Journal of
CHEMICAL INFORMATION
AND MODELING

Article

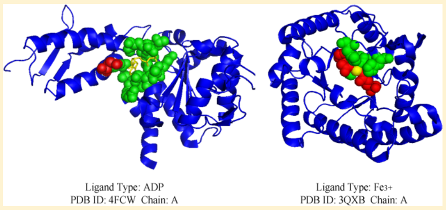
pubs.acs.org/jcim

Identification of Protein−Ligand Binding Sites by Sequence Information and Ensemble Classifier

Yijie Ding,[†] Jijun Tang,[†,‡] and Fei Guo*,[†]

[†]School of Computer Science and Technology, Tianjin University, No. 135, Yaguan Road, Tianjin Haihe Education Park, Tianjin 300350, China

[‡]Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29208, United States



ABSTRACT: Identifying protein−ligand binding sites is an important process in drug discovery and structure-based drug design. Detecting protein−ligand binding sites is expensive and time-consuming by traditional experimental methods. Hence, computational approaches provide many effective strategies to deal with this issue. Recently, lots of computational methods are based on structure information on proteins. However, these methods are limited in the common scenario, where both the sequence of protein target is known and sufficient 3D structure information is available. Studies indicate that sequence-based computational approaches for predicting protein−ligand binding sites are more practical. In this paper, we employ a novel computational model of protein−ligand binding sites prediction, using protein sequence. We apply the Discrete Cosine Transform (DCT) to extract feature from Position-Specific Score Matrix (PSSM). In order to improve the accuracy, Predicted Relative Solvent Accessibility (PRSA) information is also utilized. The predictor of protein−ligand binding sites is built by employing the ensemble weighted sparse representation model with random under-sampling. To evaluate our method, we conduct several comprehensive tests (12 types of ligands testing sets) for predicting protein−ligand binding sites. Results show that our method achieves better Matthew's correlation coefficient (MCC) than other outstanding methods on independent test sets of ATP (0.506), ADP (0.511), AMP (0.393), GDP (0.579), GTP (0.641), Mg^{2+} (0.317), Fe^{3+} (0.490) and HEME (0.640). Our proposed method outperforms earlier predictors (the performance of MCC) in 8 of the 12 ligands types.

INTRODUCTION

Detection of protein−ligand binding sites is important for understanding the function of protein and drug discovery. Experimental methods are of high cost, and lots of 3D structures (proteins) are unknown. BioLiP[2] indicated that about 40% of proteins did not have relevant Ligand-Binding Site (LBS) information in the Protein Data Bank (PDB).[1] Many computational methods have been developed to provide complementary information for traditional methods.[3−34] The ligand-binding site predictor can be classified into sequence-based, structure-based and hybrid (integrated sequence and structure informatin) methods.

The sequences of proteins can be utilized for identifying protein−ligand binding sites. Rate4Site[14] and ConSurf[31] employed Multiple-Sequence Alignment (MSA) (alignment-based) to detect hot spots. SVMPred[13] and NsitePred[15] utilized sequence information, evolutionary profiles (Position Specific Scoring Matrix (PSSM)), predicted secondary structure and solvent accessibility (based on sequence) built prediction model via Support Vector Machine (SVM).[35] TargetS[12] used the PSSM of protein and predicted protein secondary structure to built an improved AdaBoost model, which is based on random under-sampling and ensemble scheme. TargetATPsite[16] employed image sparse representation to extract feature from PSSM. And an ensemble SVM was used as the predictor. TargetATP[17] also utilized PSSM as input feature and a modified AdaBoost ensemble scheme as model to predict protein−ATP binding sites. Meta DNA Binding Site (MetaDBSite)[11] integrated the prediction result from six available online web servers: DNA Interaction Sites Identified from Sequence (DISIS),[36] DNA Binding Residues (DNABindR),[37] BindN,[38] BindN Random Forest (BindN-RF),[39] DNA Protein-Binding (DP-Bind)[40] and DNA Binding Sites Prediction (DBS-PRED),[41] and it solely used sequence information on proteins. DNA Binding Residues (DNABR)[20] used the Random Forest (RF) classifier and sequence-based features (physicochemical properties of amino acids) built prediction model. University of Tokyo Proteins (UTProt) Galaxy[6] calculated PSSM profile of the protein sequence as the feature vector, and constructed Machine Learning (ML) model to predict binding sites. They also compared the predictive performance of SVM, Neural Network and RF, respectively.

The structure-based methods containing Ligands-binding Sites detection (LIGSITE),[23] Computed Atas of Surface Topography of proteins (CASTp),[24] POCKET,[26] Fpocket,[27] SURFNET,[25] SITEHOUND[29] and Q-SiteFinder[28] et al. Above methods utilized the protein 3D information to detect the potential pockets. LIGSITE[23] identified binding site with a series of simple operations on a cubic grid. CASTp[24] can located and measured pockets and voids on 3D protein structures. POCKET[26] used a modification of the marching

Received: May 31, 2017
Published: November 10, 2017

cubes algorithm modeled surfaces of these pockets. Fpocket used Structure Based Virtual Screening (SBVS) approaches to detect pocket and cavity on protein surfaces. The SURFNET[25] generated molecular gaps and surfaces between surfaces from 3D PDB-format file. SITEHOUND[29] detected regions of ligand binding sites by a probe molecule. Q-SiteFinder[28] located energetically favorable binding sites via the interaction energy between a simple van der Waals probe and the protein. FunFOLD[18] employed only a 3D model and list of templates to detect binding sites. CHED[19] was based on the structure of the apo state. The transition metal-binding sites were identified with a selectivity more than 95%. Most methods (above structure-based methods) are used to identify the pocket of target protein.

Also, some methods integrated sequence and structure information to improve the performance of prediction. For example, Consensus Approach (COACH)[10] derived LBSs from structure-related templates and evolution information on proteins. LIGSITE^{csc}[32] extended the LIGSITE[23] by integrating the degree of involved surface residues. SURFNET-ConSurf[34] also combined evolution information with pocket detection. ConCavity[33] combined evolutionary information on sequence with structure information for detecting the cavities of protein surface. HemeNet[22] and HemeBind[21] combined the structure-based and sequence-based models, which is significantly better than the individual classifier alone for specifically predicting HEME binding residues.

Although prediction accuracies of structure-based and hybrid methods are better than sequence-based approaches. The hybrid and structure-based are limited in the sufficient 3D structure. Hence, sequence-based methods are more practical. In this study, we focus on sequence-based methods for detecting protein–ligand binding sites.

Inspired by previous work,[16,42–46,55] we use the Discrete Cosine Transform (DCT)[44] to extract compressed features from PSSM. The PSSM of sequence is the major information for predicting the binding sites. The conservation or variation of sequence is determined by many factors (including preserving the 3D structure and stability, reducing amyloid aggregation, and also conservation of functions) during evolution. These factors influence protein binding with partners of other proteins, nucleotides, ions, lipids, or nutrients etc. Hence, the PSSM (containing evolutionary information) may pickup the signals/features important for ligand binding. Predicted Relative Solvent Accessibility (PRSA) information is also utilized to improve the accuracy of prediction. The number of ligand-binding residues (minority class) is significantly fewer than that of nonbinding residues (majority class). Sample rescaling is the most straightforward strategy for dealing with the issue of class imbalance. To handle the imbalanced problem, we employ the ensemble weighted sparse representation model with random under-sampling. The performance of our method is evaluated on 12 different types of ligands, containing 5 types of nucleotides, 5 types of metal ions, DNA and HEME. The 12 types of ligands both include training sets and independent testing sets. Experiments of independent testing show that our proposed method obtains better results compared with other methods.

■ METHOD

In order to detect protein–ligand binding sites by Machine Learning (ML) approaches, the main challenge is to extract the crucial information on protein–ligand binding sites. The identification of protein–ligand binding sites could be regarded as a traditional binary classification problem. The binding residue is not isolated, we consider w contiguous residues as a window, including the target residue and $(w - 1)/2$ neighboring residues on both sides of the target residue. We utilize PSSM to represent the evolutionary conservatism of protein sequence. And we apply the DCT[44,45] to extract feature from PSSM of each residue. For this subsequence of w residues, encoding with a multidimensional vector can be built on the PSSM and predicted relative solvent accessibility. At last, the ML is used to build prediction model for detecting protein–ligand binding sites.

Extracting Feature from PSSM. The evolutionary profiles of protein sequence could be described by PSSM, which is generated by PSI-BLAST[47] (BLAST+[48] options: -num_iterations 3 -db nr -inclusion_ethresh 0.001). The PSSM is a matrix of dimensions $L \times 20$ (L rows and 20 columns), formulated as follows:

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

Then, we extract the attribute of a target residue s ($1 \le s \le L$, the out of bounds value, replaced by 0) via a window with w residues, and obtain a total of $w \times 20$ original PSSM scores, formulated as follows:

$$\text{PSSM}_s = \begin{bmatrix} P_{s-(w-1)/2,1} & P_{s-(w-1)/2,2} & \cdots & P_{s-(w-1)/2,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{s-1,1} & P_{s-1,2} & \cdots & P_{s-1,20} \\ P_{s,1} & P_{s,2} & \cdots & P_{s,20} \\ P_{s+1,1} & P_{s+1,2} & \cdots & P_{s+1,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{s+(w-1)/2,1} & P_{s+(w-1)/2,2} & \cdots & P_{s+(w-1)/2,20} \end{bmatrix}_{w\times 20} \tag{2}$$

Inspired by previous work,[42–46] we use the DCT[44] to extract compressed feature from PSSM$_s$ of target residue s. The DCT is a separable transformation for converting a discrete signal into elementary frequency components. Moreover, the DCT is widely used in lossy data compression for its capability to concentrate information into a small number of coefficients. Here, we use 2 dimensions DCT (2D-DCT) to compress PSSM$_s$. Given an input matrix Mat = PSSM$_s \in \Re^{w\times 20}$, its 2D-DCT transformation is defined as

$$\text{DCT}\,(i, j) = \alpha_i\alpha_j \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \text{Mat}(m, n)\cos\frac{\pi(2m+1)i}{2M} \times \cos\frac{\pi(2n+1)j}{2N} \tag{3a}$$

$$\alpha_i = \begin{cases} \sqrt{1/M}, & i = 0 \\ \sqrt{2/M}, & 1 \le i \le M - 1 \end{cases} \tag{3b}$$

$$\alpha_j = \begin{cases} \sqrt{1/N}, & j = 0 \\ \sqrt{2/N}, & 1 \le j \le N - 1 \end{cases} \tag{3c}$$

where $0 \le i < M$, $0 \le j < N$.

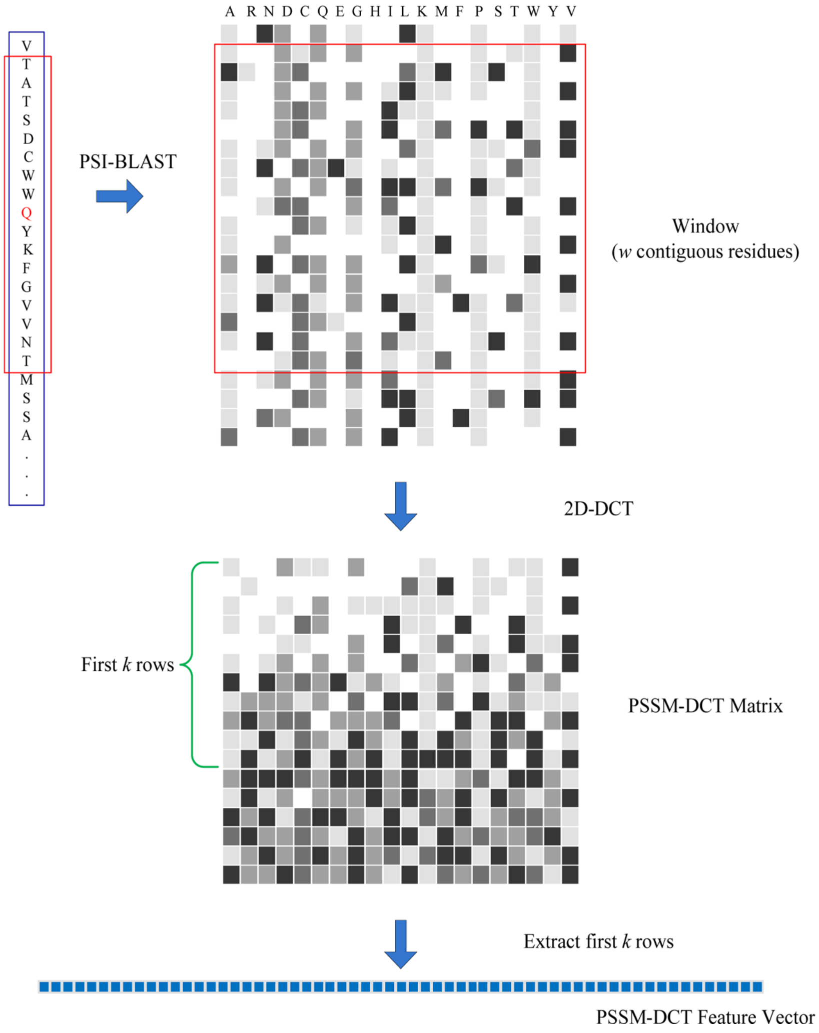


Figure 1. Schematic diagram for extracting PSSM-DCT feature.

A major characteristic of DCT is the conversion of information density from evenly to unevenly distribution. Most of natural signals (PSSM) are concentrated in the low-frequency part of the compressed PSSM, which distribute in the upper left corner. In our work, the final PSSM-DCT descriptor is obtained by retaining the first k rows for $k \times 20$ coefficients. The schematic diagram of PSSM-DCT is shown in Figure 1.

Predicted Relative Solvent Accessibility. Solvent accessibility is particularly significant, and it is closely related to the spatial features and protein folding. In fact, there is an inseparable relationship between solvent accessibility and protein−ligand interactions. Moreover, Ahmad et al.[49] has demonstrated the important role of solvent accessibility to residues in predicting protein−DNA interactions. We utilize the Solvent Accessibility prediction by Nearest Neighbor method (SANN)[50] program (downloaded at http://lee.kias.re.kr/newton/sann/) to obtained the Predicted Relative Solvent Accessibility (PRSA) characteristics of each residue by the corresponding sequence.

Weighted Sparse Representation based Classifier. With the development of Compressed Sensing (CS) theory, most researchers have recently paid attention to sparse representation[51] in pattern recognition and image processing. The Sparse Representation based Classification (SRC)[52,53] utilizes training samples to represent a new testing sample. To represent testing samples, SRC builds a linear combination of training set via computing sparse representation matrix. Then, it calculates reconstruction residuals of each class by above sparse representation matrix. At last, the test sample would be determined to the class with minimal reconstruction error. Related SRC methods have been applied to some biological problems, such as prediction of protein−protein interactions.[54,55]

There are C classes training samples, and n_c training samples are from the c-th class (columns of $\mathbf{X}^c = [X_1^c, ..., X_{n_c}^c] \in \Re^{m \times n_c}$). The m is the dimension of sample. So, we get the matrix of training sample ($\mathbf{X} = [\mathbf{X}^1, ..., \mathbf{X}^C] \in \Re^{m \times n}$). $n = \sum_{c=1}^{C} n_c$ is the total number of training samples. The new (test) sample $\mathbf{y}$ from the same class will approximately lie in the linear span of the training samples associated with class c as follows:

$$\mathbf{y}^c = \mathbf{X}^c \boldsymbol{\alpha}_0^c \tag{4}$$

Though c is unknown, the linear representation of $\mathbf{y}$ can be rewritten in terms of whole training set representation as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha}_0 \tag{5}$$

where coefficient vector $\boldsymbol{\alpha}_0 = [0, \alpha_0^c, 0]^T$, the nonzero coefficients associate with the c-th class.

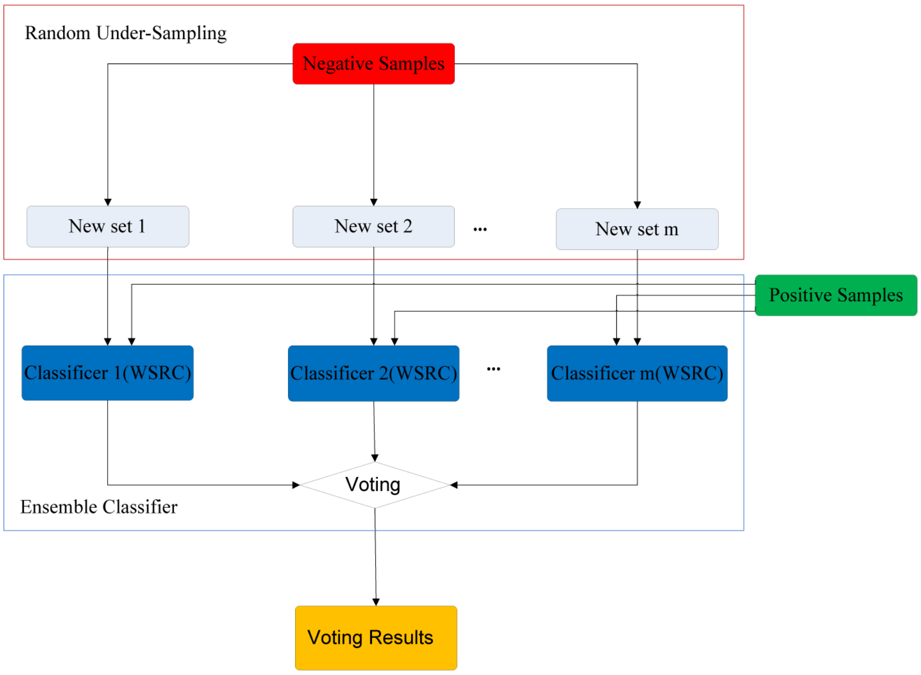


Figure 2. Overview of the Ensemble Classifier.

The SRC aims to search the $\boldsymbol{\alpha}$ vector, which can satisfy both eq 5 and minimize the l_0-norm as follows:

$$\hat{\boldsymbol{\alpha}}_{\mathbf{0}} = \arg\min \|\boldsymbol{\alpha}\|_0 \tag{6a}$$

$$\text{subject to } \mathbf{y} = \mathbf{X}\boldsymbol{\alpha} \tag{6b}$$

But the problem (6a) of finding the sparsest solution of linear equations is NP-hard. The CS reveals that if the solution $\boldsymbol{\alpha}$ is sparse enough, solving the convex l_1-minimization problem is approximate to l_0-minimization as follows:

$$\hat{\boldsymbol{\alpha}}_{\mathbf{1}} = \arg\min \|\boldsymbol{\alpha}\|_1 \tag{7a}$$

$$\text{subject to } \mathbf{y} = \mathbf{X}\boldsymbol{\alpha} \tag{7b}$$

To deal with occlusion, eq 7a can be extended to the stable l_1-minimization problem as follows:

$$\hat{\boldsymbol{\alpha}}_{\mathbf{1}} = \arg\min \|\boldsymbol{\alpha}\|_1 \tag{8a}$$

$$\text{subject to } \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2 \le \epsilon \tag{8b}$$

where $\epsilon > 0$ denotes to the tolerance of reconstruction error.

The algorithm of classifier assigns the test sample $\mathbf{y}$ to class c as follows:

$$\min_c r_c(\mathbf{y}) = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}_{\mathbf{1}}{}^c\|_2 \tag{9}$$

where $r_c(\mathbf{y})$ denotes the residuals between itself $(\mathbf{y})$ and $\mathbf{X}\hat{\boldsymbol{\alpha}}_{\mathbf{1}}^{c}$ (class c), c = 1, ..., C. SRC assigns it $(\mathbf{y})$ to the class which has minimal residuals.

Lu et al.[56] present the Weighted Sparse Representation based Classification (WSRC) method. They used all the training data as dictionary, and imposed the locality on the l_1 regularization. It solves the following weighted l_1-minimization problem:

$$\hat{\boldsymbol{\alpha}}_{\mathbf{1}} = \arg\min \|\mathbf{W}\boldsymbol{\alpha}\|_1 \tag{10a}$$

$$\text{subject to } \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2 \le \epsilon \tag{10b}$$

$$\text{diag}(\mathbf{W}) = [\text{dist}(\mathbf{y}, \mathbf{x}_1^1), ..., \text{dist}(\mathbf{y}, \mathbf{x}_{\mathbf{n}_c}^C)]^T \tag{10c}$$

where $\mathbf{W}$ is a block-diagonal matrix, $\text{dist}(\mathbf{y},\mathbf{x}_\mathbf{i}^{\ c}) = \exp\ (-\|\mathbf{y} - \mathbf{x}_\mathbf{i}^{\ c}\|/2\sigma^2)$, $\mathbf{y}$, $\mathbf{x}_\mathbf{i}^{\ c}$ are two samples and σ is the Gaussian kernel width. i is the sample index of training set in class c.

The values of Gaussian Distance (GD) are calculated for the weight of each training sample. The WSRC algorithm is listed as follows:

Algorithm 1 Weighted Sparse Representation based Classifier (WSRC)

```
Input: The matrix $\mathbf{X} \in \mathfrak{R}^{m \times n}$ of training samples, a test sample $\mathbf{y} \in \mathfrak{R}^m$; Two parameters:
    tolerance of reconstruction error $\epsilon$, gaussian kernel width $\sigma$;
Output: The prediction label of $\mathbf{y}$;
Normalize the variables (each dimension) of $\mathbf{X}$ to $0 - 1$ by the min-max normalization;
Make up matrix $\mathbf{W}$ via calculating the value of gaussian distance between each sample
    (in matrix $\mathbf{X}$) and $\mathbf{y}$;
Solve the $l_1$-minimization problem (10a);
Calculate each reconstruction error of $C$ classes: $r_c(\mathbf{y}) = \|\mathbf{y} - \mathbf{X}\hat{\alpha_\mathbf{1}}^c\|, c = 1, ..., C$;
Assign $\mathbf{y}$ to the class $c$ by the rule: $identity(\mathbf{y}) = arg\ min_c(r_c(\mathbf{y}))$;
```

For the classification problem (prediction of binding site is the problem of binary classification), the WSRC does not output the predictive probability of each class, but the output of WSRC is only each residual of C classes. If we want to get probability of each class, it is not available. Because minimal residuals of class has more possibility. To map result in the range of [0,1], we define three types of score value for binding sites prediction as follows:

$$\text{score}_1(\mathbf{y})_{\text{binding}} = 1 - \frac{r_{\text{binding}}(\mathbf{y})}{r_{\text{nonbinding}}(\mathbf{y}) + r_{\text{binding}}(\mathbf{y})} \tag{11a}$$

$$\text{score}_2(\mathbf{y})_{\text{binding}} = 2^{-r_{\text{binding}}(\mathbf{y})/r_{\text{nonbinding}}(\mathbf{y})} \tag{11b}$$

$$\text{score}_3(\mathbf{y})_{\text{binding}} = \frac{1}{1 + e^{-(r_{\text{nonbinding}}(\mathbf{y}) - r_{\text{binding}}(\mathbf{y}))}} \tag{11c}$$

where $r_{\text{binding}}(\mathbf{y})$ and $r_{\text{nonbinding}}(\mathbf{y})$ are the reconstruction error of WSRC assigning the test sample $\mathbf{y}$ to binding and nonbinding site (in this study, C = 2, so $r_1(\mathbf{y}) = r_{\text{binding}}(\mathbf{y})$ and $r_2(\mathbf{y}) = r_{\text{nonbinding}}(\mathbf{y})$), respectively. We will evaluate the performance of above three types of probability mapping functions.

Table 1. Detailed Compositions of 12 Different Data Sets

		Training Set		Independent Test Set		
Ligand Category	Ligand Type	No. of Sequences	numP, numN[a]	No. of Sequences	numP, numN[a]	Total No. of Sequences
Nucleotides	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 Ions	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

[a] numP and numN represent the numbers of positive (binding residues) samples and negative (nonbinding residues) samples, respectively.

Ensemble Classifier and Random Under-Sampling. To handle the imbalanced classification problem, we employ ensemble classifier.[57,58] The number of nonbinding examples (majority class) is much more than that of binding examples (minority class). The bootstrap resampling approach[57,58] is employed to improve the accuracy of prediction. The majority class is repeatedly random under sampled (RUS) with m times, and the size of subset is equal to the size of minority class. Then, we could get m subsets from the set of nonbinding examples. m new training sets are collected by combining m subsets with the set of binding examples. Training m classifiers $\{f(\mathbf{x})^i\}_{i=1}^{m}$ by using m new training sets. The final result of new sample $\mathbf{y}$ is determined by averaging outputs of m classifiers. We calculate the each probability value $\{\text{score}(\mathbf{y})^i_{\text{binding}}\}_{i=1}^{m}$ of the m outputs and get the final score $P(\mathbf{y})$ as follows:

$$P(\mathbf{y}) = \frac{1}{m}\sum_{i=1}^{m} \text{score}(\mathbf{y})^i_{\text{binding}}, \qquad i = 1, ..., m \tag{12}$$

where $P(\mathbf{y})$ is the probability factor of new sample $\mathbf{y}$, $\text{score}(\mathbf{y})^i_{\text{binding}}$ is the probability value of i-th base classifier. All of above feature vector should be normalized to a range of [0,1], using the min−max normalization.

Figure 2 shows the overview of the proposed component ensemble classifier. The method of Ensemble Classifier with Random Under-Sampling (EC-RUS) can handle the imbalanced classification problem and improve the generalization performance of model.

■ RESULTS

We evaluate our method on several protein−ligand binding sites data sets, including five types of nucleotides (ATP, ADP, AMP, GTP and GDP), five types of metal ions (Ca^{2+}, Zn^{2+}, Mg^{2+}, Mn^{2+} and Fe^{3+}), DNA and HEME. First, we analyze the performance of binding site features (such as PSSM, PSSM-DCT and PRSA) and models (WSRC and SVM). In addition, we compare our proposed method with other methods on the training sets of above 12 types by 5-fold cross-validation. Then, we use these training sets to construct models and predict the corresponding independent data sets of 12 types, respectively.

Data sets of Protein−Ligand Binding Sites. Most ligand-binding sites prediction methods use protein 3D structures from Protein Data Bank (PDB)[1] as templates. Some studies[2,59−61] filtered out the ligand-protein interaction from the PDB and several other databases of purified ligand-

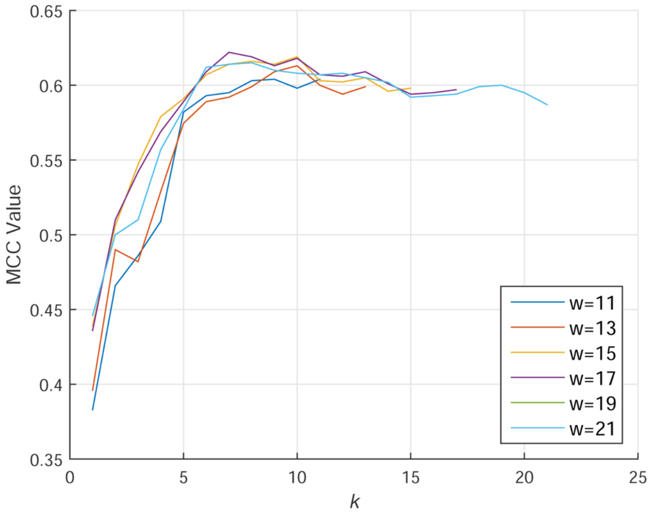


Figure 3. MCC values (on independent testing set of GTP) of our method with different values of window size w and first k rows (PSSM-DCT) by EC-RUS (utilizing 7 WSRC models).

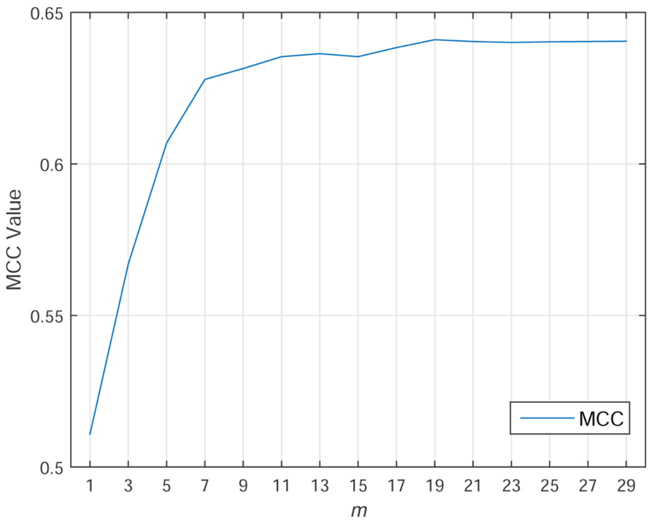


Figure 4. MCC values (on independent testing set of GTP) of our method with different number of base classifiers (PSSM-DCT + PRSA).

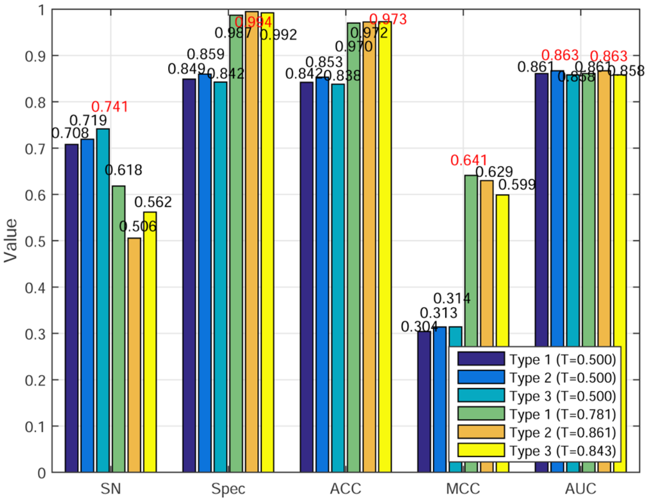


Figure 5. Results for different probability mapping functions on independent testing set of GTP. Type 1, 2 and 3 represent eqs 11a, 11b and 11c, respectively.

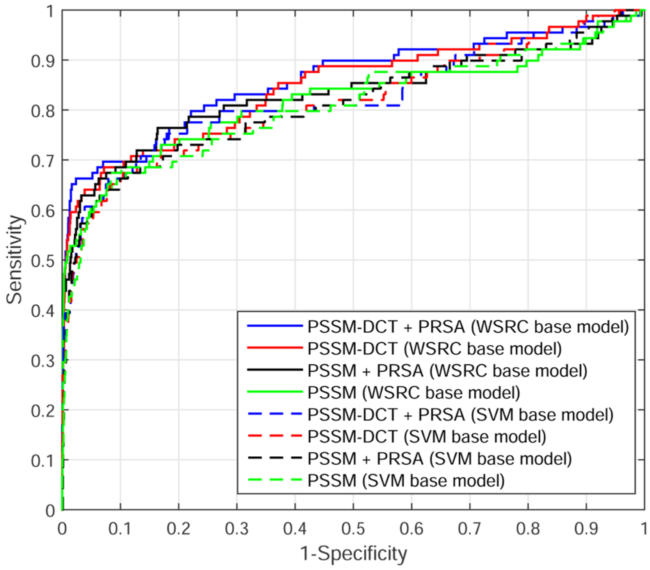


Figure 6. ROC (on independent testing set of GTP) of different feature and classifiers.

protein interaction, including LigASite,[59] FireDB,[60] BioLiP[2] and PDBbind.[61] Yu et al.[12] constructed training and independent validation data sets based on the BioLip database[2] rather than on PDB. Yu et al. considered 12 different types of ligands, containing 5 types of nucleotides, 5 types of metal ions, DNA and HEME. The 12 types of ligands both include training sets and independent testing sets. We evaluate our method by cross-validation on training sets. Moreover, independent test is often utilized to test the generalization capability. Table 1 summarizes the detailed compositions of 12 different data sets. The source code and all data sets are available at https://github.com/6gbluewind/protein_ligand_binding_site.

Evaluation Measurements. There are some parameters are employed to evaluate the performance: accuracy (ACC), sensitivity (SN), specificity (Spec), Matthew's correlation coefficient (MCC), defined as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{13a}$$

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

$$\mathrm{Spec} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{13c}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}} \tag{13d}$$

true positive (TP) denotes the number of true protein−ligand binding sites with predicted correctly; true negative (TN) denotes the number of true nonbinding sites with predicted correctly; false negative (FN) denotes the number of true protein−ligand binding sites, which are assigned to be nonbinding; false positive (FP) denotes the number of true nonbinding sites, which are assigned to be binding sites.

Table 2. Comparison of the Prediction Performance with Different Features and Classifiers on Independent Testing Set of GTP

EC-RUS	Feature	Threshold	SN (%)	Spec (%)	ACC (%)	MCC	AUC
WSRC	PSSM-DCT + PRSA	0.500	70.8	84.9	84.2	0.304	0.861
		0.781[a]	61.8	98.7	97.0	0.641	0.861
	PSSM-DCT	0.500	71.9	81.6	81.2	0.274	0.848
		0.823[a]	53.9	99.1	97.1	0.622	0.848
	PSSM+PRSA	0.500	76.4	82.9	82.6	0.310	0.832
		0.841[a]	42.7	99.8	97.2	0.619	0.832
	PSSM	0.500	74.2	79.2	79.0	0.263	0.818
		0.817[a]	47.2	99.7	97.3	0.631	0.818
SVM	PSSM-DCT + PRSA	0.500	74.2	83.8	83.3	0.309	0.824
		0.869[a]	39.3	99.5	96.8	0.546	0.824
	PSSM-DCT	0.500	70.8	80.5	80.1	0.259	0.814
		0.822[a]	48.3	98.1	95.9	0.495	0.814
	PSSM + PRSA	0.500	73.0	79.5	79.2	0.261	0.811
		0.811[a]	43.8	98.8	96.3	0.513	0.811
	PSSM	0.500	70.8	78.5	78.1	0.241	0.808
		0.835[a]	42.7	98.8	96.2	0.490	0.808

[a] Threshold is identified by maximizing the MCC value of predictions on the independent testing set.

DOI: 10.1021/acs.jcim.7b00307
J. Chem. Inf. Model. 2017, 57, 3149−3161

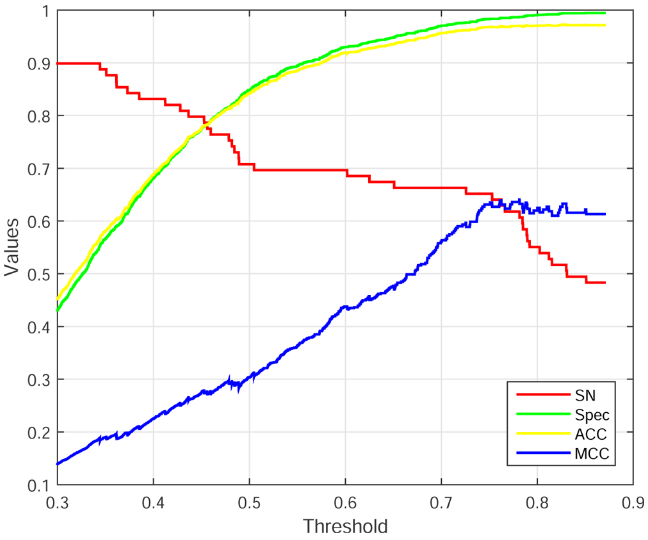


Figure 7. Results for different thresholds of probability on independent testing set of GTP.

Table 3. Performance of Single Classifier and Ensemble Classifier on Independent Testing Set of GTP

Type	Model	SN (%)	Spec (%)	ACC (%)	MCC	AUC
single subclassifier	WSRC, T = 0.500[a]	78.6	78.7	78.7	0.280	0.852
	WSRC, T = 0.807[a]	58.4	98.5	96.7	0.599	0.852
	SVM, T = 0.500[a]	72.4	81.2	80.1	0.283	0.816
	SVM, T = 0.866[a]	37.2	98.4	95.5	0.497	0.816
EC-RUS	EC-RUS(WSRC, T = 0.500)[a]	70.8	84.9	84.2	0.304	0.861
	EC-RUS(WSRC, T = 0.781)[a]	61.8	98.7	97.0	0.641	0.861
	EC-RUS(SVM, T = 0.500)[a]	74.2	83.8	83.3	0.309	0.824
	EC-RUS(SVM, T = 0.869)[a]	39.3	99.5	96.8	0.546	0.824
single classifier	WSRC, T = 0.500[a]	56.8	99.6	97.6	0.618	0.856
	WSRC, T = 0.465[a]	58.4	99.5	97.7	0.621	0.856
	SVM, T = 0.500[a]	44.9	98.8	96.4	0.561	0.842
	SVM, T = 0.300[a]	48.9	98.0	96.6	0.583	0.842

[a] The feature is PSSM-DCT + PRSA.

The Area Under the Receiver Operating Characteristic (AUC) is an evaluation method for a predictor in a binary classification system.

Experimental Environment. The simulation is carried out on a computer with Windows operating system and 3.6 GHz 4-core 8 threads CPU, 16 GB memory. In this study, we set two WSRC parameters as $\sigma = 1.5$ and $\epsilon = 0.5$, respectively.

Selecting Optimal Parameters of PSSM-DCT, Number of Base Classifiers and Probability Mapping Function. Under the imbalanced learning scenario, over pursuing the overall accuracy is not appropriate and can be deceiving for evaluating the performance of a predictor. Therefore, the MCC provides the overall measurement of the quality of binary prediction. We report the evaluation by choosing the Threshold (T) of probability value, which maximizes the MCC value of prediction. Different value of window size w and first k rows (PSSM-DCT) may lead to different performance. We evaluate value of k from 1 to w (size of sliding window) rows, with a step of 1 row, on GTP data set by independent test validation. We select the optimal value w and k by highest MCC value and find that 17 and 7 are the best parameters of window size w and first k rows, respectively. The result on GTP data set is shown in Figure 3. On the curve, the MCC (w = 17) increases when value increases from 1 to 7. But it slightly declines when size increases from 7 up to 17. The best MCC is 0.622, when window size w is 17 and k is 7 (first 7 rows of PSSM-DCT). So, we select the optimal k as 7 in our study. And the optimal dimension of PSSM-DCT is 7 × 20 = 140. In addition, the classifier is EC-RUS (utilizing 7 WSRC models).

The number of base classifiers will also affect the performance of prediction. Therefore, we evaluate the MCC performance variations of an ensembled classifier on GTP data set by independent test validation. Moreover, the feature include PSSM-DCT and PRSA. The number of base classifiers (m) is gradually varying from 1 to 29, with a step of 2. We select the optimal number of base classifiers by highest MCC value, and find that 19 is the best parameter of m. Figure 4 shows the performance variation curves of MCC. The first maximum MCC value is achieved when m = 19, and no improvement can be observed with larger values of m. Hence, we select 19 as the optimal m in our experiment.

To make a decision of temporary probability mapping function (WSRC) from eqs 11a, 11b and 11c, we test above functions on independent testing set of GTP. In addition, the feature is PSSM-DCT + PRSA. The results of different probability mapping functions on independent testing set of GTP are shown in Figure 5. Obviously, the performance of three different probability mapping functions are almost same. However, the type 1 (0.641) function achieves better performance of MCC than type 2(0.629) and 3 (0.599) under maximizing the value of MCC. Thus, we select eq 11a as the probability mapping function of WSRC.

Performance Analysis. To analyze the significance of different features, we test the features of PSSM, PSSM-DCT and PRSA combined with WSRC and SVM models, respectively. Evaluation is carried out on the GTP data set, which contains training and independent testing sets. We use training set to build the model and test it on the independent testing set. The prediction result is shown in Table 2 and Figure 6.

The AUC of PSSM-DCT + PRSA (WSRC), PSSM-DCT (WSRC), PSSM + PRSA (WSRC) and PSSM (WSRC) are 0.861, 0.848, 0.832 and 0.818, respectively. The performance of PSSM-DCT + PRSA (WSRC) is better than other feature by the WSRC model. Because the DCT algorithm can compress PSSM and remove some noise, the performance of PSSM-DCT is better than that of PSSM. Moreover, the AUC values of PSSM-DCT + PRSA (SVM), PSSM-DCT (SVM), PSSM + PRSA (SVM) and PSSM (SVM) are 0.824, 0.814, 0.811 and 0.808, respectively. Obviously, the WSRC model achieves better performance than SVM. In addition, the PSSM-DCT + PRSA (WSRC) achieves the best performance (0.641) of MCC.

Figure 7 shows the trend (including sensitivity, specificity, accuracy and MCC) on different thresholds of probability.

Table 4. Performance of Proposed Method on Training Sets of 12 Types of Ligands over 5-Fold Cross-Validation

Ligand Type	Model	Threshold	SN (%)	Spec (%)	ACC (%)	MCC	AUC
ATP	TargetS[12]	0.500	48.4	98.2	96.2	0.492	0.887
	EC-RUS[a]	0.500	84.1	84.9	84.9	0.347	0.912
	EC-RUS[a]	0.814	58.6	97.9	96.4	0.537	0.912
ADP	TargetS[12]	0.500	56.1	98.8	97.2	0.591	0.907
	EC-RUS[a]	0.500	87.8	87.7	87.7	0.395	0.939
	EC-RUS[a]	0.852	62.2	98.6	97.3	0.610	0.939
AMP	TargetS[12]	0.500	38.0	98.2	96.0	0.386	0.856
	EC-RUS[a]	0.500	81.5	79.7	79.8	0.263	0.888
	EC-RUS[a]	0.835	46.7	98.3	96.6	0.460	0.888
GDP	TargetS[12]	0.430	63.9	98.7	97.2	0.644	0.908
	EC-RUS[a]	0.500	86.1	89.8	89.7	0.435	0.937
	EC-RUS[a]	0.816	67.2	98.9	97.6	0.676	0.937
GTP	TargetS[12]	0.500	48.0	98.7	96.9	0.506	0.858
	EC-RUS[a]	0.500	79.5	85.7	85.5	0.309	0.896
	EC-RUS[a]	0.842	49.5	99.2	97.6	0.562	0.896
Ca^{2+}	TargetS[12]	0.690	19.2	99.7	98.4	0.320	0.784
	EC-RUS[a]	0.500	73.9	73.8	73.8	0.118	0.812
	EC-RUS[a]	0.861	14.7	99.7	98.6	0.220	0.812
Mg^{2+}	TargetS[12]	0.810	26.4	99.8	99.0	0.383	0.798
	EC-RUS[a]	0.500	73.8	79.4	79.3	0.125	0.839
	EC-RUS[a]	0.864	25.8	99.8	99.1	0.354	0.839
Mn^{2+}	TargetS[12]	0.740	40.8	99.5	98.7	0.445	0.901
	EC-RUS[a]	0.500	83.4	86.6	86.6	0.201	0.921
	EC-RUS[a]	0.841	31.0	99.6	98.9	0.358	0.921
Fe^{3+}	TargetS[12]	0.810	51.8	99.6	98.8	0.592	0.922
	EC-RUS[a]	0.500	87.1	90.1	90.0	0.278	0.940
	EC-RUS[a]	0.809	53.1	99.2	98.6	0.489	0.940
Zn^{2+}	TargetS[12]	0.830	50.0	99.6	98.9	0.557	0.938
	EC-RUS[a]	0.500	88.7	90.8	90.8	0.279	0.958
	EC-RUS[a]	0.860	45.6	99.3	98.7	0.440	0.958
DNA	TargetS[12]	0.490	41.7	94.5	89.9	0.362	0.824
	EC-RUS[a]	0.500	81.9	71.8	72.3	0.259	0.852
	EC-RUS[a]	0.763	48.7	95.1	92.6	0.378	0.852
HEME	TargetS[12]	0.650	50.5	98.3	94.4	0.579	0.887
	EC-RUS[a]	0.500	85.0	83.6	83.7	0.416	0.922
	EC-RUS[a]	0.846	60.3	97.5	95.1	0.591	0.922

[a] The EC-RUS model is built by WSRC and the feature is PSSM-DCT + PRSA.

Though the threshold of probability rises, values of specificity, accuracy and MCC are synchronous rising. The trend of sensitivity are opposite.

To evaluate the performance of ensemble classifier and random under-sampling, we test the single subclassifier (once random sampling), EC-RUS (repeatedly random sampling and ensemble classifier) and single classifier (using whole training set without random sampling) on independent testing set of GTP. The results are listed in Table 3. The max MCC (WSRC) of single subclassifier (once random sampling), EC-RUS (repeatedly random sampling and ensemble classifier) and single classifier (using whole training set without random sampling) are 0.599, 0.641 and 0.621, respectively. Because single classifier is built via whole training set, the performance of single classifier is better than single subclassifier (once random sampling). Obviously, the result of EC-RUS is the best. The EC-RUS is trained by repeatedly random under-sampling and ensemble classifier. The strategy of EC-RUS is useful to improve performance of prediction.

Results on Training Sets. We test the performance of our proposed method via 5-fold cross-validation on 12 training sets. Our method is compared with TargetS,[12] and results on the training sets are listed in Table 4. TargetS chose thresholds under maximizing the value of MCC. The MCC of our method are 0.537, 0.610, 0.460, 0.676 and 0.562 on ATP, ADP, AMP, GDP and GTP, respectively. TargetS achieves MCC values of 0.492, 0.591, 0.386, 0.644 and 0.506, respectively. Obviously, the performance of the proposed method is better than TargetS on nucleotides (ATP, ADP, AMP, GDP and GTP). Although the values of MCC are decreased by 0.1, 0.029, 0.087, 0.103 and 0.117, the AUC values of our proposed method are improved by 0.028, 0.041, 0.020, 0.018, and 0.20 compared with TargetS on metal ions (Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+} and Zn^{2+}), respectively. In addition, our method achieves MCC values of 0.378 (improved by 0.016) on DNA and 0.591 (improved by 0.012) on HEME, respectively. The threshold of TargetS were under maximizing the value of MCC.

Comparison with Existing Predictors on Independent Test Sets. In this section, our proposed method is compared with other existing methods on independent test sets of nucleotides, as shown in Table 5. Existing methods are proposed by Yu et al. (TargetS),[12] Chen et al. (SVMPred, NsitePred)[13,15] and alignment-based baseline predictor, respectively. It can be observed that the best MCC of 0.506

Table 5. Comparison with Existing Predictors on Independent Test Sets of Nucleotides

Ligand Type	Model	SN (%)	Spec (%)	ACC (%)	MCC	AUC
ATP	TargetS[12][c]	50.1	98.3	96.5	0.502	0.898
	NsitePred[15][c]	50.8	97.3	95.5	0.439	–[d]
	SVMPred[13][c]	47.3	96.7	94.9	0.387	0.877
	alignment-based[c]	30.6	97.0	94.5	0.265	–[d]
	EC-RUS(WSRC, T = 0.805)[a]	45.4	98.8	96.8	0.506	0.871
	EC-RUS(SVM, T = 0.841)[b]	44.3	98.2	96.2	0.443	0.876
ADP	TargetS[12][c]	46.9	98.9	97.2	0.507	0.896
	NsitePred[15][c]	46.2	97.6	96.0	0.419	–[d]
	SVMPred[13][c]	46.1	97.2	95.5	0.382	0.875
	alignment-based[c]	31.8	97.4	95.1	0.284	–[d]
	EC-RUS(WSRC, T = 0.811)[a]	44.4	99.2	97.6	0.511	0.872
	EC-RUS(SVM, T = 0.823)[b]	39.6	99.1	97.4	0.459	0.876
AMP	TargetS[12][c]	34.2	98.2	95.9	0.359	0.830
	NsitePred[15][c]	33.9	97.6	95.3	0.321	–[d]
	SVMPred[13][c]	32.1	96.4	94.1	0.255	0.798
	alignment-based[c]	19.6	97.3	94.5	0.178	–[d]
	EC-RUS(WSRC, T = 0.850)[a]	24.9	99.5	97.0	0.393	0.815
	EC-RUS(SVM, T = 0.865)[b]	33.9	98.6	96.3	0.353	0.814
GDP	TargetS[12][c]	56.2	98.1	96.2	0.550	0.896
	NsitePred[15][c]	55.7	97.9	96.1	0.536	–[d]
	SVMPred[13][c]	49.5	97.6	95.4	0.466	0.870
	alignment-based[c]	41.2	97.8	95.3	0.415	–[d]
	EC-RUS(WSRC, T = 0.870)[a]	36.6	99.9	97.1	0.579	0.872
	EC-RUS(SVM, T = 0.839)[b]	50.0	99.1	96.9	0.587	0.897
GTP	TargetS[12][c]	57.3	98.8	96.9	0.617	0.855
	NsitePred[15][c]	58.4	95.7	94.0	0.448	–[d]
	SVMPred[13][c]	48.3	91.7	89.7	0.276	0.821
	alignment-based[c]	52.8	97.9	95.9	0.516	–[d]
	EC-RUS(WSRC, T = 0.781)[a]	61.8	98.7	97.0	0.641	0.861
	EC-RUS(SVM, T = 0.869)[b]	39.3	99.5	96.8	0.546	0.824

[a] The EC-RUS model is built by WSRC (the threshold under maximizing the value of MCC) and the feature is PSSM-DCT + PRSA. [b] The EC-RUS model is built by SVM (the threshold under maximizing the value of MCC) and the feature is PSSM-DCT + PRSA. [c] Results excerpted from ref 12. [d] "–" means not available.

Table 6. Comparison with Existing Predictors on Independent Test Sets of Metal Ions

Ligand Type	Model	SN (%)	Spec (%)	ACC (%)	MCC	AUC
Ca^{2+}	TargetS[12][c]	13.8	99.8	98.8	0.243	0.767
	FunFOLD[18][c]	12.2	99.6	98.1	0.196	–[d]
	CHED[19][c]	18.7	98.2	97.1	0.142	–[d]
	alignment-based[c]	20.3	98.6	97.5	0.175	–[d]
	EC-RUS(WSRC, T = 0.839)[a]	17.3	99.6	98.7	0.225	0.779
	EC-RUS(SVM, T = 0.772)[b]	35.1	95.3	94.6	0.145	0.792
Mg^{2+}	TargetS[12][c]	18.3	99.8	98.8	0.294	0.706
	FunFOLD[18][c]	22.0	99.1	98.3	0.215	–[d]
	CHED[19][c]	14.6	98.3	97.3	0.103	–[d]
	alignment-based[c]	14.1	99.2	98.2	0.147	–[d]
	EC-RUS(WSRC, T = 0.870)[a]	20.1	99.8	99.1	0.317	0.780
	EC-RUS(SVM, T = 0.843)[b]	32.9	98.7	98.1	0.234	0.787
Mn^{2+}	TargetS[12][c]	40.1	99.5	98.7	0.449	0.888
	FunFOLD[18][c]	23.3	99.8	98.7	0.376	–[d]
	CHED[19][c]	35.0	98.1	97.3	0.253	–[d]
	alignment-based[c]	26.6	99.0	98.0	0.257	–[d]
	EC-RUS(WSRC, T = 0.829)[a]	35.8	99.6	98.9	0.403	0.888
	EC-RUS(SVM, T = 0.866)[b]	51.3	97.8	97.3	0.310	0.891
Fe^{3+}	TargetS[12][c]	48.3	99.3	98.7	0.479	0.945
	FunFOLD[18][c]	47.2	99.1	98.4	0.432	–[d]
	CHED[19][c]	49.2	97.0	96.3	0.279	–[d]
	alignment-based[c]	30.0	99.2	98.3	0.300	–[d]
	EC-RUS(WSRC, T = 0.832)[a]	44.3	99.6	99.0	0.490	0.936
	EC-RUS(SVM, T = 0.839)[b]	66.0	97.7	97.3	0.393	0.943
Zn^{2+}	TargetS[12][c]	46.4	99.5	98.7	0.527	0.936
	FunFOLD[18][c]	36.5	99.5	98.6	0.436	–[d]
	CHED[19][c]	37.9	98.0	97.1	0.280	–[d]
	alignment-based[c]	29.7	99.0	98.0	0.297	–[d]
	EC-RUS(WSRC, T = 0.855)[a]	48.9	99.2	98.6	0.437	0.958
	EC-RUS(SVM, T = 0.887)[b]	69.9	97.3	97.0	0.392	0.961

[a] The EC-RUS model is built by WSRC (the threshold under maximizing the value of MCC) and the feature is PSSM-DCT + PRSA. [b] The EC-RUS model is built by SVM (the threshold under maximizing the value of MCC) and the feature is PSSM-DCT + PRSA. [c] Results excerpted from ref 12. [d] "–" means not available.

(ATP), 0.511 (ADP), 0.393 (AMP) and 0.641 (GTP) is obtained from our proposed model (EC-RUS) WSRC)). And the EC-RUS (SVM) obtains the best MCC of 0.587 on GDP. Comparing with TargetS,[12] our method achieves MCC improvement of 0.004 (0.506 over 0.502), 0.004 (0.511 over 0.507), 0.034 (0.393 over 0.359), 0.029 (0.579 over 0.550) and 0.024 (0.641 over 0.617) on independent test sets of ATP, ADP, AMP, GDP and GTP, respectively. The thresholds of TargetS, NsitePred and SVMPred were under maximizing the value of MCC.

We also evaluate our proposed method on independent test sets of metal ions (Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+} and Zn^{2+}), as shown in Table 6. Existing methods are proposed by Yu et al. (TargetS),[12] Roche et al. (FunFOLD),[18] Babor et al. (CHED)[19] and alignment-based baseline predictor, respectively. FunFOLD and CHED are taken as ligand-specific predictors for comparison on the five types of metal ion ligands. Yu et al. retrained the FunFOLD and CHED on the data set of each metal ion ligand. Although the MCC values are decreased by 0.018 (0.243 to 0.225), 0.046 (0.449 over 0.403) and 0.09 (0.527 over 0.437) compared with TargetS on Ca^{2+}, Mn^{2+} and Zn^{2+}, respectively. Obviously, the MCC value of the proposed method is better than those of TargetS on Mg^{2+} (improved by 0.023) and Fe^{3+} (improved by 0.011). The reasons that proposed method along with TargetS do not perform well on Ca^{2+}, Mn^{2+} and Zn^{2+} are as follows: (1) the volume of metal ions are smaller compared to nucleotides, so the number of binding residues are less; (2) the performance of PRSA is not good on small volume ligands; (3) the TargetS constructed ligand-specific model and helped to improve prediction

Table 7. Comparison with Existing Predictors on Independent Test Sets of DNA

Ligand Type	Model	SN (%)	Spec (%)	ACC (%)	MCC	AUC
DNA	TargetS[12c]	41.3	96.5	93.3	0.377	0.836
	MetaDBSite[11c]	58.0	76.4	75.2	0.192	–[d]
	DNABR[20c]	40.7	87.3	84.6	0.185	–[d]
	alignment-based[c]	26.6	94.3	90.5	0.190	–[d]
	EC-RUS(WSRC, *T* = 0.787)[a]	31.5	97.8	95.2	0.319	0.814
	EC-RUS(SVM, *T* = 0.707)[b]	43.7	94.3	92.3	0.287	0.828

[a] The EC-RUS model is built by WSRC (the threshold under maximizing the value of MCC) and the feature is PSSM-DCT + PRSA. [b] The EC-RUS model is built by SVM (the threshold under maximizing the value of MCC) and the feature is PSSM-DCT + PRSA. [c] Results excerpted from ref 12. [d] "–" means not available.

Table 8. Comparison with Existing Predictors on Independent Test Sets of HEME

Ligand Type	Model	SN (%)	Spec (%)	ACC (%)	MCC	AUC
HEME	TargetS(*T* = 0.65)[c] [12]	49.8	99.0	95.9	0.598	0.907
	TargetS(*T* = 0.18)[c] [12]	69.3	90.4	89.1	0.426	0.907
	HemeBind[21c]	86.2	90.7	90.6	0.537	–[d]
	alignment-based[c]	51.4	97.3	94.4	0.507	–[d]
	EC-RUS(WSRC, *T* = 0.500)[a]	83.5	87.5	87.3	0.453	0.935
	EC-RUS(WSRC, *T* = 0.859)[a]	55.8	99.0	96.4	0.640	0.935
	EC-RUS(SVM, *T* = 0.500)[b]	76.8	92.2	91.3	0.508	0.933
	EC-RUS(SVM, *T* = 0.821)[b]	57.6	98.7	96.2	0.632	0.933

[a] The EC-RUS model is built by WSRC and the feature is PSSM-DCT + PRSA. [b] The EC-RUS model is built by SVM and the feature is PSSM-DCT + PRSA. [c] Results excerpted from ref 12. [d] "–" means not available.

Table 9. Statistical Significance of the Differences among the Predictive Performances (MCC) for Three Methods on 12 Independent Test Sets[a]

	P values	Accepted (*h*0/*h*1)[b]
TargetS ∼ our method	0.9470	*h*0
TargetS ∼ alignment-based	0.0032	*h*1
our method ∼ alignment-based	0.0055	*h*1

[a] The *P* values are computed by *t*-test. [b] The null hypothesis (*h*0) is that the means of two samples are equal. The alternative hypothesis (*h*1) is that the means of two samples are significance differences. The significance level (alpha) is 0.05.

performance on metal ions. Moreover, the thresholds of TargetS, FunFOLD and CHED were under maximizing the value of MCC.

On independent test sets of DNA, we compare our method with TargetS,[12] MetaDBSite,[11] DNABR[20] and alignment-based predictor, as shown in Table 7. Their prediction MCC values are 0.377, 0.192, 0.185 and 0.190, respectively. And our proposed method achieves 0.319 of MCC by EC-RUS(WSRC). Although our performance of MCC is lower than that of TargetS, it is better than those of MetaDBSite, DNABR and alignment-based.

Table 10. Evaluation (under the threshold of training sets) on the 12 Independent Test Data Sets

Ligand Type	Threshold	SN (%)	Spec (%)	ACC (%)	MCC	AUC
ATP	0.814	43.4	98.9	96.8	0.497	0.871
ADP	0.852	38.3	99.4	97.6	0.486	0.872
AMP	0.835	25.8	99.4	96.9	0.383	0.815
GDP	0.816	38.7	99.7	97.0	0.559	0.872
GTP	0.842	49.4	99.4	97.1	0.616	0.861
Ca^{2+}	0.861	10.6	99.8	98.8	0.190	0.779
Mg^{2+}	0.864	21.2	99.8	99.1	0.311	0.780
Mn^{2+}	0.841	31.0	99.7	99.0	0.394	0.888
Fe^{3+}	0.809	46.2	99.2	98.6	0.417	0.936
Zn^{2+}	0.860	45.5	99.3	98.6	0.427	0.958
DNA	0.763	33.4	97.4	94.9	0.313	0.814
HEME	0.846	56.7	98.9	96.3	0.638	0.935

[a] The model is built by EC-RUS(WSRC) and the feature is PSSM-DCT + PRSA.

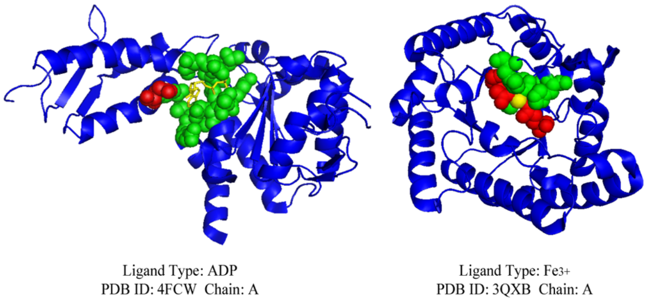


Figure 8. Representative protein–ligand complex: left is 4FCW-A, right is 3QXB-A.

Table 11. Performance of Our Model via 5-Fold Cross-Validation on 198 Drug–Target Pairs Data Set

Ligand Type	Model	SN (%)	Spec (%)	ACC (%)	MCC	AUC
Drug	EC-RUS(WSRC, *T* = 0.500)[a]	81.1	82.1	82.1	0.379	0.889
	EC-RUS(WSRC, *T* = 0.738)[a]	62.5	95.0	92.9	0.504	0.889
	EC-RUS(SVM, *T* = 0.500)[b]	77.2	79.8	79.6	0.331	0.860
	EC-RUS(SVM, *T* = 0.688)[b]	49.8	95.4	92.4	0.423	0.860

[a] The EC-RUS model is built by WSRC and the feature is PSSM-DCT + PRSA. [b] The EC-RUS model is built by SVM and the feature is PSSM-DCT + PRSA.

Table 12. Running Time (seconds) of Subclassifier and EC-RUS(WSRC) on 12 Independent Test Sets

Data Set	SS[a] (WSRC)	EC-RUS(WSRC)	Data Set	SS[a] (WSRC)	EC-RUS(WSRC)
ATP	195.8	3720.2	Ca^{2+}	1112.7	21141.3
ADP	311.4	5916.6	Mg^{2+}	1432.3	27213.7
AMP	74.2	1409.8	Mn^{2+}	156.8	2979.2
GDP	29.9	568.1	Fe^{3+}	42.2	801.8
GTP	13.7	260.3	Zn^{2+}	942.2	17901.8
DNA	316.4	6011.6	HEME	128.6	2443.4

[a] The SS denotes single subclassifier.

On independent test sets of HEME, we compare our method with TargetS,[12] HemeBind[21] and alignment-based predictor, as shown in Table 8. Our method achieves MCC of 0.640 through EC-RUS(WSRC). In addition to PSSM feature, the HemeBind used several other types of 3D structure features such as Relative Accessible Surface Area (RASA),[62] Depth index (DPX)[63] and protrusion index (CX).[64] So, the SN (86.2%) of HemeBind is better than those of other methods.

We use two-sample *t*-test to evaluate the significance differences of MCC performance. The null hypothesis ($h0$) is that the means of two samples (2 sets of MCC values from two methods) are equal. The alternative hypothesis ($h1$) is that the means of two samples are significance differences. If $h0$ is accepted, the differences of MCC is not significant, which means the MCC performance of our method is not improved significantly compared with other methods. If $h1$ is accepted, the improvement of our method is significant. Moreover, the significance level (alpha) is 0.05. The results of *t*-test are listed in Table 9. The difference between TargetS and our method is not significant (P value: 0.9470). Comparing with alignment-based method, our method shows significantly better prediction accuracy (P value: 0.0055). The reason for the difference (between TargetS and our method) is not significant is that TargetS constructed ligand-specific model and helped to improve prediction performance on metal ions. The volume of metal ions are smaller compared to other ligands. So, we will take into account metal ion specificity in our further work. Although the difference is not significant, our proposed method outperforms TargetS (MCC values) in eight of the 12 ligands types.

Tables 5, 6, 7 and 8 list the results of our model and other outstanding methods under maximizing the value of MCC, which is just a method of evaluation. In Yu's study,[12] those methods (comparison) were all under maximizing the value of MCC. In real prediction of binding sites, we could not know the threshold of probability value (independent test data sets) under maximizing the value of MCC. Thus, we could use the threshold of training sets. In Table 10, we report the evaluation by choosing the threshold (T) of training sets. Although the values of MCC go down slightly (e.g., the MCC of ATP is from 0.506 to 0.497), the model is effective on independent test data sets.

Examples of 4FCW-A and 3QXB-A belong to independent test sets of ADP and Fe^{3+}. We use corresponding training sets to build models and predict two examples, respectively. They are shown in Figure 8. The blue object is the protein sequence (contain helix, fold and loop structures), and yellow object is the ligand. The green region is the true prediction and the red region is the false prediction. Our method can predict majority protein−ligand binding sites.

The Performance of Drug Binding Site Prediction. Zhang et al.[65] used multiple computational approaches for pocket prediction. They collected drug-target pairs from DrugPort[66] (http://www.ebi.ac.uk/thornton-srv/databases/drugport/). They selected only one complex structure for every drug-target pair (the single chain with ligands bound), and obtained 217 drug-target pairs and 96 types of drug molecules. CD-HIT program[67] was used to removed the redundancy of protein sequences with 40% similarity threshold (abtained 198 drug-target complexes). We define the drug binding sites with distance between target sites and drug molecules (any heavy atom) less than 6 Å. We test our method on this data set via 5-fold cross-validation. The results of our model are listed in Table 11. The EC-RUS(WSRC) achieves better performance of MCC (0.504).

Running Time. The running time (on 12 independent test sets) of our model (WSRC) depends on the sizes of training set and test set. To identify any test sample, we first calculate the Gaussian distances between test and each training sample. Moreover, the l_1-minimization problem of linear sparse coefficients also need to be solved. So, the running time of WSRC is time-consuming (comparing with SVM). However, the performance of prediction is better than SVM on most ligand data sets. The results of running time are shown in Table 12. The items of table include running time of single subclassifier and EC-RUS (19 subclassifiers). The EC-RUS is carried out by serial program. In the future, EC-RUS will be parallelization for reducing the running time.

■ DISCUSSION

Lots of computational methods have been developed for detecting protein−ligand binding sites, but the effectiveness and efficiency of previous methods could be improved. Many methods did not take into account information compression, therefore we propose PSSM-DCT feature and EC-RUS-(WSRC) model. Although the performance of our proposed method is lower than that of TargetS[12] on independent test sets of Ca^{2+}, Mn^{2+}, Zn^{2+} and DNA. The PSSM-DCT + PRSA feature combined with EC-RUS (WSRC) model achieves the best performance of MCC on independent test sets of nucleotides (ATP, ADP, AMP, GDP and GTP), Mg^{2+}, Fe^{3+} and HEME. The main reason is that TargetS constructed a ligand-specific model to improve accuracy. We plan to develop more effective ligand-specific features or employ the better machine learning model (such as gradient boosting decision tree or deep learning) for further improving protein−ligand binding prediction performances in our future work.

■ CONCLUSIONS

In this paper, we employ a novel method to detect protein−ligand binding sites using sequences of proteins. Our model is constructed via EC-RUS(WSRC) model and ensemble feature representation scheme (PSSM-DCT and PRSA). For the evaluation, our method is tested on 12 different types of ligands data sets. The result shows that our model obtains the best MCC on independent test sets of ATP (0.506), ADP (0.511), AMP (0.393), GDP (0.579), GTP (0.641), Mg^{2+} (0.317), Fe^{3+} (0.490) and HEME (0.640). Although our performance of MCC is lower than that of TargetS, it is better than those of other methods on Ca^{2+} (0.225), Mn^{2+} (0.403), Zn^{2+} (0.437) and DNA (0.319). Compared with these state-of-the-art methods, our method is able to achieve comparable or even better prediction results on the data sets. Test results indicate that our proposed method is useful for detecting protein−ligand binding sites, thus it may reduce the cost of biological experiments.

■ AUTHOR INFORMATION

Corresponding Author

*F. Guo. E-mail: fguo@tju.edu.cn.

ORCID

Fei Guo: 0000-0003-2911-7643

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is supported by a grant from the National Science Foundation of China (NSFC 61772362 and 61402326), Peiyang Scholar Program of Tianjin University (no. 2016XRG-0009), and the Tianjin Research Program of Application Foundation and Advanced Technology (16JCQNJC00200). The authors also thank Professor Dong-jun Yu for kindly providing the data sets on his website.

■ REFERENCES

(1) 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* (D1), 345−356.

(2) Yang, J.; Roy, A.; Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **2013**, *41* (D1), 1096−1103.

(3) Madala, P. K.; Fairlie, D. P.; Bodén, M. Matching cavities in g protein-coupled receptors to infer ligand-binding sites. *J. Chem. Inf. Model.* **2012**, *52* (5), 1401−1410.

(4) Desaphy, J.; Azdimousa, K.; Kellenberger, E.; Rognan, D. Comparison and Druggability Prediction of Protein-Ligand Binding Sites from Pharmacophore-Annotated Cavity Shapes. *J. Chem. Inf. Model.* **2012**, *52* (8), 2287−2299.

(5) Weill, N.; Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein-Ligand Binding Sites. *J. Chem. Inf. Model.* **2010**, *50* (1), 123−135.

(6) Komiyama, Y.; Banno, M.; Ueki, K.; Saad, G.; Shimizu, K. Automatic generation of bioinformatics tools for predicting protein-ligand binding sites. *Bioinformatics* **2016**, *32* (6), 901−907.

(7) Guo, F.; Li, S. C.; Wei, Z. X.; Zhu, D. M.; Shen, C.; Wang, L. S. Structural Neighboring Property for Identifying Protein-Protein Binding Sites. BMC Sys. *BMC Syst. Biol.* **2015**, *9* (Suppl 5), S3.

(8) Li, Z.; Zhao, Y. L.; Pan, G. F.; Tang, J. J.; Guo, F. A Novel Peptide Binding Prediction Approach for HLA-DR Molecule Based on Sequence and Structural Information. *BioMed Res. Int.* **2016**, *2016*, 1−10.

(9) Ding, Y. J.; Tang, J. J.; Guo, F. Identification of Residue-Residue Contacts Using a Novel Coevolution-Based Method. *Curr. Proteomics* **2016**, *13* (2), 122−129.

(10) Yang, J.; Roy, A.; Zhang, Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, *29* (20), 2588−2595.

(11) Si, J.; Zhang, Z.; Lin, B.; Schroeder, M.; Huang, B. MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. BMC Sys. *BMC Syst. Biol.* **2011**, *5* (1), S7.

(12) Yu, D. J.; Hu, J.; Yang, J.; Shen, H. B.; Tang, J. H.; Yang, J. Y. Designing Template-Free Predictor for Targeting Protein-Ligand Binding Sites with Classifier Ensemble and Spatial Clustering. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2013**, *10* (4), 994−1008.

(13) Chen, K.; Mizianty, M. J.; Kurgan, L. ATPsite: sequence-based prediction of ATP-binding residues. *Proteome Sci.* **2011**, *9* (Suppl 1), S4.

(14) Pupko, T.; Bell, R. E.; Mayrose, I.; Glaser, F.; Bental, N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **2002**, *18* (1), S71.

(15) Chen, K.; Mizianty, M. J.; Kurgan, L. Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. *Bioinformatics* **2012**, *28* (3), 331−341.

(16) 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* (11), 974−985.

(17) Yu, D. J.; Hu, J.; Tang, Z. M.; Shen, H. B.; Yang, J.; Yang, J. Y. Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling. *Neurocomputing* **2013**, *104*, 180−190.

(18) 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 Bioinf.* **2011**, *12* (1), 160−189.

(19) Babor, M.; Gerzon, S.; Raveh, B.; Sobolev, V.; Edelman, M. Prediction of transition metal-binding sites from apo protein structures. *Proteins: Struct., Funct., Genet.* **2008**, *70* (1), 208.

(20) 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. Bioinf.* **2012**, *9* (6), 1766−1775.

(21) Liu, R.; Hu, J. J. HemeBIND: a novel method for heme binding residue prediction by combining structural and sequence information. *BMC Bioinf.* **2011**, *12* (1), 207.

(22) Liu, R.; Hu, J. J. Computational prediction of heme-binding residues by exploiting residue interaction network. *PLoS One* **2011**, *6* (10), e25560.

(23) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics Modell.* **1997**, *15* (6), 359−363.

(24) 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.

(25) Laskowski, R. A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graphics* **1995**, *13* (5), 323−330.

(26) Levitt, D. G.; Banaszak, L. J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graphics* **1992**, *10* (4), 229−234.

(27) Le Guilloux, V. L.; Schmidtke, P.; Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10* (1), 168.

(28) Laurie, A. T.; Jackson, R. M. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **2005**, *21* (9), 1908−1916.

(29) Hernandez, M.; Ghersi, D.; Sanchez, R. SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res.* **2009**, *37* (Suppl 2), W413−W416.

(30) Hoffmann, B.; Zaslavskiy, M.; Vert, J. P.; Stoven, V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinf.* **2010**, *11* (1), 99.

(31) Armon, A.; Graur, D.; BenTal, N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **2001**, *307* (1), 447−463.

(32) Huang, B.; Schroeder, M. LIGSITE csc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6* (1), 19.

(33) 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* (12), e1000585.

(34) Glaser, F.; Morris, R. J.; Najmanovich, R. J.; Laskowski, R. A.; Thornton, J. M. A method for localizing ligand binding pockets in protein structures. *Proteins: Struct., Funct., Genet.* **2006**, *62* (2), 479−488.

(35) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20* (3), 273−297.

(36) Ofran, Y.; Mysore, V.; Rost, B. Prediction of dna-binding residues from sequence. *Bioinformatics* **2007**, *23* (13), i347−i353.

(37) Yan, C.; Terribilini, M.; Wu, F. H.; Jernigan, R. L.; Dobbs, D.; Honavar, V. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinf.* **2006**, *7* (1), 262.

(38) 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* (Suppl 2), W243−W248.

(39) Wang, L.; Yang, M. Q.; Yang, J. Y. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics* **2009**, *10* (Suppl 1), S1.

(40) Hwang, S.; Gou, Z. K.; Kuznetsov, I. B. DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* **2007**, *23* (5), 634−636.

(41) 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* (4), 477−486.

(42) Nanni, L.; Brahnam, S.; Lumini, A. Wavelet images and Chou's pseudo amino acid composition for protein classification. *Amino Acids* **2012**, *43* (2), 657−665.

(43) Nanni, L.; Brahnam, S.; Lumini, A. High performance set of PseAAC and sequence based descriptors for protein classification. *J. Theor. Biol.* **2010**, *266* (1), 1−10.

(44) Ahmed, N.; Natarajan, T.; Rao, K. R. Discrete cosine transform. *IEEE Trans. Comput.* **1974**, *C-23* (1), 90−93.

(45) Nanni, L.; Lumini, A.; Brahnam, S. An empirical study of different approaches for protein classification. *Sci. World J.* **2014**, *2014*, 236717.

(46) Ding, Y. J.; Tang, J. 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* (10), 1623−1636.

(47) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J. H.; 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* (17), 3389−3402.

(48) Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *BMC Bioinf.* **2009**, *10* (1), 421−429.

(49) Ahmad, S.; Gromiha, M. M.; Sarai, A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins: Struct., Funct., Genet.* **2003**, *50* (4), 629−635.

(50) Joo, K.; Lee, S. J.; Lee, J. Sann: solvent accessibility prediction of proteins by nearest neighbor method. *Proteins: Struct., Funct., Genet.* **2012**, *80* (7), 1791−1797.

(51) Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; Ma, Y. Robust Face Recognition via Sparse Representation. *IEEE T. Pattern Anal.* **2009**, *31* (2), 210−227.

(52) Wright, J.; Ganesh, A.; Zhou, Z.; Wagner, A.; Ma, Y. Demo: Robust face recognition via sparse representation. *IEEE Int. Conf. Automat. Face Gesture Recogn.* **2008**, *31* (2), 1−2.

(53) Liao, B.; Jiang, Y.; Yuan, G.; Zhu, W.; Cai, L. J.; Cao, Z. Learning a weighted meta-sample based parameter free sparse representation classification for microarray data. *PLoS One* **2014**, *9* (8), e104314.

(54) Huang, Y. A.; You, Z. H.; Chen, X.; Chan, K.; Luo, X. Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. *BMC Bioinf.* **2016**, *17* (1), 184−194.

(55) Huang, Y. A.; You, Z. H.; Gao, X.; Wong, L.; Wang, L. Using Weighted Sparse Representation Model Combined with Discrete Cosine Transformation to Predict Protein-Protein Interactions from Protein Sequence. *BioMed Res. Int.* **2015**, *2015*, e902198.

(56) Lu, C. Y.; Min, H.; Gui, J.; Zhu, L.; Lei, Y. K. Face recognition via Weighted Sparse Representation. *J. Vis. Commun. Image R.* **2013**, *24* (2), 111−116.

(57) Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7* (1), 1−26.

(58) Tao, D.; Tang, X.; Li, X.; Wu, X. Asymmetric Bagging and Random Subspace for Support Vector Machines-Based Relevance Feedback in Image Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28* (7), 1088−1099.

(59) Dessailly, B. H.; Lensink, M. F.; Orengo, C. A.; Wodak, S. J. LigASite-a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.* **2008**, *36* (Suppl 1), D667−D673.

(60) Lopez, G.; Valencia, A.; Tress, M. FireDB-a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.* **2007**, *35* (Suppl 1), D219−D223.

(61) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47* (12), 2977−2980.

(62) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22* (12), 2577−2637.

(63) Pintar, A.; Carugo, O.; Pongor, S. DPX: for the analysis of the protein core. *Bioinformatics* **2003**, *19* (2), 313−314.

(64) Pintar, A.; Carugo, O.; Pongor, S. CX, an algorithm that identifies protruding atoms in proteins. *Bioinformatics* **2002**, *18* (7), 980−984.

(65) Zhang, Z.; Li, Y.; Lin, B.; Schroeder, M.; Huang, B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* **2011**, *27* (15), 2083−2088.

(66) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36* (Suppl 1), D901−D906.

(67) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22* (13), 1658−1659.

■ NOTE ADDED AFTER ASAP PUBLICATION

The definition of matrix **W** below eq 10c has been updated in the version published ASAP on November 21, 2017. The corrected version was published ASAP on November 22, 2017.