genes

Article

SXGBsite: Prediction of Protein–Ligand Binding Sites Using Sequence Information and Extreme Gradient Boosting

Ziqi Zhao, Yonghong Xu * and Yong Zhao

School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China; Lightness7@outlook.com (Z.Z.); zhaoyong@ysu.edu.cn (Y.Z.)

* Correspondence: xyh@ysu.edu.cn; Tel.: +86-131-0254-7218

Received: 5 September 2019; Accepted: 19 November 2019; Published: 22 November 2019

Abstract: The prediction of protein–ligand binding sites is important in drug discovery and drug design. Protein–ligand binding site prediction computational methods are inexpensive and fast compared with experimental methods. This paper proposes a new computational method, SXGBsite, which includes the synthetic minority over-sampling technique (SMOTE) and the Extreme Gradient Boosting (XGBoost). SXGBsite uses the position-specific scoring matrix discrete cosine transform (PSSM-DCT) and predicted solvent accessibility (PSA) to extract features containing sequence information. A new balanced dataset was generated by SMOTE to improve classifier performance, and a prediction model was constructed using XGBoost. The parallel computing and regularization techniques enabled high-quality and fast predictions and mitigated overfitting caused by SMOTE. An evaluation using 12 different types of ligand binding site independent test sets showed that SXGBsite performs similarly to the existing methods on eight of the independent test sets with a faster computation time. SXGBsite may be applied as a complement to biological experiments.

Keywords: protein–ligand binding site; SMOTE; Extreme Gradient Boosting; discrete cosine transform (DCT); discrete wavelet transform (DWT)

1. Introduction

Accurate prediction of protein–ligand binding sites is important for understanding protein function and drug design [1–4]. The experiment-based three-dimensional (3D) structure recognition of protein–ligand complexes and binding sites is relatively expensive and time consuming [5,6]. Computational methods can predict binding sites rapidly and can be applied as a supplement to experimental methods. Structure-based methods, sequence-based methods, and hybrid methods are the commonly applied computation methods [7,8].

The structure-based methods are usually applied to predict ligand binding sites with known 3D protein structures [2,9–11]. We focused on the sequence-based method without 3D structure information, and only a few structure-based methods are listed due to the rapid update of these different methods. Pockets on the protein surface can be identified by computing geometric measures, such as LIGSITE^{CSC} [2,12], CASTp [13–16], LigDig [17], and Fpocket [18,19]. LIGSITE^{CSC} [2,12] identifies pockets through the number of surface–solvent–surface events and clusters. CASTp [13–16] locates and measures pockets on 3D protein structures and annotates functional information for specific residues. Unlike traditional protein-centric approaches, LigDig [17] is a ligand-centric approach that identifies pockets using information from PDB [20], UniProt [21], PubChem [22], ChEBI [23], and KEGG [24]. Fpocket [18,19] identifies pockets using structure-based virtual screening (SBVS). RF-Score-VS [25] improves the performance of SBVS and can be used in the open source ODDT [26,27]. FunFOLD [1] introduces cluster identification and residue selection to automatically predict ligand binding residues. CHED [28] constructs a model to predict metal-binding sites using geometric information and machine learning methods. The integration of sequence information in structure-based methods helps improve prediction performance [29–31]. ConCavity [29] integrates sequence evolution information and structure information to recognize pockets. COACH [30] and HemeBIND [31] construct prediction models and identify ligand binding sites using sequence and structural information features based on machine learning methods. In general, structure-based methods and hybrid methods enable high-quality predictions when 3D structures of protein–ligand complexes are known [8].

Sequence-based methods can predict protein–ligand binding sites with unknown 3D structures [5,32–34]. MetaDBSite [32] integrates six methods, including DISIS [35], DNABindR [36], BindN [37], BindN-rf [38], DP-Bind [39], and DBS-PRED [40], and produces better results than each of the methods alone. DNABR [5] introduces sequence characteristics based on the random forest method [41] to study the sequence characteristics that delineate the physicochemical properties of amino acids. Both SVMPred [33] and NsitePred [34] construct support vector machine (SVM) [42] prediction models using multiple features including position-specific scoring matrix (PSSM), predicted solvent accessibility (PSA), predicted secondary structure (PSS), and predicted dihedral angles. TargetS [7] considers the ligand-specific binding propensity feature and builds models using a scheme of under-sampling and ensemble SVMs. EC-RUS [8] selects position-specific scoring matrix discrete cosine transform (PSSM-DCT) and PSA as features, constructs prediction models using under-sampling and ensemble classifiers, and compares the prediction quality of weighted sparse representation based classifier (WSRC) [43] and SVM.

One machine learning model in the ensemble classifiers is usually built with a dataset generated by under-sampling, and a new model is built after the end of the building process of the previous model. In this paper, this process is called the serial method, and performs well among the sequence-based methods at present but requires more computation time [8,44]. Here, we propose a new parallel method for predicting protein–ligand binding site residues using the evolutionary conservation information of homologous proteins. The main information source used for predictions is the PSSM of sequences. The prediction model of binding residues is constructed by XGBoost machine learning method [45] with the synthetic minority over-sampling technique (SMOTE) [46], and this method reduces the computation time while ensuring prediction quality. We compared the prediction qualities of different feature combination schemes of PSSM-DCT [8,47–49], PSSM-discrete wavelet transform (DWT) [49–51] and PSA [52], and PSSM-DCT + PSA scheme was selected. For the dataset imbalance problem, XGBoost with SMOTE was applied to construct the protein–ligand binding site prediction models, and the optimal parameters were determined by five-fold cross-validation and a grid search method. The models were validated on 12 different types of protein–ligand binding site datasets. The SXGBsite process is shown in Figure 1.

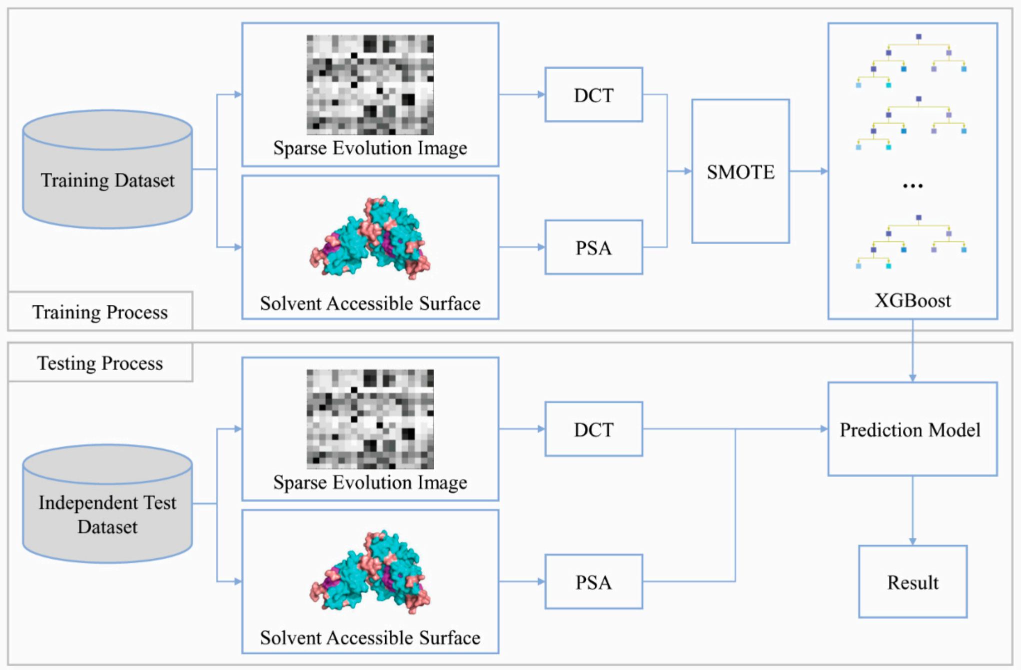

Figure 1. SXGBsite Flowchart. During the training process, the position-specific scoring matrix (PSSM) feature of residues was represented by the sparse evolution image, discrete cosine transform (DCT) compressed the PSSM feature to obtain the PSSM-DCT feature, and the predicted solvent accessibility (PSA) feature was used to improve the prediction quality. SMOTE generated a new balanced training set with the training set of PSSM-DCT + PSA features, and the prediction model of binding residues was constructed by the balanced training set and XGBoost. During the testing process, the unbalanced independent test set, which also extracted the PSSM-DCT + PSA features, was input into the prediction model to obtain the result.

2. Materials and Methods

2.1. Benchmark Datasets

The benchmark datasets were constructed based on the BioLip database [53] developed by Yu et al. [7], including the training and independent test datasets of 12 different ligands. The 12 types of ligands used were five types of nucleotides, five types of metal ions, DNA and Heme (Table 1). The source code and datasets are available at https://github.com/Lightness7/SXGBsite.

Table 1. Composition of datasets for the 12 different ligands [7].

Ligand Category	Ligand Type	Training Dataset		Independent Test Dataset		Total No. Sequences
		No. Sequences	(numP,numN)	No. Sequences	(numP,numN)	
Nucleotide	ATP	221	(3021,72334)	50	(647,16639)	271
	ADP	296	(3833,98740)	47	(686,20327)	343
	AMP	145	(1603,44401)	33	(392,10355)	178
	GDP	82	(1101,26244)	14	(194,4180)	96
	GTP	54	(745,21205)	7	(89,1868)	61
Metal Ion	Ca^{2+}	965	(4914,287801)	165	(785,53779)	1130
	Zn^{2+}	1168	(4705,315235)	176	(744,47851)	1344
	Mg^{2+}	1138	(3860,350716)	217	(852,72002)	1355
	Mn^{2+}	335	(1496,112312)	58	(237,17484)	393
	Fe^{3+}	173	(818,50453)	26	(120,9092)	199
	DNA	335	(6461,71320)	52	(973,16225)	387
	HEME	206	(4380,49768)	27	(580,8630)	233

Note: numP, positive (binding residues) sample numbers; numN, negative (non-binding residues) sample numbers; ATP, adenosine triphosphate; ADP, adenosine diphosphate; AMP, adenosine monophosphate; GDP, guanosine diphosphate; GTP, guanosine triphosphate.

2.2. Feature Extraction

2.2.1. Position-Specific Scoring Matrix

The position-specific scoring matrix (PSSM) encodes the evolution information of the protein sequence. The PSSM of each sequence was obtained using PSI-BLAST [54] in the database of non-redundant protein sequences (nr) with three iterations and the E-value of 0.001. The PSSM is a matrix of $L \times 20$, where L rows represent L amino acid residues in the protein sequence, 20 columns represent the probability that each residue is mutated to 20 native residues, as follows:

$$\mathrm{PSSM} = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,20} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{L,1} & P_{L,2} & \cdots & P_{L,20} \end{bmatrix} \tag{1}$$

The PSSM feature of contiguous residues was extracted with a sliding window with size w. The window was centered on the target residue and contained $(w-1)/2$ adjacent residues on both sides. The size of the PSSM feature matrix was $w \times 20$, and the residue sparse evolution image [8,48] is shown in Figure 2. The window size $w = 17$ was selected after testing different values of w, and the dimensions of the PSSM feature were $17 \times 20 = 340$.

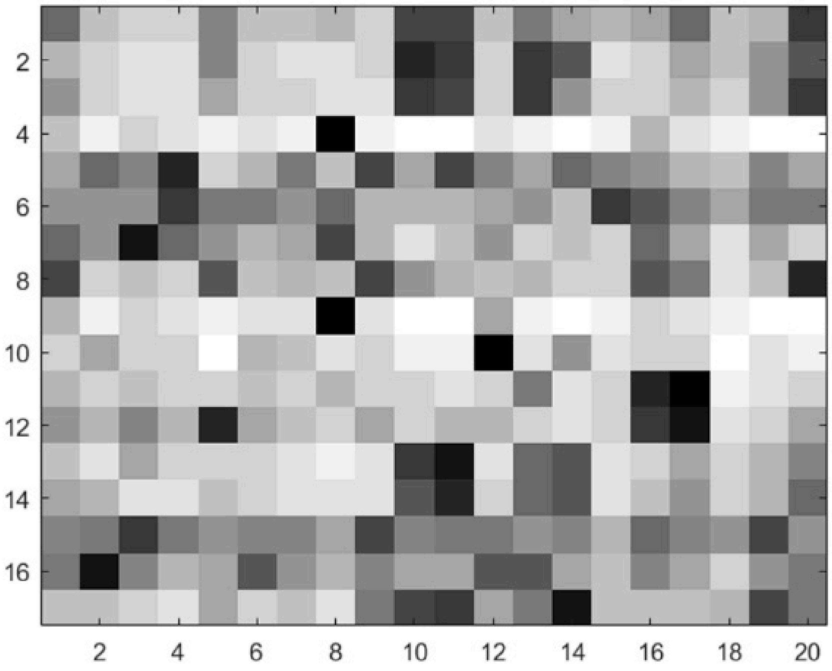

Figure 2. Residue sparse evolution image.

2.2.2. Discrete Cosine Transform

Discrete Cosine Transform (DCT) [47] is widely applied for lossy data compression of signals and images. In this study, we used DCT to concentrate the information of PSSM into a few coefficients. For a given input matrix $Mat \in \Re^{m \times n}$, DCT is defined as:

$$DCT(i,j) = a_i a_j \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} Mat(m,n) \cos\frac{\pi(2m+1)i}{2M} \times \cos\frac{\pi(2n+1)j}{2N}, \quad 0 \le i \le M, \quad 0 \le j \le N, \tag{2}$$

where

$$a_i = \begin{cases} \frac{1}{\sqrt{M}}, & i = 0 \\ \sqrt{\frac{2}{M}}, & 1 \le i \le M-1 \end{cases}$$
$$a_j = \begin{cases} \frac{1}{\sqrt{N}}, & j = 0 \\ \sqrt{\frac{2}{N}}, & 1 \le j \le N-1 \end{cases} \tag{3}$$

The compressed PSSM feature of the residue was obtained by using DCT on the PSSM feature matrix. Most of the information after PSSM-DCT was concentrated in the low-frequency part of the compressed PSSM. The first r rows of the compressed PSSM were reserved as the PSSM-DCT feature, and the dimensions of the PSSM-DCT feature were $r \times 20$.

2.2.3. Discrete Wavelet Transform

Discrete Wavelet Transform (DWT) [49] can decompose discrete sequences into high- and low-frequency coefficients. Four-level DWT [50] was applied to acquire the first five discrete cosine coefficients, the standard deviation, mean, and maximum and minimum values of different scales, as shown in Figure 3. The PSSM-DWT feature of the residue was obtained from the PSSM feature via four-level DWT, and the PSSM-DWT feature had 1040 dimensions.

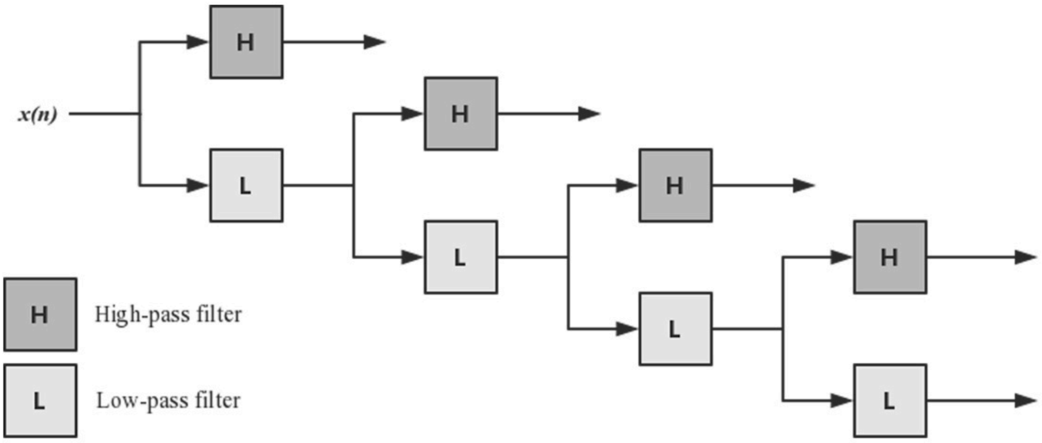

Figure 3. Four-level discrete wavelet transform (DWT).

2.2.4. Predicting Solvent Accessibility

Solvent accessibility [52] is related to the spatial arrangement and packing of residues during the protein folding process, which is an effective feature for protein–ligand binding site prediction [8,33,34]. We used the solvent accessibility prediction of proteins by nearest neighbor method (Sann) to obtain the PSA feature of residues [55], and the PSA feature had three dimensions.

2.3. *SMOTE Over-Sampling*

As a common method for tackling unbalanced data, SMOTE over-samples the minority class by synthesizing new samples, under-samples the majority class, and provides better classifier performance within the receiver operating characteristic (ROC) space [45,56]. A balanced sample set is generated from the unbalanced sample set through feature extraction by SMOTE. After a series of tests, a new sample set with better results was constructed with the same positive and negative sample number: 19,000.

2.4. *Extreme Gradient Boosting Algorithm*

Extreme Gradient Boosting (XGBoost) algorithm [46] is an improvement on the Gradient Boosting algorithm [57] by Chen et al. and is characterized by fast calculation and high prediction accuracy.

XGBoost is widely used by data scientists in multiple applications and has provided advanced results [58,59]. The training set after feature extraction and SMOTE x_i $(x_i = \{x_1, x_2, \ldots, x_m\}, i = 1, 2, \ldots, n\ )$ was input into the K additive functions of XGBoost to build the model. The prediction result of the independent test set y_i $(y_i = \{0, 1\}, i = 1, 2, \ldots, s$, where 0 represents non-binding residues and 1 represents binding residues) was output as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{4}$$

where f_k is each independent tree function with leaf weights and F is the tree ensemble containing each function of the tree. XGBoost avoids large models with the following regularized objective formula:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{5}$$

where l is a differentiable convex loss function that measures the closeness of the prediction $\hat{y}_i$ and the target y_i, and Ω is a regular term that penalizes model complexity by greedily adding f_t to improve the tree ensemble model. The regular term avoids overfitting by penalizing leaf weights, and the Ω penalty function is as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{6}$$

where T is the number of leaves, ω is the leaf weights, and the regularization coefficients γ and λ are constants. The traditional GBDT only uses the first-order information of the loss function, whereas the second-order Taylor expansion was introduced into the loss function of XGBoost to optimize the function rapidly [60]. The simplified objective function of step t is:

$$\widetilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{7}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ represent the first- and second-order gradient statistics of the loss function, respectively. $I_j = \{i \big| q(x_i = j)\}$ is defined as a sample set of leaf j, simplified Equation (7) is:

$$\begin{aligned} \widetilde{\mathcal{L}}^{(t)} &= \sum_{i=1}^{n} [g_i f_t(x_i) + \tfrac{1}{2} h_i f_t^2(x_i)] + \gamma T + \tfrac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \\ &= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i)\omega_j + \tfrac{1}{2}(\sum_{i \in I_j} h_i + \lambda)\omega_j^2] + \gamma T \end{aligned} \tag{8}$$

The optimal weight ω_j^* of leaf j and the corresponding objective function value are calculated by:

$$\omega_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{9}$$

$$\widetilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{10}$$

The above equation provides the best split of the node. Supposing I_L and I_R are the left and right split nodes of the sample set I of the leaf, $I = I_L \cup I_R$, respectively, the loss reduction after splitting is expressed as:

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma \tag{11}$$

To prevent overfitting, XGBoost uses shrinkage and column (feature) subsampling techniques, as well as the regularized objective [57].

3. Results

The performance of classification was evaluated on the specificity (SP), sensitivity (SN), accuracy (ACC), and Matthews correlation coefficient (MCC). The overall prediction quality of a binary model was evaluated using the area under the receiver operating characteristic curve (AUC). The formulas used to determine SN, SP, ACC, and MCC are, respectively, as follows:

$$\mathrm{SP} = \frac{TN}{TN + FP} \tag{12}$$

$$\mathrm{SN} = \frac{TP}{TP + FN} \tag{13}$$

$$\mathrm{ACC} = \frac{TP + TN}{TP + FP + TN + FN} \tag{14}$$

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}} \tag{15}$$

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

3.1. Parameter Selection

ACC is insufficient for performance evaluation in unbalanced learning [7,8], MCC is suitable for quality assessment in sequence-based predictions [3], and AUC is usually used to assess the overall prediction quality of models. The value of MCC changes with the threshold, whereas the AUC value is not affected by the threshold value. We evaluated the prediction performance using MCC and AUC, and the threshold of the probability value was selected by maximizing the value of MCC. The value of MCC was used to select the first *r* rows of the PSSM-DCT matrix as feature on the guanosine triphosphate (GTP) training and independent test sets. PSSM-DCT obtained the optimal value of MCC when *r* was 9, as shown in Figure 4, and the dimensions of the PSSM-DCT feature were 9 × 20 = 180.

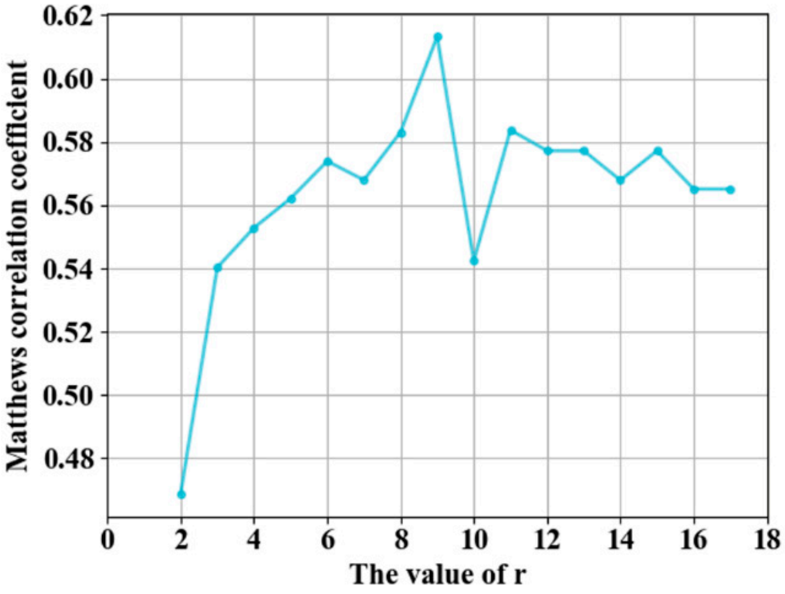

Figure 4. Adjustment of parameters *r*.

The size of the positive and negative sample sets after SMOTE is usually an integer multiple of the positive sample size in the original dataset, and the prediction quality may be affected by the amplification ratio of the positive sample sets. In this study, a fixed-size positive and negative sample set was generated by SMOTE to improve the prediction quality, and the optimal sample number was selected according to the value of MCC on the GTP training and independent test sets. The best value of MCC was obtained when the number of positive and negative samples was 19,000, as shown in Figure 5.

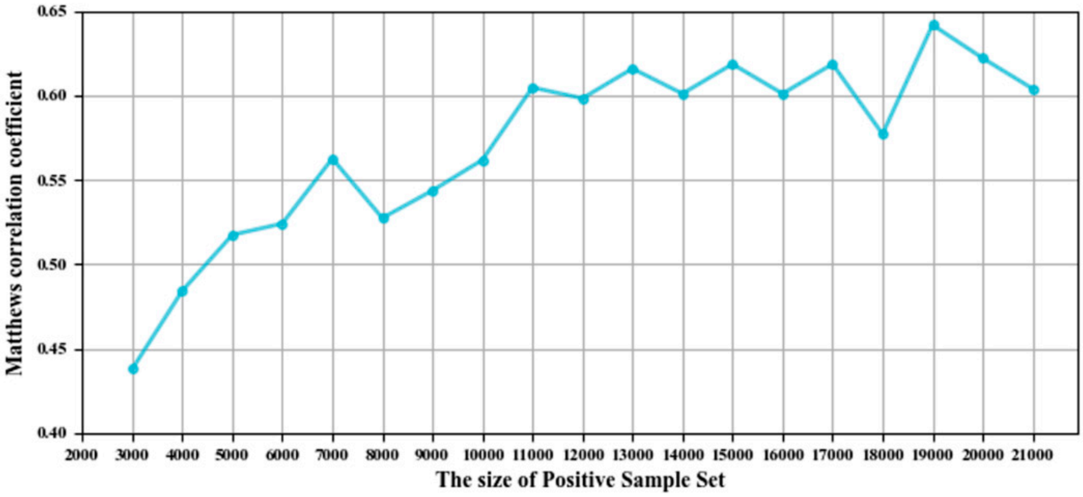

Figure 5. The values of the Matthews correlation coefficient (MCC) corresponding to the number of samples after SMOTE.

The parameters of XGBoost were adjusted with five-fold cross-validation and a grid search method on the GTP training set.

3.2. Method Selection

Different feature combinations of PSSM, PSSM-DCT, PSSM-DWT, and PSA were used to evaluate the prediction performance using the GTP training and independent test sets, PSSM-DCT + PSA produced the optimal MCC and AUC values (Table 2), and receiver operating characteristic curve (ROC) of different feature combinations is shown in Figure 6. As shown in Table 2, PSSM performed better in terms of AUC than PSSM-DCT and PSSM-DWT, and PSA (3-D) improved PSSM (340-D), PSSM-DCT (180-D), and PSSM-DWT (1040-D) by 0.14, 0.22 and 0.09, respectively. The relationship between the increase in AUC and the feature dimensions indicated that the prediction quality using PSA improved more for features with fewer dimensions (PSSM and PSSM-DCT). PSSM + PSA and PSSM-DCT + PSA performed almost the same in terms of AUC, and we tended to improve prediction quality by over-sampling in the comparison of feature combinations. The prediction qualities of PSSM and PSSM + PSA were more dependent on threshold moving, and the difference in MCC between the default threshold (0.500) and the maximum MCC threshold demonstrated the effect of threshold moving.

Table 2. Comparison of different feature combinations on the GTP independent test set (average of 10 replicate experiments in SXGBsite with adjusted parameters).

Feature	Threshold	SN (%)	SP (%)	ACC (%)	MCC	AUC
PSSM	0.500	34.8	99.7	96.8	0.536	0.855
	0.139	46.1	99.5	97.0	0.596	0.855
PSSM-DCT	0.500	43.8	99.7	97.1	0.605	0.848
	0.612	42.7	99.8	97.2	0.611	0.848
PSSM-DWT	0.500	41.6	99.7	97.0	0.586	0.830
	0.458	43.8	99.7	97.1	0.605	0.830
PSSM + PSA	0.500	37.1	99.9	97.0	0.581	0.869
	0.109	52.8	99.4	97.2	0.636	0.869
PSSM-DCT + PSA	0.500	49.4	99.6	97.3	0.642	0.870
	0.421	50.6	99.6	97.4	0.650	0.870
PSSM-DWT + PSA	0.500	46.1	99.7	97.3	0.630	0.839
	0.370	49.4	99.6	97.3	0.642	0.839
PSSM-DCT + PSSM-DWT + PSA	0.500	44.9	99.6	97.1	0.607	0.850
	0.545	44.9	99.8	97.3	0.629	0.850

Note: ACC, accuracy; MCC, Matthews correlation coefficient; AUC, the area under the receiver operating characteristic curve.

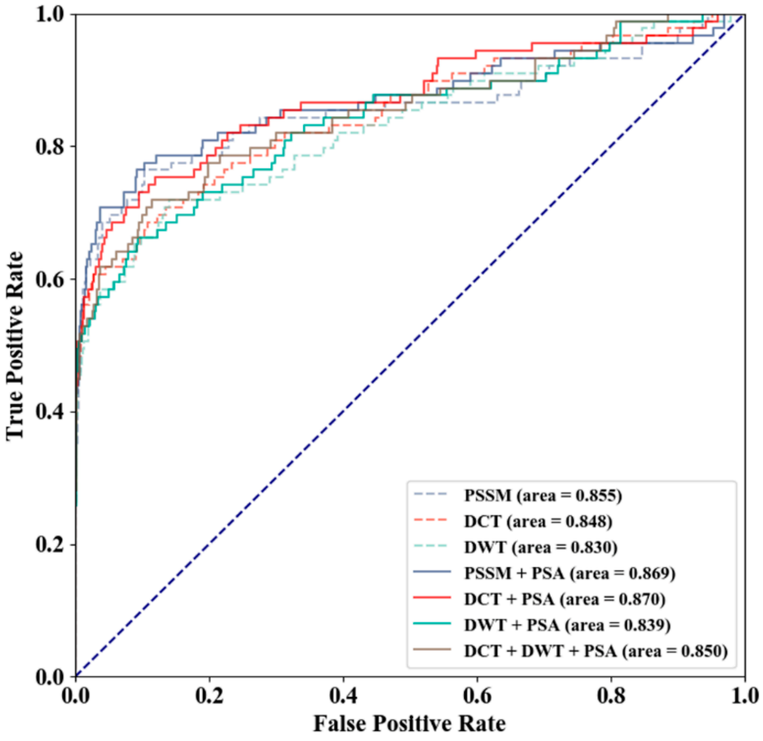

Figure 6. Receiver Operating Characteristic Curve (ROC) of Different Feature Combinations.

Three sampling schemes were used on the GTP training set to obtain three different training sets, including the entire GTP training set, the training set after random under-sampling (RUS), and the training set after SMOTE over-sampling. On the GTP independent test set, the prediction qualities of the models constructed by the three training sets are shown in Table 3, and receiver operating characteristic curve (ROC) of different sampling and classification algorithms is shown in Figure 7. SMOTE + XGBoost achieved the best prediction quality, performing better than SMOTE + SVM.

Table 3. Comparison of different sampling and classification algorithms on the GTP independent test set (average of 10 replicate experiments in XGBoost with adjusted parameters).

Scheme	Threshold	SN (%)	SP (%)	ACC (%)	MCC	AUC
XGBoost	0.500	30.3	99.8	96.7	0.512	0.842
	0.153	37.1	99.7	96.9	0.556	0.842
RUS + XGBoost	0.500	68.5	84.5	83.8	0.288	0.827
	0.914	51.7	97.9	95.8	0.504	0.827
SMOTE + XGBoost	0.500	49.4	99.6	97.3	0.642	0.870
	0.421	50.6	99.6	97.4	0.650	0.870
SMOTE + SVM	0.500	51.7	99.3	97.1	0.616	0.838
	0.714	49.4	99.5	97.2	0.628	0.838

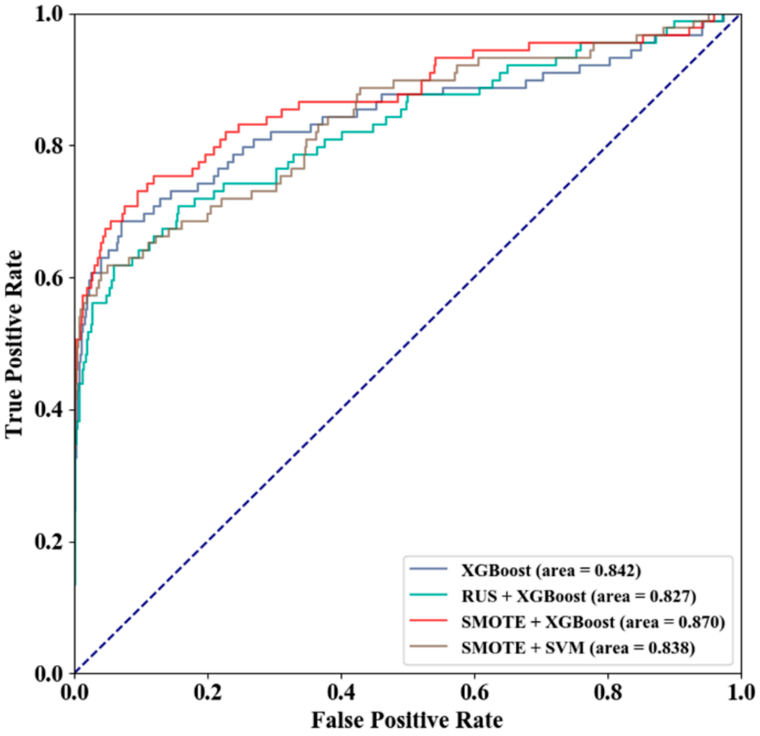

Figure 7. ROC of Different Sampling and Classification Algorithms.

3.3. Results of Training Sets

The performance of SXGBsite was evaluated using five-fold cross-validation on the training sets. The results with the threshold of 0.5 and the maximized the MCC value are listed in Table 4. The five-fold cross-validation results are basically consistent with the maximized MCC threshold results of TargetS and EC-RUS. Regardless of the impact of the threshold, the results in Table 4 show the different characteristics of the two schemes for the class imbalance problem by comparing the default threshold (0.500) results of SXGBsite and EC-RUS, which use the same features. The RUS + ensemble classifiers scheme was more sensitive to positive samples and had information loss for negative samples. The SMOTE + XGBoost scheme reduced the information loss, the positive samples in the training set were mostly synthesized, and the sensitivity to positive samples was lower.

Table 4. Performance of SXGBsite (average of 10 replicate experiments) on the training sets after five-fold cross-validation.

Ligand	Predictor	Threshold	SN (%)	SP (%)	ACC (%)	MCC	AUC
ATP	TargetS [1]	0.500	48.4	98.2	96.2	0.492	0.887
	EC-RUS [2]	0.500	84.1	84.9	84.9	0.347	0.912
		0.814	58.6	97.9	96.4	0.537	0.912
	SXGBsite	0.500	53.4	96.3	94.6	0.413	0.886
		0.775	40.3	98.6	96.4	0.448	0.886
ADP	TargetS [1]	0.500	56.1	98.8	97.2	0.591	0.907
	EC-RUS [2]	0.500	87.8	87.7	87.7	0.395	0.939
		0.852	62.2	98.6	97.3	0.610	0.939
	SXGBsite	0.500	61.6	96.2	94.9	0.459	0.907
		0.832	46.4	98.9	97.0	0.521	0.907
AMP	TargetS [1]	0.500	38.0	98.2	96.0	0.386	0.856
	EC-RUS [2]	0.500	81.5	79.7	79.8	0.263	0.888
		0.835	46.7	98.3	96.6	0.460	0.888
	SXGBsite	0.500	37.0	97.8	95.8	0.347	0.851
		0.636	32.3	98.6	96.4	0.366	0.851
GDP	TargetS [1]	0.430	63.9	98.7	97.2	0.644	0.908
	EC-RUS [2]	0.500	86.1	89.8	89.7	0.435	0.937
		0.816	67.2	98.9	97.6	0.676	0.937
	SXGBsite	0.500	59.4	99.3	97.7	0.664	0.930
		0.653	57.0	99.5	97.9	0.678	0.930
GTP	TargetS [1]	0.500	48.0	98.7	96.9	0.506	0.858
	EC-RUS [2]	0.500	79.5	85.7	85.5	0.309	0.896
		0.842	49.5	99.2	97.6	0.562	0.896
	SXGBsite	0.500	42.4	99.4	97.6	0.540	0.883
		0.685	40.7	99.7	97.8	0.572	0.883
Ca^{2+}	TargetS [1]	0.690	19.2	99.7	98.4	0.320	0.784
	EC-RUS [2]	0.500	73.9	73.8	73.8	0.118	0.812
		0.861	14.7	99.7	98.6	0.220	0.812
	SXGBsite	0.500	32.8	95.0	94.2	0.135	0.757
		0.818	16.3	99.1	98.1	0.167	0.757
Mg^{2+}	TargetS [1]	0.810	26.4	99.8	99.0	0.383	0.798
	EC-RUS [2]	0.500	73.8	79.4	79.3	0.125	0.839
		0.864	25.8	99.8	99.1	0.354	0.839
	SXGBsite	0.500	46.1	95.9	95.5	0.196	0.819
		0.926	26.3	99.7	99.0	0.326	0.819
Mn^{2+}	TargetS [1]	0.740	40.8	99.5	98.7	0.445	0.901
	EC-RUS [2]	0.500	83.4	86.6	86.6	0.201	0.921
		0.841	31.0	99.6	98.9	0.358	0.921
	SXGBsite	0.500	45.0	98.3	97.7	0.297	0.888
		0.759	36.1	99.1	98.5	0.329	0.888
Fe^{3+}	TargetS [1]	0.810	51.8	99.6	98.8	0.592	0.922
	EC-RUS [2]	0.500	87.1	90.1	90.0	0.278	0.940
		0.809	53.1	99.2	98.6	0.489	0.940
	SXGBsite	0.500	48.2	99.1	98.5	0.440	0.913
		0.496	50.1	99.1	98.5	0.454	0.913
Zn^{2+}	TargetS [1]	0.830	50.0	99.6	98.9	0.557	0.938
	EC-RUS [2]	0.500	88.7	90.8	90.8	0.279	0.958
		0.860	45.6	99.3	98.7	0.440	0.958
	SXGBsite	0.500	59.7	96.5	96.1	0.299	0.892
		0.894	38.5	99.2	98.5	0.363	0.892
DNA	TargetS [1]	0.490	41.7	94.5	89.9	0.362	0.824
	EC-RUS [2]	0.500	81.9	71.8	72.3	0.259	0.852
		0.763	48.7	95.1	92.6	0.378	0.852
	SXGBsite	0.500	41.0	92.3	89.6	0.255	0.827
		0.420	49.8	89.2	87.2	0.270	0.827
HEME	TargetS [1]	0.650	50.5	98.3	94.4	0.579	0.887
	EC-RUS [2]	0.500	85.0	83.6	83.7	0.416	0.922
		0.846	60.3	97.5	95.1	0.591	0.922
	SXGBsite	0.500	59.3	96.2	93.8	0.520	0.900
		0.805	45.3	98.9	95.4	0.555	0.900

[1] Results excerpted from Yu et al. [7]. [2] Results excerpted from Ding et al. [8].

3.4. Comparison with Existing Methods

In terms of the independent test sets of the five nucleotides, SXGBsite is compared with TargetS, SVMPred, NsitePred, EC-RUS, and the alignment-based baseline predictor in Table 5. The results of TargetS, SVMPred, NsitePred, and EC-RUS are the threshold of maximizing the MCC value. In terms of the ATP, ADP, AMP, GDP, and GTP independent test sets, the metrics of the best prediction quality refer to the AUC of TargetS and the MCC of EC-RUS. The differences between SXGBsite and TargetS for the AUC are 0.018 (0.880 to 0.898), 0.011 (0.885 to 0.896), 0.007 (0.823 to 0.830), 0.002 (0.894 to 0.896), and 0.015 (0.870 over 0.855), respectively, and the differences between SXGBsite and EC-RUS for the MCC are 0.043 (0.463 to 0.506), 0.023 (0.488 to 0.511), 0.065 (0.328 to 0.393), 0.003 (0.576 to 0.579), and 0.009 (0.650 over 0.641), respectively. The difference between SXGBsite and the best prediction quality is small for the AUC and relatively large for the MCC.

Table 5. SXGBsite (average of 10 replicate experiments) compared with the existing methods on five nucleotide independent test sets.

Ligand	Predictor	SN (%)	SP (%)	ACC (%)	MCC	AUC
ATP	TargetS [1]	50.1	98.3	96.5	0.502	0.898
	NsitePred [1]	50.8	97.3	95.5	0.439	-
	SVMPred [1]	47.3	96.7	94.9	0.387	0.877
	alignment-based [1]	30.6	97.0	94.5	0.265	-
	EC-RUS [2]	45.4	98.8	96.8	0.506	0.871
	SXGBsite (T = 0.500)	54.6	95.7	94.2	0.397	0.880
	SXGBsite (T = 0.718)	43.7	98.5	96.5	0.463	0.880
ADP	TargetS [1]	46.9	98.9	97.2	0.507	0.896
	NsitePred [1]	46.2	97.6	96.0	0.419	-
	SVMPred [1]	46.1	97.2	95.5	0.382	0.875
	alignment-based [1]	31.8	97.4	95.1	0.284	-
	EC-RUS [2]	44.4	99.2	97.6	0.511	0.872
	SXGBsite (T = 0.500)	53.1	96.9	95.6	0.399	0.885
	SXGBsite (T = 0.844)	37.3	99.5	97.7	0.488	0.885
AMP	TargetS [1]	34.2	98.2	95.9	0.359	0.830
	NsitePred [1]	33.9	97.6	95.3	0.321	-
	SVMPred [1]	32.1	96.4	94.1	0.255	0.798
	alignment-based [1]	19.6	97.3	94.5	0.178	-
	EC-RUS [2]	24.9	99.5	97.0	0.393	0.815
	SXGBsite (T = 0.500)	36.0	97.5	95.4	0.325	0.823
	SXGBsite (T = 0.486)	37.1	97.4	95.3	0.328	0.823
GDP	TargetS [1]	56.2	98.1	96.2	0.550	0.896
	NsitePred [1]	55.7	97.9	96.1	0.536	-
	SVMPred [1]	49.5	97.6	95.4	0.466	0.870
	alignment-based [1]	41.2	97.8	95.3	0.415	-
	EC-RUS [2]	36.6	99.9	97.1	0.579	0.872
	SXGBsite (T = 0.500)	46.4	99.0	96.7	0.551	0.894
	SXGBsite (T = 0.687)	40.2	99.7	97.1	0.576	0.894
GTP	TargetS [1]	57.3	98.8	96.9	0.617	0.855
	NsitePred [1]	58.4	95.7	94.0	0.448	-
	SVMPred [1]	48.3	91.7	89.7	0.276	0.821
	alignment-based [1]	52.8	97.9	95.9	0.516	-
	EC-RUS [2]	61.8	98.7	97.0	0.641	0.861
	SXGBsite (T = 0.500)	49.4	99.6	97.3	0.642	0.870
	SXGBsite (T = 0.421)	50.6	99.6	97.4	0.650	0.870

[1] Results excerpted from Yu et al. [7]. [2] Results excerpted from Ding et al. [8]. - denotes unavailable.

On the independent test sets of the five metal ions, SXGBsite is compared with TargetS, FunFOLD, CHED, EC-RUS, and the alignment-based baseline predictor in Table 6. The results of TargetS, FunFOLD, CHED, and EC-RUS are the threshold of maximizing the MCC value. In terms of the independent test sets of Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}, and Zn^{2+}, the differences between SXGBsite and the best prediction quality for the AUC are 0.021 (0.758 to 0.779), 0.001 (0.779 to 0.780), 0.032 (0.856 to 0.888), 0.054 (0.891 to 0.945), and 0.052 (0.906 to 0.958), respectively, and the differences between SXGBsite and the best prediction quality for the MCC are 0.046 (0.197 to 0.243), 0.026 (0.291 to 0.317), 0.067 (0.382 to 0.449), 0.094 (0.396 to 0.490), and 0.137 (0.390 to 0.527), respectively. SXGBsite showed good prediction performance on the Mg^{2+} independent test set, and the reasons for the unsatisfactory performance on the metal ion independent test sets may be as follows: (1) TargetS uses the ligand-specific binding propensity feature to improve the prediction quality, and the features used in this study did not perform well for predicting metal ion binding residues; and (2) the volume of metal ions is smaller than that of nucleotides, which means that there are fewer binding residues (positive samples), and the lack of positive samples affected the prediction quality of the model.

Table 6. SXGBsite (average of 10 replicate experiments) compared with the existing methods on the five metal ion independent test sets.

Ligand	Predictor	SN (%)	SP (%)	ACC (%)	MCC	AUC
Ca^{2+}	TargetS [1]	13.8	99.8	98.8	0.243	0.767
	FunFOLD [1]	12.2	99.6	98.1	0.196	-
	CHED [1]	18.7	98.2	97.1	0.142	-
	alignment-based [1]	20.3	98.6	97.5	0.175	-
	EC-RUS [2]	17.3	99.6	98.7	0.225	0.779
	SXGBsite (T = 0.500)	32.6	95.6	94.9	0.139	0.758
	SXGBsite (T = 0.832)	13.3	99.7	98.7	0.197	0.758
Mg^{2+}	TargetS [1]	18.3	99.8	98.8	0.294	0.706
	FunFOLD [1]	22.0	99.1	98.3	0.215	-
	CHED [1]	14.6	98.3	97.3	0.103	-
	alignment-based [1]	14.1	99.2	98.2	0.147	-
	EC-RUS [2]	20.1	99.8	99.1	0.317	0.780
	SXGBsite (T = 0.500)	41.0	96.3	95.8	0.177	0.779
	SXGBsite (T = 0.917)	19.8	99.8	99.1	0.291	0.779
Mn^{2+}	TargetS [1]	40.1	99.5	98.7	0.449	0.888
	FunFOLD [1]	23.3	99.8	98.7	0.376	-
	CHED [1]	35.0	98.1	97.3	0.253	-
	alignment-based [1]	26.6	99.0	98.0	0.257	-
	EC-RUS [2]	35.8	99.6	98.9	0.403	0.888
	SXGBsite (T = 0.500)	44.3	98.3	97.7	0.299	0.856
	SXGBsite (T = 0.797)	34.2	99.5	98.8	0.382	0.856
Fe^{3+}	TargetS [1]	48.3	99.3	98.7	0.479	0.945
	FunFOLD [1]	47.2	99.1	98.4	0.432	-
	CHED [1]	49.2	97.0	96.3	0.279	-
	alignment-based [1]	30.0	99.2	98.3	0.300	-
	EC-RUS [2]	44.3	99.6	99.0	0.490	0.936
	SXGBsite (T = 0.500)	42.5	99.0	98.3	0.361	0.891
	SXGBsite (T = 0.670)	38.7	99.4	98.7	0.396	0.891
Zn^{2+}	TargetS [1]	46.4	99.5	98.7	0.527	0.936
	FunFOLD [1]	36.5	99.5	98.6	0.436	-
	CHED [1]	37.9	98.0	97.1	0.280	-
	alignment-based [1]	29.7	99.0	98.0	0.297	-
	EC-RUS [2]	48.9	99.2	98.6	0.437	0.958
	SXGBsite (T = 0.500)	62.4	96.7	96.3	0.323	0.906
	SXGBsite (T = 0.833)	41.0	99.2	98.6	0.390	0.906

[1] Results excerpted from Yu et al. [7]. [2] Results excerpted from Ding et al. [8]. - denotes unavailable.

Compared with TargetS, MetaDBSite, DNABR, EC-RUS, and the alignment-based baseline predictor on the DNA independent test set (Table 7), SXGBsite achieved an MCC value lower than those of TargetS and EC-RUS, and an inferior AUC value to TargetS.

Table 7. SXGBsite (average of 10 replicate experiments) compared with the existing methods on the DNA independent test set.

Ligand	Predictor	SN (%)	SP (%)	ACC (%)	MCC	AUC
DNA	TargetS [1]	41.3	96.5	93.3	0.377	0.836
	MetaDBSite [1]	58.0	76.4	75.2	0.192	-
	DNABR [1]	40.7	87.3	84.6	0.185	-
	alignment-based [1]	26.6	94.3	90.5	0.190	-
	EC-RUS [2]	31.5	97.8	95.2	0.319	0.814
	SXGBsite (T = 0.500)	36.5	95.1	92.8	0.256	0.826
	SXGBsite (T = 0.408)	46.2	92.8	91.0	0.269	0.826

[1] Results excerpted from Yu et al. [7]. [2] Results excerpted from Ding et al. [8]. - denotes unavailable.

Compared with TargetS, HemeBind, EC-RUS, and the alignment-based baseline predictor on the Heme independent test set (Table 8), SXGBsite achieved inferior MCC and AUC values to EC-RUS.

Table 8. SXGBsite (average of 10 replicate experiments) compared with the existing methods on the HEME independent test set.

Ligand	Predictor	SN (%)	SP (%)	ACC (%)	MCC	AUC
HEME	TargetS (T = 0.650) [1]	49.8	99.0	95.9	0.598	0.907
	TargetS(T = 0.180) [1]	69.3	90.4	89.1	0.426	0.907
	HemeBind [1]	86.2	90.7	90.6	0.537	-
	alignment-based [1]	51.4	97.3	94.4	0.507	-
	EC-RUS (T = 0.500) [2]	83.5	87.5	87.3	0.453	0.935
	EC-RUS (T = 0.859) [2]	55.8	99.0	96.4	0.640	0.935
	SXGBsite (T = 0.500)	61.6	97.7	95.5	0.600	0.933
	SXGBsite (T = 0.700)	52.1	99.0	96.2	0.618	0.933

[1] Results excerpted from Yu et al. [7]. [2] Results excerpted from Ding et al. [8]. - denotes unavailable.

The prediction performance of SXGBsite was similar to those of the best two methods, TargetS and EC-RUS, on the independent test sets of the five nucleotides, Mg^{2+}, DNA, and Heme. Both TargetS and EC-RUS are serial combinations of under-sampling and ensemble classifiers, which requires long calculation times. SXGBsite is a method of over-sampling and a single XGBoost classifier to quickly build high quality prediction models.

3.5. Running Time Comparison

The running time comparison of SXGBsite, EC-RUS (SVM), and EC-RUS (WSRC) on the independent test sets is provided in Table 9, and the benchmark in this study is the EC-RUC (SVM) running time. EC-RUS is a sequence-based method that was proposed by Ding et al., and its prediction quality was excellent. Ding et al. selected 19 sub-classifiers in the ensemble classifier, compared the results of ensemble SVMs and ensemble WSRCs, and concluded that ensemble WSRCs are more time-consuming than ensemble SVMs. Both SXGBsite and EC-RUS used the feature of PSSM-DCT + PSA, and the prediction model was built by SMOTE + XGBoost and RUS + ensemble classifiers, respectively. Due to having the same features, the results in Table 9 also show the running time comparison of SMOTE + XGBoost and RUS + ensemble classifiers, which means that two schemes for the class imbalance problem.

Table 9. Comparison of running time between SXGBsite and EC-RUS (SVM and WSRC) (seconds).

Dataset	SXGBsite [1]	EC-RUS (SVM) [2]	EC-RUS (WSRC) [2]	Dataset	SXGBsite [1]	EC-RUS (SVM) [2]	EC-RUS (WSRC) [2]
ATP	134.5	1746.3	7018.4	Ca^{2+}	273.6	6366.5	25627.2
ADP	146.2	4602.8	10940.5	Mg^{2+}	290.9	6558.6	31094.1
AMP	118.5	647.5	2298.1	Mn^{2+}	124.9	439.5	2806.8
GDP	90.4	284.6	685.8	Fe^{3+}	110.6	173.3	1065.9
GTP	92.6	115.8	334.6	Zn^{2+}	215.9	4284.6	20220.6
DNA	131.4	4508.5	9083.6	HEME	104.6	3459.9	2940.5

[1] The PSSM-DCT + PSA feature of SXGBsite is 183-D. [2] The PSSM-DCT + PSA feature of EC-RUS (SVM) is 143-D. SVM, support vector machine; WSRC, weighted sparse representation based classifier.

3.6. Comparison with Existing Methods on the PDNA-41 Independent Test Set

Different from the previous protein–DNA binding site dataset, PDNA-543 (9549 binding residues and 134,995 non-binding residues) and PDNA-41 (734 binding residues and 14,021 non-binding residues) are datasets constructed by Hu et al. [61]. SXGBsite constructed the prediction model by the PDNA-543 training set, obtained prediction results on the PDNA-41 independent test set, and the comparison of SXGBsite with BindN [37], ProteDNA [62], BindN+ [63], MetaDBSite [32], DP-Bind [39], DNABind [64], TargetDNA [61], and EC-RUS(DNA) [44] is provided in Table 10. SXGBsite achieved the best MCC (0.272) under *Sen* ≈ *Spec*, and achieved MCC after EC-RUS(DNA) and TargetDNA under *FPR* ≈ 5% (*FPR* = 1 - SP). The best MCC (0.279) of SXGBsite is achieved under *FPR* ≈ 10%.

Table 10. SXGBsite (average of 10 replicate experiments) compared with the existing methods on the PDNA-41 independent test set.

Predictor	SN (%)	SP (%)	ACC (%)	MCC	AUC
BindN [1]	45.64	80.90	79.15	0.143	-
ProteDNA [1]	4.77	99.84	95.11	0.160	-
BindN + (*FPR* ≈ 5%) [1]	24.11	95.11	91.58	0.178	-
BindN + (*Spec* ≈ 85%) [1]	50.81	85.41	83.69	0.213	-
MetaDBSite [1]	34.20	93.35	90.41	0.221	-
DP-Bind [1]	61.72	82.43	81.40	0.241	-
DNABind [1]	70.16	80.28	79.78	0.264	-
TargetDNA (*Sen* ≈ *Spec*) [1]	60.22	85.79	84.52	0.269	-
TargetDNA (*FPR* ≈ 5%) [1]	45.50	93.27	90.89	0.300	-
EC-RUS (DNA) (*Sen* ≈ *Spec*) [2]	61.04	77.25	76.44	0.193	-
EC-RUS (DNA) (*FPR* ≈ 5%) [2]	27.25	97.31	94.58	0.315	-
SXGBsite (*Sen* ≈ *Spec*)	60.35	85.94	84.67	0.272	0.825
SXGBsite (*FPR* ≈ 5%)	35.01	95.01	92.03	0.265	0.825

[1] Results excerpted from Hu et al. [61]. [2] Results excerpted from Shen et al. [44]. - denotes unavailable.

3.7. Case Study

The prediction results of SXGBsite are shown in the 3D models in Figure 8, and the protein–ligand complexes of 2Y4K-A and 2Y6P-A belong to the independent test sets of GDP and Mg^{2+}, respectively.

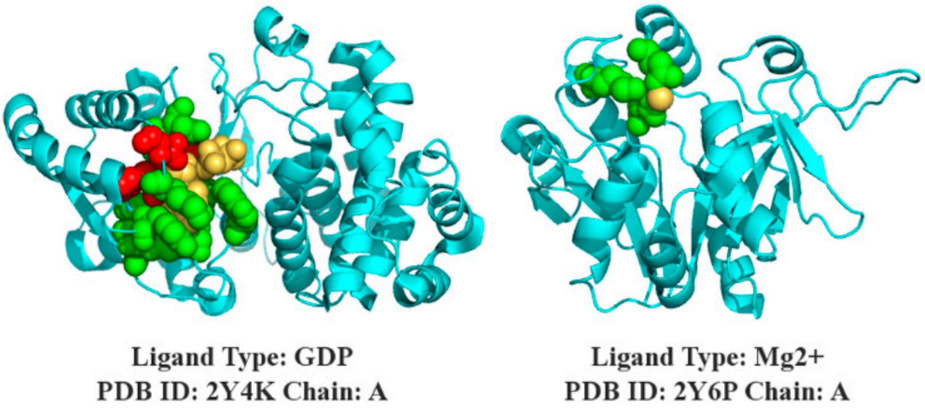

Figure 8. Prediction of SXGBsite. The cyan indicates the helix, the folding and ring structure of the protein sequence, and the yellow indicates the ligand; and true and false predictions are indicated in green and red, respectively.

4. Discussion

Many excellent computational methods are available in the field of protein–ligand binding site prediction; however, prediction efficiency can still be improved [8]. As the actual acquired protein–ligand binding site data show many fewer binding sites than non-binding sites, we selected unbalanced datasets of 12 different ligand types constructed by Yu et al. as the benchmark datasets. The adverse effects of unbalanced data on predictions are usually mitigated by over- or under-sampling methods, which are widely applied, and ensemble classifiers are often used together to overcome the loss of information caused by under-sampling. Both TargetS and EC-RUS performed well on the independent test sets built by Yu et al. by applying the scheme of under-sampling and ensemble classifiers. Although the loss of information by multiple under-sampling can be reduced by ensemble classifiers, serial combinations of multiple machine learning algorithms and high-dimensional features increase the computation time.

SXGBsite uses the features of PSSM-DCT + PSA and XGBoost with SMOTE to build prediction models, and Extreme Gradient Boosting algorithm developed by Chen et al. [46] was applied to solve overfitting and large sample sets caused by over-sampling. XGBoost's regularization technology overcomes the overfitting problem, and parallel computing can be used to quickly construct prediction models with large sample sets, which constitute the basis of SXGBsite. The threshold moving was used in this study to obtain the best MCC for comparison with other existing methods. The use of both threshold moving and sampling methods complicated the interpretation of the results, and the AUC measure without threshold change was used to better evaluate the prediction quality difference between SMOTE + XGBoost and RUS + ensemble classifiers. On the independent test sets of five nucleotides, Mg^{2+}, DNA, and Heme, the difference between the AUC of SXGBsite and the best AUC was within 0.020. Considering the decrease in the running time, we think that the difference in AUC is acceptable. On the independent test sets of 12 ligands, the new method proposed here produced a higher prediction quality with a shorter computation time using the two features and a single classifier, and produced similar results to the best-performing TargetS and EC-RUS on 8 of the 12 independent test sets.

5. Conclusions

This paper proposes a new computational method, SXGBsite. Sequence information was used for the protein–ligand binding site prediction, and features extracted by PSSM-DCT+PSA and XGBoost with SMOTE were used to construct the prediction model. On the independent test sets of 12 different ligands, SXGBsite performed similarly to the best methods on the datasets with less computation time, which could be a complement of biological experiments as well as cost reductions. The features we used did not perform well on the metal ion datasets, and adding features with better prediction performance is the next step of the study.

Author Contributions: Conceptualization, Z.Z., and Y.X.; methodology, Z.Z.; software, Z.Z.; validation, Z.Z., and Y.Z.; writing—original draft preparation, Z.Z.; and writing—review and editing, Z.Z., Y.X., and Y.Z.

Funding: This research received no external funding.

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

References

1. Roche, D.B.; Tetchner, S.J.; McGuffin, L.J. FunFOLD: An improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinform.* **2011**, *12*, 160. [CrossRef]
2. Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **1997**, *15*, 359–363. [CrossRef]
3. Roche, D.B.; Brackenridge, D.A.; McGuffin, L.J. Proteins and their interacting partners: An introduction to protein–ligand binding site prediction methods. *Int. J. Mol. Sci.* **2015**, *16*, 29829–29842. [CrossRef] [PubMed]
4. Rose, P.W.; Prlić, 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*, D345–D356. [CrossRef] [PubMed]
5. Ma, X.; Guo, J.; Liu, H.D.; Xie, J.M.; Sun, X. Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1766–1775. [CrossRef]
6. Ding, Y.; Tang, J.; Guo, F. Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information. *Int. J. Mol. Sci.* **2016**, *17*, 1623. [CrossRef]
7. Yu, D.J.; Hu, J.; Yang, J.; Shen, H.B.; Tang, J.; Yang, J.Y. Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2013**, *10*, 994–1008. [CrossRef]
8. Ding, Y.; Tang, J.; Guo, F. Identification of protein–ligand binding sites by sequence information and ensemble classifier. *J. Chem. Inf. Model.* **2017**, *57*, 3149–3161. [CrossRef]
9. Levitt, D.G.; Banaszak, L.J. POCKET: A computer graphies method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* **1992**, *10*, 229–234. [CrossRef]
10. Laskowski, R.A. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph. Model.* **1995**, *13*, 323–330. [CrossRef]
11. Xie, Z.R.; Hwang, M.J. Methods for Predicting Protein–Ligand Binding Sites. In *Molecular Modeling of Proteins*; Kukol, A., Ed.; Springer: New York, NY, USA, 2015; Volume 1215, pp. 383–398.
12. Huang, B.; Schroeder, M. LIGSITEcsc: Predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6*, 19. [CrossRef] [PubMed]
13. Liang, J.; Woodward, C.; Edelsbrunner, H. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7*, 1884–1897. [CrossRef] [PubMed]
14. Binkowski, T.A.; Naghibzadeh, S.; Liang, J. CASTp: Computed atlas of surface topography of proteins. *Nucleic Acids Res.* **2003**, *31*, 3352–3355. [CrossRef] [PubMed]
15. Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* **2006**, *34*, W116–W118. [CrossRef]
16. Tian, W.; Chen, C.; Lei, X.; Zhao, J.; Liang, J. CASTp 3.0: Computed atlas of surface topography of proteins. *Nucleic Acids Res.* **2018**, *46*, W363–W367. [CrossRef]
17. Fuller, J.C.; Martinez, M.; Henrich, S.; Stank, A.; Richter, S.; Wade, R.C. LigDig: A web server for querying ligand–protein interactions. *Bioinformatics* **2014**, *31*, 1147–1149. [CrossRef]
18. Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinform.* **2009**, *10*, 168. [CrossRef]
19. Schmidtke, P.; Le Guilloux, V.; Maupetit, J.; Tuffery, P. Fpocket: Online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* **2010**, *38*, 582–589. [CrossRef]

20. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef]
21. UniProt Consortium. UniProt: A hub for protein information. *Nucleic Acids Res.* **2015**, *43*, 204–212. [CrossRef]
22. Bolton, E.E.; Wang, Y.; Thiessen, P.A.; Bryant, S.H. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*; Wheeler, R.A., Spellmeyer, D.C., Eds.; Elsevier: Amsterdam, The Netherlands, 2008; Volume 4, pp. 217–241.
23. Hastings, J.; de Matos, P.; Dekker, A.; Ennis, M.; Harsha, B.; Kale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M.; et al. The ChEBi reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Res.* **2013**, *41*, 456–463. [CrossRef] [PubMed]
24. Okuda, S.; Yamada, T.; Hamajima, M.; Itoh, M.; Katayama, T.; Bork, P.; Goto, S.; Kanehisa, M. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.* **2008**, *36*, 423–426. [CrossRef] [PubMed]
25. Wójcikowski, M.; Ballester, P.J.; Siedlecki, P. Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.* **2017**, *7*, 46710. [CrossRef]
26. Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A new open-source player in the drug discovery field. *J. Cheminform.* **2015**, *7*, 26. [CrossRef] [PubMed]
27. Wójcikowski, M.; Zielenkiewicz, P.; Siedlecki, P. DiSCuS: An open platform for (not only) virtual screening results management. *J. Chem. Inf. Model* **2014**, *54*, 347–354. [CrossRef]
28. Babor, M.; Gerzon, S.; Raveh, B.; Sobolev, V.; Edelman, M. Prediction of transition metal-binding sites from apo protein structures. *Proteins* **2008**, *70*, 208–217. [CrossRef]
29. Capra, J.A.; Laskowski, R.A.; Thornton, J.M.; Singh, M.; Funkhouser, T.A. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.* **2009**, *5*, e1000585. [CrossRef]
30. Yang, J.; Roy, A.; Zhang, Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, *29*, 2588–2595. [CrossRef]
31. Liu, R.; Hu, J. HemeBIND: A novel method for heme binding residue prediction by combining structural and sequence information. *BMC Bioinform.* **2011**, *12*, 207. [CrossRef]
32. Si, J.; Zhang, Z.; Lin, B.; Schroeder, M.; Huang, B. MetaDBSite: A meta approach to improve protein DNA-binding sites prediction. *BMC Syst. Biol.* **2011**, *5*, S7. [CrossRef]
33. Chen, K.; Mizianty, M.J.; Kurgan, L. ATPsite: Sequence-based prediction of ATP-binding residues. *Proteome Sci.* **2011**, *9*, S4. [CrossRef] [PubMed]
34. Chen, K.; Mizianty, M.J.; Kurgan, L. Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. *Bioinformatics* **2011**, *28*, 331–341. [CrossRef] [PubMed]
35. Ofran, Y.; Mysore, V.; Rost, B. Prediction of DNA-binding residues from sequence. *Bioinformatics* **2007**, *23*, i347–i353. [CrossRef] [PubMed]
36. Yan, C.; Terribilini, M.; Wu, F.; Jernigan, R.L.; Dobbs, D.; Honavar, V. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinform.* **2006**, *7*, 262. [CrossRef] [PubMed]
37. Wang, L.; Brown, S.J. BindN: A web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.* **2006**, *34*, W243–W248. [CrossRef]
38. Wang, L.; Yang, M.Q.; Yang, J.Y. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genom.* **2009**, *10*, S1. [CrossRef]
39. Hwang, S.; Gou, Z.; Kuznetsov, I.B. DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* **2007**, *23*, 634–636. [CrossRef]
40. Ahmad, S.; Gromiha, M.M.; Sarai, A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* **2004**, *20*, 477–486. [CrossRef]
41. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
42. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
43. Lu, C.Y.; Min, H.; Gui, J.; Zhu, L.; Lei, Y.K. Face recognition via weighted sparse representation. *J. Vis. Commun. Image Represent.* **2013**, *24*, 111–116. [CrossRef]
44. Shen, C.; Ding, Y.; Tang, J.; Song, J.; Guo, F. Identification of DNA–protein binding sites through multi-scale local average blocks on sequence information. *Molecules* **2017**, *22*, 2079. [CrossRef] [PubMed]

45. Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. In Proceedings of the 22nd Acm sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
46. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
47. Ahmed, N.; Natarajan, T.; Rao, K.R. Discrete cosine transform. *IEEE T. Comput.* **1974**, *100*, 90–93. [CrossRef]
48. Yu, D.J.; Hu, J.; Huang, Y.; Shen, H.B.; Qi, Y.; Tang, Z.M.; Yang, J.Y. TargetATPsite: A template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble. *J. Comput. Chem.* **2013**, *34*, 974–985. [CrossRef] [PubMed]
49. Nanni, L.; Lumini, A.; Brahnam, S. An empirical study of different approaches for protein classification. *Sci. World J.* **2014**, *2014*, 1–17. [CrossRef]
50. Nanni, L.; Brahnam, S.; Lumini, A. Wavelet images and Chou's pseudo amino acid composition for protein classification. *Amino Acids* **2012**, *43*, 657–665. [CrossRef]
51. Wang, Y.; Ding, Y.; Guo, F.; Wei, L.; Tang, J. Improved detection of DNA-binding proteins via compression technology on PSSM information. *PLoS ONE* **2017**, *12*, e0185587. [CrossRef]
52. Ahmad, S.; Gromiha, M.M.; Sarai, A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins* **2003**, *50*, 629–635. [CrossRef]
53. Yang, J.; Roy, A.; Zhang, Y. BioLiP: A semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.* **2012**, *41*, D1096–D1103. [CrossRef]
54. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
55. Joo, K.; Lee, S.J.; Lee, J. Sann: Solvent accessibility prediction of proteins by nearest neighbor method. *Proteins* **2012**, *80*, 1791–1797. [CrossRef] [PubMed]
56. Hu, J.; He, X.; Yu, D.J.; Yang, X.B.; Yang, J.Y.; Shen, H.B. A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction. *PLoS ONE* **2014**, *9*, e107676. [CrossRef] [PubMed]
57. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
58. Deng, L.; Sui, Y.; Zhang, J. XGBPRH: Prediction of Binding Hot Spots at Protein–RNA Interfaces Utilizing Extreme Gradient Boosting. *Genes* **2019**, *10*, 242. [CrossRef] [PubMed]
59. Wang, H.; Liu, C.; Deng, L. Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. *Sci. Rep.* **2018**, *8*, 14285. [CrossRef]
60. Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *Ann Stat* **2000**, *28*, 337–407. [CrossRef]
61. Hu, J.; Li, Y.; Zhang, M.; Yang, X.; Shen, H.B.; Yu, D.J. Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *14*, 1389–1398. [CrossRef]
62. Chu, W.Y.; Huang, Y.F.; Huang, C.C.; Cheng, Y.S.; Huang, C.K.; Oyang, Y.J. ProteDNA: A sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. *Nucleic Acid Res.* **2009**, *37*, 396–401. [CrossRef]
63. Wang, L.; Huang, C.; Yang, M.Q.; Yang, J.Y. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.* **2010**, *4*, 1–9. [CrossRef]
64. Li, B.Q.; Feng, K.Y.; Ding, J.; Cai, Y.D. Predicting DNA-binding sites of proteins based on sequential and 3D structural information. *Mol. Genet. Genom.* **2014**, *289*, 489–499. [CrossRef] [PubMed]