{x}_t, \mathbf{x}_i) - \sum_{y_t\neq y_i} K_{train}^*(\mathbf{x}_t, \mathbf{x}_i)) \tag{15}$$

where $score_t$ denotes the score of training point t. If a sample t has a larger score. This sample may has a greater contribution to model. We normalize scores into fuzzy membership values (0–1), as follows:

$$D_{tt} = \frac{1}{1 + exp(-score_t)},\ t = 1, 2, ..., N \tag{16}$$

5. Conclusions

FKRR-MVSF achieves better results on independent datasets (MCC: 0.676). Eliminating noise points can improve the predictive performance of the model. In the future, we aim to use other fuzzy membership functions to build fuzzy models for filtering the noise points. As pointed out in PseAAC-based methods [13,33,39,40,55–60], we will establish a web-server for our model. The related code and datasets can be download from: https://figshare.com/s/e80f1a96b7b7bbf8062b.

Author Contributions: Y.Z., L.P. and Y.D. conceived the study. Y.Z. and Y.D. performed the experiments and analyzed the data. Y.Z., Y.D., J.T., F.G. and L.P. drafted the manuscript. All authors read and approved the manuscript.

Funding: This research was funded by State Key Research Project: Ferment Equipment Intelligent monitor and Early-warning Diagnosis System (grant number 2018YFD0400902), National Science Foundation of China (grant number 61873112) and Natural Science Research Project of Jiangsu Higher Eduction Institutions of China (grant number 19KJB520014).

Acknowledgments: The authors would like to thank all the guest editors and anonymous reviewers for their constructive advices.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chou, K.C.; Tomasselli, A.G.; Heinrikson, R.L. Prediction of the Tertiary Structure of a Caspase-9/Inhibitor Complex. *FEBS Lett.* **2000**, *470*, 249–256. [CrossRef]
2. Chou, K.C.; Jones, D.; Heinrikson, R.L. Prediction of the tertiary structure and substrate binding site of caspase-8. *FEBS Lett.* **1997**, *419*, 49–54. [CrossRef]
3. Chou, K.C. Insights from modelling the 3D structure of the extracellular domain of α7 nicotinic acetylcholine receptor. *Biochem. Biophys. Res. Commun.* **2004**, *319*, 433–438. [CrossRef] [PubMed]
4. Xie, H.L.; Fu, L.; Nie, X. 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*, 735–742. [CrossRef] [PubMed]
5. Xu, Y.; Ding, J.; Wu, L. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS ONE* **2013**, *8*, e55844. [CrossRef] [PubMed]
6. Chen, W.; Feng, P.; Ding, H.; Lin, H. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. *Anal. Biochem.* **2015**, *490*, 26–33. [CrossRef]
7. Chou, K.C. Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* **2015**, *11*, 218–234. [CrossRef]
8. Jia, J.; Liu, Z.; Xiao, X.; Liu, B. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.* **2016**, *394*, 223–230. [CrossRef]
9. Jia, J.; Liu, Z.; Xiao, X.; Liu, B. iCar-PseCp: Identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC. *Oncotarget* **2016**, *7*, 34558–34570. [CrossRef]
10. Liu, Z.; Xiao, X.; Qiu, W.R. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Anal. Biochem.* **2015**, *474*, 69–77. [CrossRef]
11. Xiao, X.; Min, J.L.; Lin, W.Z.; Liu, Z.; Cheng, X. iDrug-Target: Predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. *J. Biomol. Struct. Dyn.* **2015**, *33*, 2221–2233. [CrossRef] [PubMed]
12. Jia, J.; Liu, Z.; Xiao, X. iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. *J. Theor. Biol.* **2015**, *377*, 47–56. [CrossRef] [PubMed]
13. Chen, W.; Feng, P.M.; Lin, H. iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* **2013**, *41*, e68. [CrossRef] [PubMed]
14. Chou, K.C. Prediction of protein cellular attributes using pseudo amino acid composition. *PROTEINS Struct. Funct. Genet.* **2001**, *43*, 246–255. [CrossRef] [PubMed]
15. Chen, W.; Lei, T.; Jin, D.; Lin, H. PseKNC: A flexible web-server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.* **2014**, *456*, 53–60. [CrossRef]
16. Wei, L.; Luan, S.; Nagai, L.; Su, R.; Zou, Q. Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* **2019**, *35*, 1326–1333. [CrossRef]
17. Zou, Q.; Xing, P.; Wei, L.; Liu, B. Gene2vec: Gene Subsequence Embedding for Prediction of Mammalian N6-Methyladenosine Sites from mRNA. *RNA* **2019**, *25*, 205–218. [CrossRef]
18. Jia, C.; Zuo, Y.; Zou, Q. O-GlcNAcPRED-II: An integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinformatics* **2018**, *34*, 2029–2036. [CrossRef]
19. Zeng, X.; Liu, L.; Lu, L.; Zou, Q. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* **2018**, *34*, 2425–2432. [CrossRef]
20. Xuan, P.; Han, K.; Guo, Y.; Li, J.; Li, X.; Zhong, Y.; Zhang, Z.; Ding, J. Prediction of potential disease-associated microRNAs by using neural network. *Mol. Ther. -Nucleic Acids* **2019**, *16*, 566–575.
21. Liu, B.; Jiang, S.; Zou, Q. HITS-PR-HHblits: Protein remote homology detection by combining pagerank and hyperlink-induced topic search. *Brief. Bioinform.* **2019**. [CrossRef] [PubMed]
22. Wei, L.; Ding, Y.; Su, L.; Tang, J.; Zou, Q. Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.* **2018**, *117*, 212–217. [CrossRef]
23. Ru, X.; Li, L.; Zou, Q. Incorporating Distance-based Top-n-gram and Random Forest to Identify Electron Transport Proteins. *J. Proteome Res.* **2019**, *18*, 2931–2939. [CrossRef] [PubMed]
24. Qu, K.; Guo, F.; Liu, X.; Zou, Q. Application of Machine Learning in Microbiology. *Front. Microbiol.* **2019**, *10*, 827. [CrossRef] [PubMed]

25. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
26. Liu, B.; Xu, J.; Fan, S.; Xu, R.; Zhou, J.; Wang, X. PseDNA-Pro: DNA-Binding Protein Identification by Combining Chou's PseAAC and Physicochemical Distance Transformation. *Mol. Inform.* **2015**, *34*, 8–17. [CrossRef]
27. Kumar, M.; Gromiha, M.M.; Raghava, G.P. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinform.* **2007**, *8*, 463. [CrossRef]
28. Lipman, D.J.; Zhang, J.; Madden, T.L. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
29. Liu, B.; Xu, J.; Lan, X.; Xu, R.; Zhou, J.; Wang, X.; Chou, K.C. iDNA-Prot | dis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition. *PLoS ONE* **2014**, *9*, e106691. [CrossRef]
30. Wei, L.; Tang, J.; Zou, Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. *Inf. Sci.* **2016**, *384*, 135–144. [CrossRef]
31. Nimrod, G.; Schushan, M.; Szilágyi, A.; Leslie, C.; Ben-Tal, N. iDBPs: A web server for the identification of DNA binding proteins. *Bioinformatics* **2010**, *26*, 692–693. [CrossRef] [PubMed]
32. Hussain, W.; Khan, S.D.; Rasool, N.; Khan, S.A. SPalmitoylC-PseAAC: A sequence-based model developed via Chou's five-step rule and general PseAAC for identifying S-palmitoylation sites in proteins. *Anal. Biochem.* **2019**, *568*, 14–23. [CrossRef] [PubMed]
33. Chou, K.C. Progresses in predicting post-translational modification. *Int. J. Pept. Res. Ther.* **2019**. [CrossRef]
34. Awais, M.; Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A. iPhosH-PseAAC: Identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**. [CrossRef] [PubMed]
35. Ning, Q.; Ma, Z.; Zhao, X. dForml(KNN)-PseAAC: Detecting formylation sites from protein sequences using K-nearest neighbor algorithm via Chou's 5-step rule and pseudo components. *J. Theor. Biol.* **2019**, *470*, 43–49. [CrossRef] [PubMed]
36. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review, five-step rule). *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef] [PubMed]
37. Chou, K.C. Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Curr. Med. Chem.* **2019**. [CrossRef]
38. Chou, K.C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10–19. [CrossRef]
39. Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L. Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **2015**, *43*, W65–W71. [CrossRef]
40. Liu, B.; Wu, H. Pse-in-One 2.0: An improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nat. Sci.* **2017**, *9*, 67–91. [CrossRef]
41. Rose, P.W.; Prli, A.; Bi, C.; Bluhm, W.F.; Christie, C.H.; Dutta, S.; Green, R.K.; Goodsell, D.S.; Westbrook, J.D.; Woo, J.; et al. The RCSB Protein Data Bank: Views of structural biology for basic and applied research and education. *Nucleic Acids Res.* **2015**, *43*, 345–356. [CrossRef]
42. Lou, W.; Wang, X.; Chen, F.; Chen, Y.; Jiang, B.; Zhang, H. Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes. *PLoS ONE* **2014**, *9*, e86703. [CrossRef]
43. Lin, W.; Fang, J.; Xiao, X. iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. *PLoS ONE* **2011**, *6*, e24756. [CrossRef]
44. Kumar, K.K.; Pugalenthi, G.; Suganthan, P.N. DNA-Prot: Identification of DNA Binding Proteins from Protein Sequence Information using Random Forest. *J. Biomol. Struct. Dyn.* **2009**, *26*, 679–686. [CrossRef]
45. Liu, B.; Wang, S.; Wang, X. DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation. *Sci. Rep.* **2015**, *5* 15479. [CrossRef]
46. Adilina, S.; Farid, D.; Shatabda, S. Effective DNA binding protein prediction by using key features via Chou's general PseAAC. *J. Theor. Biol.* **2019**, *460*, 64–78. [CrossRef]
47. Xu, R.; Zhou, J.; Wang, H. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Syst. Biol.* **2014**, *9*, e86703. [CrossRef]

48. Liu, X.; Gong, X.; Yu, H.; Xu, J. A Model Stacking Framework for Identifying DNA Binding Proteins by Orchestrating Multi-View Features and Classifiers. *Genes* **2018**, *9*, 394. [CrossRef]
49. Feng, Z.P.; Zhang, C.T. Prediction of membrane protein types based on the hydrophobic index of amino acids. *J. Protein Chem.* **2000**, *19*, 269–275. [CrossRef]
50. Ding, Y.J.; Tang, J.J.; Guo, F. Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinform.* **2016**, *17*, 398–410. [CrossRef]
51. Jeong, J.C.; Lin, X.; Chen, X.W. On position-specific scoring matrix for protein function prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2011**, *8*, 308–315. [CrossRef]
52. You, Z.H.; Zhu, L.; Zheng, C.H.; Yu, H.J.; Deng, S.P.; Ji, Z. Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinform.* **2014**, *15*, S9. [CrossRef]
53. Chou, K.C.; Shen, H.B. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* **2007**, *360*, 339–345. [CrossRef]
54. He, J.; Chang, S.F.; Xie, L. Fast Kernel learning for Spatial Pyramid Matching. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–7.
55. Chou, K.C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr. Proteom.* **2009**, *6*, 262–274. [CrossRef]
56. Chen, W.; Lin, H. Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences. *Mol. Biosyst.* **2015**, *11*, 2620–2634. [CrossRef]
57. Liu, B.; Yang, F.; Huang, D.S. iPromoter-2L: A two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* **2018**, *34*, 33–40. [CrossRef]
58. Chen, W.; Ding, H.; Zhou, X.; Lin, H. iRNA(m6A)-PseDNC: Identifying N6-methyladenosine sites using pseudo dinucleotide composition. *Anal. Biochem.* **2018**, *561*, 59–65. [CrossRef]
59. Chen, W.; Feng, P.; Yang, H.; Ding, H.; Lin, H. iRNA-3typeA: Identifying 3-types of modification at RNA's adenosine sites. *Mol. Ther.-Nucleic Acid* **2018**, *11*, 468–474. [CrossRef]
60. Lin, H.; Deng, E.Z.; Ding, H.; Chen, W. iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* **2014**, *42*, 12961–12972. [CrossRef]