Analytical Biochemistry 604 (2020) 113799

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Identification of ligand-binding residues using protein sequence profile alignment and query-specific support vector machine model

Jun Hu[a,b,*], Liang Rao[a], Xueqiang Fan[a], Guijun Zhang[a,*]

[a] College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China
[b] Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Zhangzhou, 363000, China

ARTICLE INFO

Keywords:
Protein-ligand binding residue
Protein sequence profile alignment
Query-specific
Ligand-specific
SVM

ABSTRACT

Information embedded in ligand-binding residues (LBRs) of proteins is important for understanding protein functions. How to accurately identify the potential ligand-binding residues is still a challenging problem, especially only protein sequence is given. In this paper, we establish a new query-specific computational method, named I-LBR, for the identification of LBRs without directly using the information of protein 3D structure. I-LBR includes two modes, named as I-LBR^{GP} and I-LBR^{LS}, for the general-purpose and ligand-specific LBR identification. For both modes, I-LBR first construct the specific training subset based on the query sequence information; then use support vector machine (SVM) algorithm to learn the LBR identification model; finally, predict the probability of each residue in query protein belongs to the class of LBR. Experimental results on four testing dataset demonstrate that I-LBR^{LS} is the better choice against I-LBR^{GP}, when the ligand type/types of the query protein binds is/are known. Comparing to other state-of-the-art LBR identification methods, I-LBR can achieve a better or comparable performance. The web-server of I-LBR and dataset used in this study are freely available for academic use at https://jun-csbio.github.io/I-LBR.

1. Introduction

Protein in cells constantly interact to several small-molecule ligands. Statistics has shown that more than 58% of the 515,776 protein chains solved in the Protein Data Bank (PDB) [1] interact one small-molecule ligand at least. These interactions between proteins and ligands can give rise to or modulate various aspects of protein functions. Previous researches [2,3] have shown that, in order to bind one ligand, a small part of surface residues of protein always comprise one local concave-shaped structure known as a 'pocket'. Identifying the ligand-binding residues (LBRs) of proteins is thus a crucial step in elucidating their functions and designing new drugs to regulate these function.

A variety of computational methods have been proposed for the identification of LBRs of proteins. Depending on the type of the input protein data, these methods can be generally categorized into structure-based and sequence-based methods. The structure-based methods is the main direction of the research fields of LBRs identification in the early stage. Generally, the structure-based methods can be further subdivided into two classes of approaches. In the first class of approaches, e.g., ConCavity [4], LIGSITE [5], and SURFNET [6], the ligand-binding pocket consists of LBRs is located by recognizing the concave-shaped structure on the surface of the 3D structure of the query protein [7]. The advantage of this class of approaches is that none of templates is required, but the false positive rate should be high, especially for the low-resolution models generated by protein structure predictors, i.e., MODELLER [8], Rosetta [9], and I-TASSER [10]. The second class of approaches, such as FINDSITE [11], FunFOLD [12], 3DLigandSite [13], COFACTOR [14], COACH [15], ATPbind [16], and DELIA [17], is to infer ligand-binding information from the know template proteins, which have similar local and/or global structure to the query protein. On average, the second class of approaches can achieve a good performance of LBRs identification, especially for the query proteins which have close homologous in the protein-ligand complex structure databases. However, the good performance of the second class cannot be maintained on these query proteins which have only distant-homologous and/or no-homologous templates on the databases.

Unlike the structure-based methods, the sequence-based methods do not require the 3D structure information of query protein [18]. Due to the functions of LBRs are important in the biological processes, the LBRs should be conserved in the evolutionary process [19]. Based on the conservation property of LBR, the sequence-based methods can achieve a comparable performance against the structure-based methods, especially using the machine-learning algorithms, such as support vector machine (SVM) [20]. To name a few, ConSurf [21],

* Corresponding authors. College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, China.
E-mail addresses: hujunum@zjut.edu.cn (J. Hu), zgj@zjut.edu.cn (G. Zhang).

https://doi.org/10.1016/j.ab.2020.113799
Received 1 May 2020; Received in revised form 23 May 2020; Accepted 26 May 2020
Available online 02 July 2020
0003-2697/

Rate4Site [22], NsitePred [23], S-SITE [15], and TargetS [18]. Although the performance of the sequence-based methods is not superior to that of the structure-based methods in general, they can give a more reasonable LBRs for these query proteins which only have distant-homologous or no-homologous template proteins in the existing protein-ligand binding databases.

In addition, depending on whether the ligand type is cared or not, these existing LBR identification methods can be also grouped into general-purpose and ligand-specific methods [16]. The general-purpose methods, which predict LBRs regardless of the ligand types, dominated the field of LBRs identification. Most of the structure-based methods are the general-purpose methods, such as ConCavity [4], LIGSITE [5], and SURFNET [6], FINDSITE [11], FunFOLD [12], 3DLigandSite [13], COFACTOR [14], and COACH [15]. Some sequence-based methods are also the general-purpose methods, including ConSurf [21] and Rate4-Site [22]. However, different ligands tend to bind diverse types of residues with observable specificities due to the specific roles, sizes, and distributions of protein-ligand interactions [24]. Hence, developing ligand-specific LBRs identification methods has attracted considerable attention to gain much more accurate performance. Many ligand-specific methods have emerged recently, such as HemeBind [25] is designed for identifying HEME-specific LBRs, FINDSITE-metal [26] is extended from FINDSITE [11] to identify metal ion-specific LBRs, AT-Pint [27], ATPsite [28], and ATPbind [16] are developed for predicting the ATP-specific LBRs, NsitePred [23] and TargetS [18] are proposed to locate the nucleotides-specific LBRs, and MetaDBSite [29], TargetDNA [30], and DNAPred [31] are designed to predict DNA-specific LBRs.

Despite the progress made in the identification of LBRs, the performance of the existing methods has, on average, not yet lived up to expectations, especially for the sequence-based methods. Furthermore, such methods which can both implement the general-purpose and ligand-specific LBRs identification are rare. In this study, to improve the performance of the sequence-based methods, we report a new query-specific computational method, called I-LBR, to identify the LBRs from the sequence information of the query protein. The proposed I-LBR contains two modes for the identification of LBRs, i.e., general-purpose and ligand-specific modes, named as I-LBR^{GP} and I-LBR^{LS}. For both modes of I-LBR, a template database, called TeD, is first pre-constructed for all proteins in the BioLiP database [32] by collecting the residues associated with the known ligands; secondly, three sequence-related profiles, i.e., position-specific frequency matrix (PSFM), predicted secondary structure probability matrix (PSSPM), and predicted solvent accessibility probability matrix (PSAPM), of query and template proteins are generated; thirdly, based on these profiles, I-LBR generates the query-specific training subset; fourthly, the SVM algorithm is employed to train the query-specific identification model; finally, I-LBR can easily obtain the probability of each residue in the query protein belongs to the class of LBR. Experimental results on four testing datasets demonstrate the effectiveness of the proposed I-LBR. The web-server of I-LBR and dataset used in this study are freely available for academic use at https://jun-csbio.github.io/I-LBR.

2. Materials and methods

To make the proposed I-LBR to be a useful LBR identification method, we basically needs to follow "Chou's 5-steps rule" [33–36] to go through the following five steps (see, e.g. Ref. [37],: (1) select or construct a valid benchmark dataset to train and test the predictor; (2) represent the samples with an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted; (3) introduce or develop a powerful algorithm to conduct the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy; (5) establish a user-friendly web-server for the predictor that is accessible to the public.

2.1. Benchmark datasets

In this study, one general-purpose dataset, which is collected by Yang et al. [15] and consists of 400 non-redundant ligand-binding proteins, is employed to tune the parameters of our proposed method (i.e., I-LBR) and named as TRAIN. Another one general-purpose dataset (named as TEST) and three ligand-specific datasets (named as ATP-TEST, GTP-TEST, and GDP-TEST) are utilized to testify the effectiveness of general-purpose and ligand-specific binding residue identification of the proposed I-LBR. The TEST dataset, which is used in COACH [15], consists of 500 non-redundant proteins that harbor 814 ligands (410 natural ligands, 238 drug-like ligands and 164 metal ions). Noted that none of proteins in TEST has a sequence identity > 30% to the proteins in TRAIN [15]. The ATP-TEST includes 41 ATP-binding proteins, which is derived from our previous work ATPbind [16] directly. In ATP-TEST, only one protein (i.e., 5BURA) has a sequence identity > 30% (i.e., 31.6%, calculated by NW-align which is a sequence alignment tool and available at https://zhanglab.ccmb.med.umich.edu/NW-align) to the proteins in TRAIN. The GDP-TEST and GTP-TEST contain 14 GDP-binding and 7 GTP-binding proteins, respectively, which are collected by Yu et al. [18]. None of proteins in GTP-TEST has a sequence identity > 30% to the proteins in TRAIN. In GDP-TEST, only one protein (3SEAA) has a sequence identity > 30% (i.e., 36.8%) to the proteins in TRAIN. Detailed information of the five datasets is shown in Table 1.

2.2. I-LBR

I-LBR is designed to identify query-specific ligand-binding residues from protein sequence information, which contains five main steps (1) template (i.e., protein with known LBRs): database preparation, (2) sequence profile generation, (3) query-specific training subset construction, (4) query-specific SVM-based model generation, and (5) ligand-binding residues prediction. In addition, I-LBR also proposes two modes (i.e., general-purpose and ligand-specific, named as I-LBR^{GP} and I-LBR^{LS}) to detect ligand-binding residues of proteins. The main difference between I-LBR^{GP} and I-LBR^{LS} is how to construct query-specific training subset (see details in the last two paragraphs of Section 2.2.3). Fig. 1 shows the flowchart of I-LBR. More details is described in the following paragraphs.

2.2.1. Template database preparation

In this study, the template database (called TeD) is pre-calculated for all proteins in the BioLiP database by collecting the residues associated with the known ligands. Concretely, for each template protein, we first count the number of ligands it binds. For each bond ligand of each template protein, we create one database record, which includes the information of the template protein sequence, ligand type, and LBRs. That is, when one template protein binds n ligands, there are n database records corresponding to this template protein.

2.2.2. Sequence profile generation

Starting from a query protein sequence with L residues, I-LBR first

Table 1
Details of five protein-ligand binding residue datasets used in this study.

Dataset	$N_{protein}$[a]	N_{posi}[b]	N_{nega}[c]	N/P ratio[d]
TRAIN	400	8045	122369	15.21
TEST	500	7687	135667	17.65
ATP-TEST	41	674	14159	21.01
GTP-TEST	7	89	1868	21.00
GDP-TEST	14	194	4180	21.55

[a] $N_{protein}$: Number of proteins in the corresponding dataset.
[b] N_{posi}: Number of LBRs in the corresponding dataset.
[c] N_{nega}: Number of non-LBRs in the corresponding dataset.
[d] N/P ratio: Ratio = N_{nega}/N_{posi}.

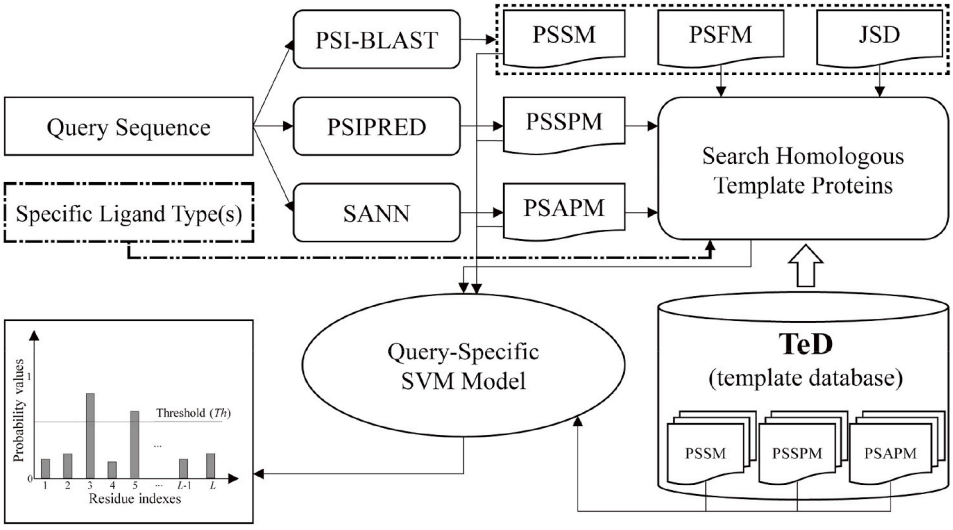

Fig. 1. Flowchart of I-LBR for protein-ligand binding residue identification. The "Specific Ligand Type (s)" is an optional input information, which is only required by the ligand-specific mode of I-LBR, i.e., I-LBR^{LS}.

generates three different profiles: (1) position-specific frequency matrix (PSFM), (2) predicted secondary structure probability matrix (PSSPM), and (3) predicted solvent accessibility probability matrix (PSAPM). To obtain PSFM of query protein, a matrix with size $L\times20$, PSI-BLAST [38] is first employed to thread the query sequence through the non-redundant NCBI sequence database [39] for constructing multiple sequence alignments. The PSFM is then calculated from the multiple sequence alignments (MSAs). To gain the PSSPM of query protein, a probability matrix with size $L\times3$, PSIPRED [40] is used in this study with its default settings. In PSSPM, each line includes the probabilities of three general secondary structure classes, i.e., coil (C), helix (H), and strand (E), of the corresponding residue. To achieve the PSAPM of query protein, SANN developed by Joo et al. [41] is utilized. The PSAPM, which is a probability matrix with size $L\times3$, contains the probabilities of three solvent accessibility classes, i.e., buried, intermediate, and expose, of each residue. Similarly, each template in TeD contains three profiles, i.e., position-specific scoring matrix (PSSM), PSSPM, and PSAPM. The PSSM, PSSPM, and PSAPM are also obtained via PSI-BLAST [38], PSIPRED [40], and SANN [41], respectively.

2.2.3. Query-specific training subset construction

In order to construct the query-specific training subset for each query protein, we want to search several homologous template proteins from TeD. To detect homologous templates, inspired by S-SITE [15], the query profiles, i.e., $PSFM^Q$, $PSSPM^Q$, and $PSAPM^Q$, are compared with the corresponding profiles, i.e., $PSSM^T$, $PSSPM^T$, and $PSAPM^T$, of each template in TeD using the Needleman-Wunsch dynamic programming algorithm [42]. For the h-th database record of each template, the score for aligning the i-th residue in the query to the j-th residue in the template is calculated as

$$S_{i,j} = \sum_{k=1}^{20} PSFM_{i,k}^{Q} PSSM_{j,k}^{T} + 1.8\sum_{k=1}^{3} PSSPM_{i,k}^{Q} PSSPM_{j,k}^{T} + 1.8\sum_{k=1}^{3} PSAPM_{i,k}^{Q} PSAPM_{j,k}^{T} + 2b_j^{T^h} B(Q_i, T_j) \quad (1)$$

where $PSFM_{i,k}^{Q}$, $PSSPM_{i,k}^{Q}$, and $PSAPM_{i,k}^{Q}$ are the i, k-th elements in the $PSFM^Q$, $PSSPM^Q$, and $PSAPM^Q$, respectively; $PSSM_{j,k}^{T}$, $PSSPM_{j,k}^{T}$, and $PSAPM_{j,k}^{T}$ are the j, k-th elements in the $PSFM^T$, $PSSPM^T$, and $PSAPM^T$, respectively; $b_j^{T^h} = 1$ if the j-th residue of template is a LBR of the h-th database record of this template, otherwise $b_j^{T^h} = 0$; the values of 1.8, 1.8, and 2.0 are empirically tuned on the TRAIN dataset; $B(Q_i, T_j)$ is the normalized BLOSUM62 [43] similarity score between the i-th residue of query protein and the j-th residue of template protein. The value of $B(Q_i, T_j)$ is calculated as

$$B(Q_i, T_j) = \frac{BM(Q_i, T_j) - BM_{min}}{BM_{max} - BM_{min}} \quad (2)$$

where $BM(Q_i, T_j)$ is the original value of the BLOSUM62 matrix [43] corresponding to residue types of $Q_i$and T_j. BM_{max} and BM_{min} are the maximum and minimum values of the elements of the BLOSUM62 matrix, respectively. Overall, in Eq. (1), the first term accounts for evolutionary conservation positions alignments, the second for the secondary structure match, the third for the solvent accessibility match, and the last term for evolutionary relation of residues in the LBRs concerning the h-th database record of the template.

The match quality between the query protein and the h-th database record of each template is calculated by

$$mq = \frac{2}{1 + e^{-(0.4A_S^G + L_c(0.1A_S^L + 0.2JSD))}} - 1 \quad (3)$$

where $A_S^G = \sum_{i=1}^{L_{ali}} S_{ali_i^Q, ali_i^T}/L$ and $A_S^L = \sum_{i=1}^{L_{ali}} b_{ali_i^T}^{T^h} S_{ali_i^Q, ali_i^T} / \sum_{i=1}^{L_{ali}} b_{ali_i^T}^{T^h}$ are the global and local alignment scores separately normalized by the query sequence length (L) and the number of the aligned residue pairs associated with the LBRs of the h-th database record of the template; L_{ali} is the number of the aligned residue pairs; ali_i^Q and ali_i^T means the indexes of two residues, which form the i-th aligned pair, in the query and template proteins; $b_{ali_i^T}^{T^h}$ is similar to that defined in Eq. (1); L_c is the fraction of the LBRs of the h-th database record of the template that are aligned to query sequence; the values of 0.4, 0.1, and 0.2 are also tuned on the TRAIN dataset; JSD is an evolutionary conservation index defined as the average Jensen–Shannon divergence score over the query residues aligned the LBRs of the h-th database record of the template, which is calculated from multiple sequence alignments [15]. Specifically, JSD is calculated as

$$JSD = \sum_{i=1}^{L_{ali}} b_{ali_i^T}^{T^h} JSD_{ali_i^Q} / \sum_{i=1}^{L_{ali}} b_{ali_i^T}^{T^h} \quad (4)$$

where $JSD_{ali_i^Q}$ represents the evolutionary conservation score of the ali_i^Q-th residue of the query protein. Here, $JSD_{ali_i^Q}$ is calculated based on multiple sequence alignment generated on the step of "sequence profile generation". The equation of calculating $JSD_{ali_i^Q}$ is shown as

$$JSD_{ali_i^Q} = \frac{1}{2}\sum_{aa\in AA} p_{ali_i^Q}(aa)\log\frac{p_{ali_i^Q}(aa)}{c_{ali_i^Q}(aa)} + \frac{1}{2}\sum_{a\in AA} q(aa)\log\frac{q(aa)}{c_{ali_i^Q}(aa)} \quad (5)$$

where AA means the set of 20 common residue types, $p_{ali_i^Q}(aa)$ is the occurring frequency of the residue type of aa in the i-th column of the multiple sequence alignment, $q(aa)$ is the occurring frequency of the aa residue type estimated on a large set of random sequences, $c_{ali_i^Q}(aa) = (p_{ali_i^Q}(aa) + q(aa))/2$. After the mq score of this record is calculated, if the ligand type of this record is cared by user, the ligand-specific mode of I-LBR, i.e., I-LBR^{LS}, will increment the value of mq by one, but the general-purpose mode of I-LBR will not. Here, all ligand types are cared by the general-purpose mode of I-LBR, i.e., I-LBR^{GP}.

All database records of the template proteins in TeD with a mq score above a threshold (T_{mq}) are chosen to construct the query-specific training subset. Note that, the value of T_{mq} is 0.5 in I-LBR^{GP} and 1.5 in I-LBR^{LS}. If the number of the database records in the constructed training subset is less than 20, the top 20 database records with the highest mq score will be selected to construct the training subset. If the number of the database records in the constructed training subset is more than 200, the top 200 database records with the highest mq score will be selected to construct the training subset and the others will be removed. It is noted that, in order to fairly evaluate the efficiency of I-LBR, none of the proteins of all selected database records has a sequence identity > 30% to the query protein in this study.

2.2.4. Query-specific SVM-Based model generation

Based on the constructed query-specific training subset above, we use the support vector machine (SVM) algorithm [44] to train the computational model for identifying the ligand-binding residues (LBRs) and non-ligand-binding residues (non-LBRs) of the query protein. Due to there may exist two or more database records in the training subset come from the same template protein, we first integrate the label information of all database records with the same protein to one protein and make sure that each residue of each protein has only one label. For each protein in the training subset, each element x of its PSSM is normalized with the logistic function $f(x) = 1/(1 + e^{-x})$. Then, a sliding window with size W is employed to extract the feature vector of each residue based on the PSSM, PSSPM, and PSAPM of the corresponding protein. More specifically, the feature vector of a residue is obtained by concatenating the PSSM, PSSPM, and PSAPM scores of its neighboring residues within the window centered at the residue. In this study we set $W = 17$ based on our previous researches [45]. Therefore, the dimension number of the feature vector of each residue is $(20 + 3 + 3) \times 17 = 442$. Finally, we obtain a query-specific training sample set S_{qst}, which consists of all LBRs and non-LBRs in the training subset.

The identification of LBRs is a typical class imbalance learning problem. As shown in Table 1, we can easily find that a severe class imbalance phenomenon does exist among all datasets: the ratio of the number of non-LBRs to that of LBRs is consistently larger than 15. Our previous study [45] shows that directly using SVM to train the identification model does not yield satisfactory performance of LBR identification. In this study, we combine the random under-sampling (RUS) and modified random over-sampling (MROS) methods to make the number of LBR and non-LBR data to be balance.

Let $S_{qst} = S_{posi} \cup S_{nega}$, S_{posi} be the positive sample (i.e., LBR) subset, S_{nega} be the negative sample (i.e., non-LBR) subset. We hope obtain a new sample set $\hat{S}_{qst} = \hat{S}_{posi} \cup \hat{S}_{nega}$, where the sample number (i.e., $|\hat{S}_{posi}|$) of $\hat{S}_{posi}$ is equal to that (i.e., $|\hat{S}_{nega}|$) of $\hat{S}_{nega}$. In order to achieve this purpose, we employ RUS to randomly remove some negative samples to generate $\hat{S}_{nega}$ and make sure that $|\hat{S}_{nega}|/|S_{posi}| = \alpha$ ($\alpha \geq 1$). Hence, the MROS methods should be used to make the sample number of S_{posi} to be $\alpha \cdot |S_{posi}|$. Due to directly copying the positive samples is not the best way absolutely, the MROS method is proposed to relieve the negative effect of the imbalance phenomenon. The MROS contains two steps which is described as follows:

Step I: Three samples, denoted as $\mathbf{x}_i$, $\mathbf{x}_j$, and $\mathbf{x}_k$, are randomly selected from S_{posi}.

Step II: According to the three randomly selected samples, an additional positive sample can be synthesized:

$$\mathbf{x}_{new} \leftarrow \mathbf{x}_i + \lambda \cdot (\mathbf{x}_j - \mathbf{x}_k) \tag{6}$$

where λ is a random value ranging from 0 to 0.3. If $|\hat{S}_{posi}| < \alpha \cdot |S_{posi}|$, $\hat{S}_{posi} \leftarrow \hat{S}_{posi} \cup \{\mathbf{x}_{new}\}$.

Steps I and II are repeated until $|\hat{S}_{posi}| = \alpha \cdot |S_{posi}|$. In this study, we have tested different values of α and found that $\alpha = 4$ is a better choice (see details in Section 3.1).

After the balanced training sample set $\hat{S}_{qst}$ obtained, we employ LIBSVM software (version libsvm-3.18) [46], which is freely downloadable package from http://www.csie.ntu.edu.tw/~cjlin/libsvm/, to implement the SVM algorithm to train the query-specific model for identifying the LBRs of the query protein.

2.2.5. Ligand-binding residues prediction

After the query-specific SVM model obtained, we generate the query sample set S_Q from the query protein by using the similar procedures described in the first paragraph of Section 2.2.4. In S_Q, each sample corresponds to one residue in the query protein. The probability (p_{Q_i}) of each residue to be LBR is gained easily, when the corresponding sample is fed into the query-specific SVM model.

2.3. Evaluation indices

The performance of LBR identification is evaluated by five routinely used indices, i.e., Recall (*Rec*), Specificity (*Spe*), Precision (*Pre*), Accuracy (*Acc*), and the Mathew's Correlation Coefficient (*MCC*):

$$Rec = \frac{TP}{TP + FN} \tag{7}$$

$$Spe = \frac{TN}{TN + FP} \tag{8}$$

$$Pre = \frac{TP}{TP + FP} \tag{9}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{10}$$

$$MCC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN) \cdot (TP + FP) \cdot (TN + FN) \cdot (TN + FP)}} \tag{11}$$

where *TN* (true negative), *TP* (true positive), *FN* (false negative), and *FP* (false positive) mean the numbers of true non-LBRs, true LBRs, false non-LBRs, and false LBRs in the prediction, respectively. In addition, the overall evaluation index *AUC*, which is the area under the Receiver Operating Characteristic (ROC) curve, is employed to evaluate the performance of LBR identification. Furthermore, the average (denoted as $Ave_{M,A}$) of *MCC* and *AUC* is also computed to testify the LBR identification performance. It is noted that the values of *TN*, *TP*, *FN*, and *FP* depend on a report-threshold *Th* in this study. If the probability (p_{Q_i}) of each residue to be LBR is larger than *Th*, the residue is predicted as LBR by the proposed I-LBR. Therefore, how to objectively determine the value of *Th* is a significant problem, especially in the situation of imbalanced learning scenario.

In this study, in order to determine the values of the parameters of I-LBR (including *Th*), we use the 10-fold cross-validation. Briefly, we randomly divided the TRAIN dataset into 10 subsets of equal size, where 9 subsets are employed to tune these parameters and the remaining subset is utilized as validation; for each combination of the parameter values in the grid space, such practice continued until all the 10 subsets of the TRAIN dataset are traversed over; an overall $Ave_{M,A}$ is finally computed on the union of 10 validation results. The parameters of I-LBR with the highest overall $Ave_{M,A}$ are finally selected for identifying LBRs from protein sequence.

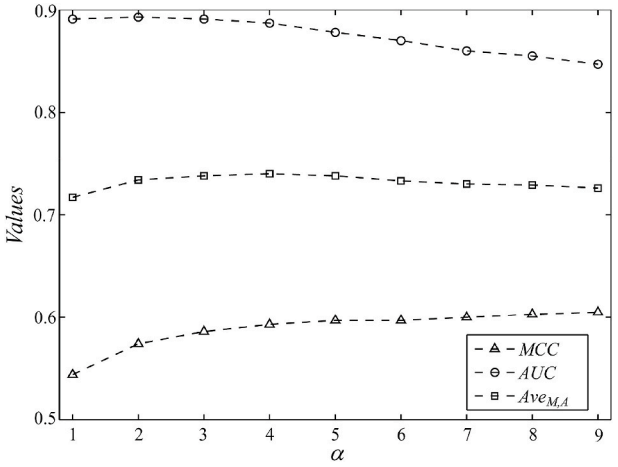

Fig. 2. The performance variation curves of *MCC*, *AUC*, and $Ave_{M,A}$ versus α under the general-purpose mode of I-LBR, i.e., I-LBR^{GP}.

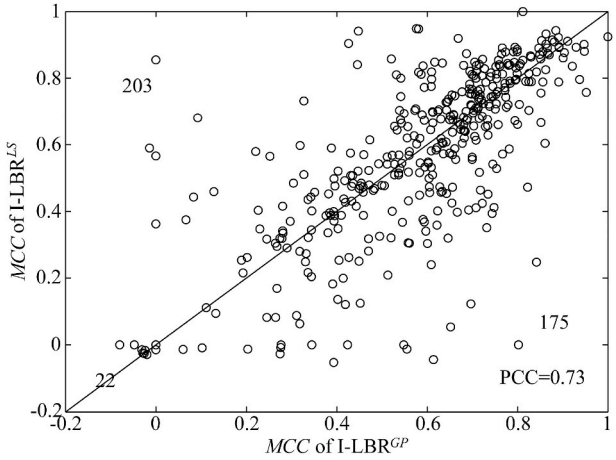

Fig. 3. Head-to-head comparisons of *MCC* values between I-LBR^{GP} and I-LBR^{LS} on the 400 proteins in TRAIN. The numbers in each panel represent the number of points in the upper and lower triangles, respectively. PCC is the Pearson's correlation coefficient between the MCCs of the two compared methods.

3. Experimental results and analysis

3.1. Tuning the value of α under the general-purpose mode of I-LBR

In this section, the parameter of α described in Section 2.2.4 is empirically tuned for search an appropriate value for training the query-specific SVM model under the general-purpose mode of I-LBR, i.e., I-LBR^{GP}. Concretely, we evaluate the performance variations by gradually varying the value of α from 1 to 9 with a step size of 1. For each value of α, the *MCC*, *AUC*, and $Ave_{M,A}$ values are computed by using 10-fold cross-validation described in Section 2.3 on the TRAIN dataset with fixing other parameters of I-LBR.

Fig. 2 shows the performance variation curves of *MCC*, *AUC*, and $Ave_{M,A}$ versus α under the general-purpose mode of I-LBR, i.e., I-LBR^{GP}. It is easily to find that the trends of *MCC* and *AUC* are completely monotonous and opposite. Therefore, we choose the value of α depending on the values of $Ave_{M,A}$. Concretely, when $\alpha \leq 4$, the overall trend of the values of $Ave_{M,A}$ tends to increase with the increment of α obvious enhancement. However, when $\alpha > 4$, the values of $Ave_{M,A}$ tend to decrease. It can be expected that the values of $Ave_{M,A}$ will further deteriorate with the increase in α when $\alpha > 9$, due to the noise and redundant information of positive samples will increase with the increment of α. In view of this, we will set $\alpha = 4$ in this study.

3.2. Performance comparison between the modes of general-purpose and ligand-specific

In this section, we compare the performance of two modes of I-LBR, i.e., I-LBR^{GP} and I-LBR^{LS}, on the TRAIN database over 10-fold cross-validation. It is noted that only the types of the ligands bond to the query proteins are used in the I-LBR^{LS}. The overall values of *Rec*, *Spe*, *Acc*, *Pre*, *MCC*, *AUC*, and $Ave_{M,A}$ of I-LBR^{GP} and I-LBR^{LS} are shown in Table 2.

From Table 2, we can find that I-LBR^{LS} is superiors to I-LBR^{GP} concerning the *Spe*, *Acc Pre*, *MCC*, and $Ave_{M,A}$ evaluation indexes. Concretely, the *Spe*, *Acc*, *Pre*, *MCC*, and $Ave_{M,A}$ values of I-LBR^{LS} are 0.987, 0.960, 0.735, 0.610, and 0.743, which are 1.13%, 0.73%, 17.79%, 2.87%, and 0.41% higher than that of I-LBR^{GP}, although I-LBR^{LS} has a lower *Rec* and *AUC*.

Table 2
Performance comparison of I-LBR^{GP} and I-LBR^{LS} over 10-fold cross-validation on the TRAIN dataset.

Mode of I-LBR	*Rec*	*Spe*	*Acc*	*Pre*	*MCC*	*AUC*	$Ave_{M,A}$
I-LBR^{GP}	0.613	0.976	0.953	0.624	0.593	0.887	0.740
I-LBR^{LS}	0.541	0.987	0.960	0.735	0.610	0.876	0.743

Fig. 3 demonstrates a head-to-head comparison of I-LBR^{LS} and I-LBR^{GP} based on *MCC*. In Fig. 3, each circle means one protein in the TRAIN dataset. Out of the 400 train proteins, there are 203 cases where I-LBR^{LS} has higher *MCC* than I-LBR^{GP}. Interestingly, the types of the ligands of the 203 proteins bond are frequently occurring in the TeD database. Depending on these results, we can assume that I-LBR^{LS} can be first employed to identify the LBRs when the known ligand types of the query protein are common in TeD. When the ligand types of the query protein are unknown or uncommon, we suggest the users employ the general-purpose mode of I-LBR, i.e., I-LBR^{GP}.

3.3. Comparison with existing predictors

In this section, we will compare the proposed method, I-LBR, with other existing LBRs identification methods to demonstrate its efficacy, including 6 general-purpose methods, i.e., ConCavity [4], FINDSITE [11], COFACTOR [47], S-SITE [15], TM-SITE [15], and COACH [15], and 4 ligand-specific methods, i.e., NsitePred [23], TargetS [48], ATPbind [16], and DELIA [49].

3.3.1. Performance comparisons on the TEST dataset

Table 3 lists the performance comparisons of the proposed I-LBR, ConCavity [4], FINDSITE [11], COFACTOR [47], S-SITE [15], TM-SITE [15], and COACH [15] on the TEST dataset. To achieve a relatively fair comparison, the *Rec*, *Pre*, and *MCC* values of the structure-based control methods, which are excerpted from the reference of COACH [15], are evaluated on the structures predicted by I-TASSER [10,50].

By observing Table 3, it is easy to find that both modes of the

Table 3
Performance comparison of I-LBR and other existing methods on the TEST dataset.

Methods	*Rec*	*Pre*	*MCC*
ConCavity[a]	0.510	0.230	0.260
FINDSITE[a]	0.490	0.440	0.420
COFACTOR[a]	0.390	0.560	0.420
TM-SITE[a]	0.490	0.570	0.480
COACH[a]	0.630	0.540	0.540
S-SITE[a]	0.580	0.450	0.450
I-LBR^{GP}	0.590	0.524	0.529
I-LBR^{LS}	0.575	0.561	0.543

[a] Data excerpted from the reference [15].

Table 4
Performance comparison of I-LBR^{LS} and other existing ATP-specific methods on the ATP-TEST dataset.

Methods	*Rec*	*Spe*	*Acc*	*Pre*	*MCC*	*AUC*	$Ave_{M,A}$
NsitePred [a]	0.467	0.977	0.954	0.492	0.456	0.852	0.654
TargetS [b]	0.516	0.989	0.967	0.689	0.580	0.872	0.726
DELIA [b]	0.506	N/A	N/A	0.730	0.593	0.901	0.747
ATPbind [a]	0.623	0.989	0.972	0.720	0.656	0.905	0.781
I-LBR^{LS} [c]	0.467	0.993	0.969	0.763	0.583	0.886	0.734

'N/A' means that the corresponding value is not provided.
[a] Data excerpted from the reference [16]; the results of ATPbind is obtained based on the protein structure predicted by I-TASSER [10].
[b] Data excerpted from the reference [17]; the results of DELIA is obtained based on the protein structure predicted by MODELLER [8].
[c] I-LBR^{LS} only care the ligand type of ATP.

proposed I-LBR can obtain *Pre* > 0.52 and *MCC* > 0.52 on the TEST dataset. The *MCC* value for I-LBR^{LS} is 0.543, with *Pre* 0.561 and *Rec* 0.575, respectively. The *MCC* and *Pre* values of I-LBR^{LS} are higher than that of I-LBR^{GS}, although I-LBR^{LS} has a lower *Rec*. However, due to the listed control methods are all for general-purpose LBR identification, it is fair to use I-LBR^{GP} to compare the ligand-binding residues identification performance between I-LBR and them. Concretely, the *Rec*, *Pre* and *MCC* values of I-LBR^{GP} are separately 1.72%, 16.44%, and 17.56% higher than that of S-SITE, which is the only sequence-based control method. The *Rec* and *MCC* values of I-LBR^{GP} are 15.69% and 103.46% higher than ConCavity, 20.41% and 25.95% higher than FINDSITE, 51.28% and 25.95% higher than COFACTOR, and 20.41% and 10.21% higher than TM-SITE, respectively. Comparing to the best control method COACH, the *Rec*, *Pre*, and *MCC* values of I-LBR^{GP} are 6.35%, 2.96%, and 2.04% lower. However, the ligand-specific mode of I-LBR, i.e., I-LBR^{LS}, could obtain the higher *Pre* and *MCC* than COACH.

3.3.2. Performance comparisons on the ATP-TEST dataset

To further testify the performance of I-LBR, Table 4 summarizes the comparisons between the I-LBR^{LS} and other four state-of-the-art ligand-specific LBR identification methods, i.e., two sequence-based methods (NsitePred [23] and TargetS [18]) and two structure-based methods (DELIA [17] and ATPbind [16]), on the ATP-TEST dataset.

By visiting Table 4, we can observe that the I-LBR^{LS} is superior to the two sequence-based methods, i.e., NsitePred [23] and TargetS [18], concerning the three overall evaluation indexes, i.e., *MCC*, *AUC*, and $Ave_{M,A}$. The *Pre*, *MCC*, *AUC*, and $Ave_{M,A}$ of I-LBR^{LS} are 0.763, 0.583, 0.886, and 0.734, which are 55.08%, 27.85%, 3.99%, and 12.23% higher than NsitePred and 10.74%, 0.52%, 1.61%, and 1.10% higher than TargetS, respectively, although I-LBR^{LS} has a slightly lower *Rec* than TargetS. However, out of the seven evaluation indexes, there are only two indexes, i.e., *Spe* and *Pre*, where I-LBR^{LS} are higher than the two structure-based methods, i.e., DELIA [17] and ATPbind [16], although their inputted protein structure data are predicted by MODELLER [8] and I-TASSER [10], respectively. The main reason is that most of the predicted structures used by DELIA and ATPbind are accurate in fold-level (TM-score > 0.5 [51]). Taking ATPbind as an example, out of the 41 testing proteins in the ATP-TEST dataset, there are 36 cases can be modeled by I-TASSER with a correct fold [16]. However, on the testing protein 5F1BC in ATP-TEST, whose I-TASSER-generated structure only has TM-score = 0.262, the *MCC* value of ATPbind is 0.183, with *Rec* = 0.231, *Spe* = 96.49, *Acc* = 0.938, and *Pre* = 0.200, while the *MCC* value of I-LBR^{LS} is 0.386, with *Rec* = 0.154, *Spe* = 1.00, *Acc* = 0.969, and *Pre* = 1.00, respectively. Hence, we suggest that the I-LBR^{LS} can be considered to use to identify the ATP-specific binding residues, when the query protein has no high-resolution structure.

3.3.3. Performance comparisons on the GTP-TEST dataset

Table 5 lists the performance comparisons of I-LBR^{LS}, NsitePred [23], and TargetS [18] on the GTP-TEST. The *MCC*, *AUC*, and $Ave_{M,A}$ of I-LBR^{LS} are 0.643, 0.955, and 0.799, which are 4.21%, 11.70%, and 8.56% higher than that of TargetS, respectively. Comparing to NsitePred, the I-LBR^{LS} achieve the higher values of the four evaluation indexes, i.e., *Rec*, *Spe*, *Acc*, and *MCC*. Since the GTP-specific and structure-based LBR method is rare, we have not compare the performance between I-LBR^{LS} with them.

Table 5
Performance comparison of I-LBR^{LS}, NsitePred, and TargetS on the GTP-TEST dataset.

Methods	*Rec*	*Spe*	*Acc*	*Pre*	*MCC*	*AUC*	$Ave_{M,A}$
NsitePred [a]	0.584	0.957	0.940	N/A	0.448	N/A	N/A
TargetS [a]	0.573	0.988	0.969	N/A	0.617	0.855	0.736
I-LBR^{LS} [b]	0.640	0.986	0.970	0.679	0.643	0.955	0.799

'N/A' means that the corresponding value is not provided.
[a] Data excerpted from the reference [18].
[b] I-LBR^{LS} only care the ligand type of GTP.

3.3.4. Performance comparisons on the GDP-TEST dataset

Table 6 demonstrates the performance comparisons of I-LBR^{LS}, NsitePred [23], and TargetS [18] on the GDP-TEST. It is easy to find that the proposed I-LBR^{LS} outperforms the two control methods, i.e., NsitePred [23], and TargetS [18]. Concretely, the *MCC* value of I-LBR^{LS} is 0.617, which is 12.18% higher than that of TargetS and 15.11% higher than that of NsitePred. In addition, the *AUC* and $Ave_{M,A}$ of I-LBR^{LS} are 0.912 and 0.765, which are 1.79% and 5.81% higher than that of TargetS, respectively. Due to the GDP-specific and structure-based LBR method cannot be found easily, we have not compare the performance between I-LBR^{LS} with them.

4. Conclusions

In this study, we report a new query-specific computational method, I-LBR, to identify the ligand-binding residues (LBRs) from protein sequence. I-LBR contains two modes, i.e., general-purpose and ligand-specific, named as I-LBR^{GP} and I-LBR^{LS}. For both modes of I-LBR, a template database, called TeD, is first pre-constructed; secondly, three sequence-related profiles, i.e., PSFM, PSSPM, and PSAPM, of query and template proteins are generated; then, based on these profiles, I-LBR generates the query-specific training subset; finally, the SVM algorithm is employed to train the query-specific identification model for obtaining the probability of each residue in the query protein belongs to the class of LBR. Our experimental results demonstrate that I-LBR can achieve a better or comparable performance against the state-of-the-art LBR identification methods. When the ligand type/types of the query protein binds is/are known, the experimental results suggest the user should select the ligand-specific mode of I-LBR, i.e., I-LBR^{LS}. In the future, we plan to further improve the performance of I-LBR mainly in three aspects. The first is that collecting more available ligand-binding proteins to increase the capacity of the template database, i.e., TeD. The second is that the protein-level features, such as PseAAC [52–55], generated by two powerful web-servers called 'Pse-in-One' [56,57] and

Table 6
Performance comparison of I-LBR^{LS}, NsitePred, and TargetS on the GDP-TEST dataset.

Methods	*Rec*	*Spe*	*Acc*	*Pre*	*MCC*	*AUC*	$Ave_{M,A}$
NsitePred [a]	0.557	0.979	0.961	N/A	0.536	N/A	N/A
TargetS [a]	0.562	0.981	0.962	N/A	0.550	0.896	0.723
I-LBR^{LS} [b]	0.572	0.989	0.970	0.698	0.617	0.912	0.765

'N/A' means that the corresponding value is not provided.
[a] Data excerpted from the reference [18].
[b] I-LBR^{LS} only care the ligand type of GDP.

BioSeq-Analysis [58] should be carefully employed in the model of I-LBR. The third is that the SVM algorithm used in this study should be replaced by the deep convolution neural network algorithm [59,60], whose effectiveness has been verify in DELIA [17], to enhance the ability of learning knowledge. In addition, we will use the graphic approaches [61,62] to study biological and medical systems to further provide an intuitive vision and useful insights for helping analyze complicated relations.

Author contributions

J.H., L.R., and G.Z. designed research; J.H. and L.R. performed research; J.H. and X.F. analyzed data; and J.H. and X.F. wrote the paper; J.H. established the web-server.

Declaration of competing interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61902352 and 61773346), the Key Laboratory of Data Science and Intelligence Application, Fujian Province University (No. D1903), and Natural Science Foundation of Zhejiang (No. LZ20F030002).

References

[1] P.W. Rose, A. Prlić, A. Altunkaya, C. Bi, A.R. Bradley, C.H. Christie, L.D. Costanzo, J.M. Duarte, S. Dutta, Z. Feng, The RCSB protein data bank: integrative view of protein, gene and 3D structural information, Nucleic Acids Res. 45 (D1) (2017) D271–D281.
[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Molecular Biology of the Cell, Garland Science., New York, 2008.
[3] R.A. Laskowski, N.M. Luscombe, M.B. Swindells, J.M. Thornton, Protein clefts in molecular recognition and function, Protein Sci.: Publ. Protein Soc. 5 (12) (1996) 2438.
[4] J.A. Capra, R.A. Laskowski, J.M. Thornton, M. Singh, T.A. Funkhouser, Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure, PLoS Comput. Biol. 5 (12) (2009) e1000585.
[5] M. Hendlich, F. Rippmann, G. Barnickel, LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins, J. Mol. Graph. Model. 15 (6) (1997) 359–363.
[6] R.A. Laskowski, SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions, J. Mol. Graph. 13 (5) (1995) 323–330.
[7] J. An, M. Totrov, R. Abagyan, Pocketome via comprehensive identification and classification of ligand binding envelopes, Mol. Cell. Proteomics 4 (6) (2005) 752–761.
[8] A. Sali, T.L. Blundell, Comparative protein modeling by satisfaction of spatial restraints, J. Mol. Biol. 234 (3) (1993) 779–815.
[9] S. Raman, B. Qian, D. Baker, R.C. Walker, Advances in Rosetta protein structure prediction on massively parallel systems, IBM J. Res. Dev. 52 (1.2) (2008) 7–17.
[10] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The I-TASSER Suite: protein structure and function prediction, Nat. Methods 12 (1) (2015) 7–8.
[11] M. Brylinski, J. Skolnick, A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation, Proc. Natl. Acad. Sci. U. S. A. 105 (1) (2008) 129–134.
[12] D.B. Roche, S.J. Tetchner, L.J. McGuffin, FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins, BMC Bioinf. 12 (1) (2011) 160.
[13] M.N. Wass, L.A. Kelley, M.J. Sternberg, 3DLigandSite: predicting ligand-binding sites using similar structures, Nucleic Acids Res. 38 (suppl_2) (2010) W469–W473.
[14] C. Zhang, P.L. Freddolino, Y. Zhang, COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information, Nucleic Acids Res. 45 (W1) (2017) W291–W299.
[15] J. Yang, A. Roy, Y. Zhang, Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment, Bioinformatics 29 (20) (2013) 2588–2595.
[16] J. Hu, Y. Li, Y. Zhang, D.-J. Yu, ATPbind: accurate protein–ATP binding site prediction by combining sequence-profiling and structure-based comparisons, J. Chem. Inf. Model. 58 (2) (2018) 501–510.
[17] C.-Q. Xia, X. Pan, H.-B. Shen, Protein-ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data, Bioinformatics 1 (1) (2020) 1.
[18] D.J. Yu, J. Hu, J. Yang, H.B. Shen, J.H. Tang, J.Y. Yang, Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering, IEEE ACM Trans. Comput. Biol. Bioinf 10 (4) (2013) 994–1008.
[19] J.A. Capra, M. Singh, Predicting functionally important residues from sequence conservation, Bioinformatics 23 (15) (2007) 1875–1882.
[20] J. Hu, X. He, D.-J. Yu, X.-B. Yang, J.-Y. Yang, H.-B. Shen, A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction, PloS One 9 (9) (2014) e107676.
[21] A. Armon, D. Graur, N. Ben-Tal, ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information, J. Mol. Biol. 307 (1) (2001) 447–463.
[22] T. Pupko, R.E. Bell, I. Mayrose, F. Glaser, N. Ben-Tal, Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues, Bioinformatics 18 (suppl_1) (2002) S71–S77.
[23] K. Chen, M.J. Mizianty, L. Kurgan, Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors, Bioinformatics 28 (3) (2012) 331–341.
[24] S. Henrich, O.M. Salo-Ahen, B. Huang, F.F. Rippmann, G. Cruciani, R.C. Wade, Computational approaches to identifying and characterizing protein binding sites for ligand design, J. Mol. Recogn. 23 (2) (2010) 209–219.
[25] R. Liu, J. Hu, HemeBIND: a novel method for heme binding residue prediction by combining structural and sequence information, BMC Bioinf. 12 (2011) 207.
[26] M. Brylinski, J. Skolnick, FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level, Proteins: Struct. Funct. Bioinf. 79 (3) (2011) 735–751.
[27] J.S. Chauhan, N.K. Mishra, G.P. Raghava, Identification of ATP binding residues of a protein from its primary sequence, BMC Bioinf. 10 (2009) 434.
[28] K. Chen, M.J. Mizianty, L. Kurgan, ATPsite: sequence-based prediction of ATP-binding residues, Proteome Sci. 9 (1) (2011) S4.
[29] J. Si, Z. Zhang, B. Lin, M. Schroeder, B. Huang, MetaDBSite: a meta approach to improve protein DNA-binding sites prediction, BMC Syst. Biol. 5 (Suppl 1) (2011) S7.
[30] J. Hu, Y. Li, M. Zhang, X. Yang, H.-B. Shen, D.-J. Yu, Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs, IEEE ACM Trans. Comput. Biol. Bioinf 14 (6) (2017) 1389–1398.
[31] Y.-H. Zhu, J. Hu, X.-N. Song, D.-J. Yu, DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines, J. Chem. Inf. Model. 59 (2019) 3057–3071.
[32] J. Yang, A. Roy, Y. Zhang, BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions, Nucleic Acids Res. 41 (D1) (2013) D1096–D1103.
[33] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (1) (2011) 236–247.
[34] K.-C. Chou, Other mountain stones can attack jade: the 5-steps rule, Nat. Sci. 12 (3) (2020) 59–64.
[35] K.-C. Chou, Proposing 5-steps rule is a notable milestone for studying molecular biology, Nat. Sci. 12 (2020) 74 03.
[36] W. Lin, X. Xiao, W. Qiu, K.-C. Chou, Use chou's 5-steps rule to predict remote homology proteins by merging grey incidence analysis and domain similarity analysis, Nat. Sci. 12 (2020) 181 03.
[37] K.C. Chou, Advances in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs, Curr. Med. Chem. 26 (26) (2019) 4918–4943.
[38] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.
[39] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, Nucleic Acids Res. 35 (suppl_1) (2006) D61–D65.
[40] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (2) (1999) 195–202.
[41] K. Joo, S.J. Lee, J. Lee, Sann: solvent accessibility prediction of proteins by nearest neighbor method, Proteins Struct. Funct. Bioinf. 80 (7) (2012) 1791–1797.
[42] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (3) (1970) 443–453.
[43] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. Unit. States Am. 89 (22) (1992) 10915–10919.
[44] V.N. Vapnik (Ed.), Statistical Learning Theory Wiley-Interscience, New York, 1998.
[45] D.J. Yu, J. Hu, Q.M. Li, Z.M. Tang, J.Y. Yang, H.B. Shen, Constructing query-driven dynamic machine learning model with application to protein-ligand binding sites prediction, IEEE Trans. NanoBioscience 14 (1) (2015) 45–58.
[46] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27.
[47] A. Roy, J. Yang, Y. Zhang, COFACTOR: an accurate comparative algorithm for structure-based protein function annotation, Nucleic Acids Res. (2012) gks372.
[48] D.J. Yu, J. Hu, J. Yang, H.B. Shen, J.H. Tang, J.Y. Yang, Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering, IEEE ACM Trans. Comput. Biol. Bioinf 10 (4) (2013) 994–1008.
[49] C.-Q. Xia, X. Pan, H.-B. Shen, Protein–ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data, Bioinformatics (2020).
[50] Y. Zhang, I-TASSER server for protein 3D structure prediction, BMC Bioinf. 9 (1) (2008) 40.
[51] J. Xu, Y. Zhang, How significant is a protein structure similarity with TM-score =

0.5? Bioinformatics 26 (7) (2010) 889–895.
[52] L.M. Liu, Y. Xu, K.C. Chou, iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC, Med. Chem. 13 (6) (2017) 552–559.
[53] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (1) (2005) 10–19.
[54] P. Du, S. Gu, Y. Jiao, PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets, Int. J. Mol. Sci. 15 (3) (2014) 3495–3506.
[55] K.C. Chou, Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology, Curr. Proteonomics 6 (4) (2009) 262–274.
[56] B. Liu, F.L. Liu, X.L. Wang, J.J. Chen, L.Y. Fang, K.C. Chou, Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Res. 43 (W1) (2015) W65–W71.
[57] B. Liu, H. Wu, K.-C. Chou, Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nat. Sci. 9 (2017) 67 04.
[58] B. Liu, BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches, Briefings Bioinf. (2017).
[59] B.B. Traore, B. Kamsu-Foguem, F. Tangara, Deep convolution neural network for image recognition, Ecol. Inf. 48 (2018) 257–268.
[60] Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, The Handbook of Brain Theory and Neural Networks, 3361 1995, p. 1995 10.
[61] K.-C. Chou, S. Forsén, Graphical rules for enzyme-catalysed rate laws, Biochem. J. 187 (3) (1980) 829–835.
[62] G. Zhou, M. Deng, An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways, Biochem. J. 222 (1) (1984) 169.