JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2018, 58, 1459−1468

pubs.acs.org/jcim

Improving Sequence-Based Prediction of Protein−Peptide Binding Residues by Introducing Intrinsic Disorder and a Consensus Method

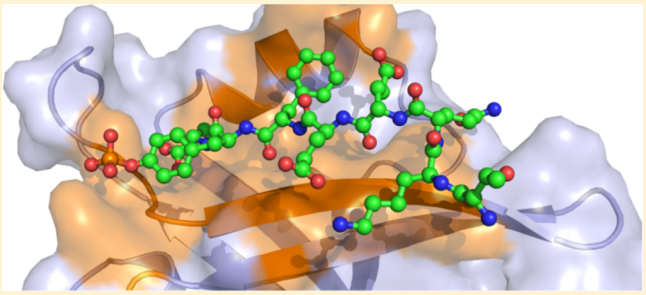
Zijuan Zhao,[†] Zhenling Peng,*[,†] and Jianyi Yang*[,‡]

[†]Center for Applied Mathematics, Tianjin University, Tianjin 300072, China

[‡]School of Mathematical Sciences, Nankai University, Tianjin 300071, China

Supporting Information

ABSTRACT: Protein−peptide interaction is crucial for many cellular processes. It is difficult to determine the interaction by experiments as peptides are often very flexible in structure. Accurate sequence-based prediction of peptide-binding residues can facilitate the study of this interaction. In this work, we developed two novel sequence-based methods SVMpep and PepBind to identify the peptide-binding residues. Recent studies demonstrate that the protein−peptide binding is closely associated with intrinsic disorder. We thus introduced intrinsic disorder in our feature design and developed the *ab initio* method SVMpep. Experiments show that intrinsic disorder contributes to 1.2−5.2% improvement in area under the receiver operating characteristic curve (AUC). Comparison to the recent sequence-based method SPRINT-Seq reveals that SVMpep improves the AUC and Matthews correlation coefficient (MCC) by at least 7.7% and 70%, respectively. In addition, by combining SVMpep with two template-based methods S-SITE and TM-SITE, we next proposed the consensus-based method PepBind. Remarkably, compared with the latest structure-based method SPRINT-Str, PepBind improves the AUC and MCC by 1.7% and 28.3%, respectively, on the same independent test set of SPRINT-Str. The success of PepBind is attributed to the improved prediction of the *ab initio* method SVMpep by introducing intrinsic disorder and the consensus prediction by combining three complementary methods. A web server that implements the proposed methods is freely available at http://yanglab.nankai.edu.cn/PepBind/.



1. INTRODUCTION

Protein−peptide interaction is essential for many cellular processes, such as programmed cell death,[1] gene expression,[2] DNA replication and repair,[3] and so on. The peptides involved in protein binding are usually flexible in structure, short in length, and weak in binding affinity,[4] which challenge the experimental detection of these interactions. Fortunately, there are many researchers working on the investigation of protein−peptide interactions. These efforts lead to a steady increase in the experimental source of protein−peptide interactions, with about 20 000 entries of protein−peptide complex structures in the BioLiP database.[5]

The functional importance, the challenges in experimental determination, and the availability of the experimental data motivate the development of computational methods for the prediction of protein−peptide binding residues. These predictors allow for high-throughput peptide-binding annotations in protein chains and therefore provide a viable solution to the investigation of protein−peptide binding events. They can be broadly classified into structure- and sequence-based methods.

The structure-based methods include PepSite,[6] Peptimap,[7] ACCLUSTER,[8] SPRINT-Str,[9] and so on. PepSite works by searching the regions that match the spatial matrix derived from known protein−peptide complex structures and employs the distance constraints to infer the binding sites. The Peptimap protocol was developed based on fragment mapping and clustering by considering the characteristics of peptide binding sites. In ACCLUSTER, 20 standard amino acids were first used as probes to scan the protein surface to generate binding poses of strong chemical interactions with the protein. This was followed by clustering of the poses and the largest cluster was used to infer the binding sites. The recent method SPRINT-Str applies the machine learning algorithm random forest to predict the binding residues, based on the sequence and the structural information extracted from the experimental data.

The structures for most proteins are not available, which motivates the development of sequence-based methods. To the best of our knowledge, there is only one sequence-based method for peptide-binding residues prediction, i.e., SPRINT-Seq[10] It is a machine learning based method with features extracted from sequence profile and predicted structural attributes. On the other hand, the past decades witnessed the development of many sequence-based approaches to the prediction of binding residues for other ligands, such as DNA/RNA,[11] Coenzyme A (COA),[12] carbohydrates,[13] and so on.

Received: January 13, 2018
Published: June 12, 2018

DOI: 10.1021/acs.jcim.8b00019
J. Chem. Inf. Model. 2018, 58, 1459−1468

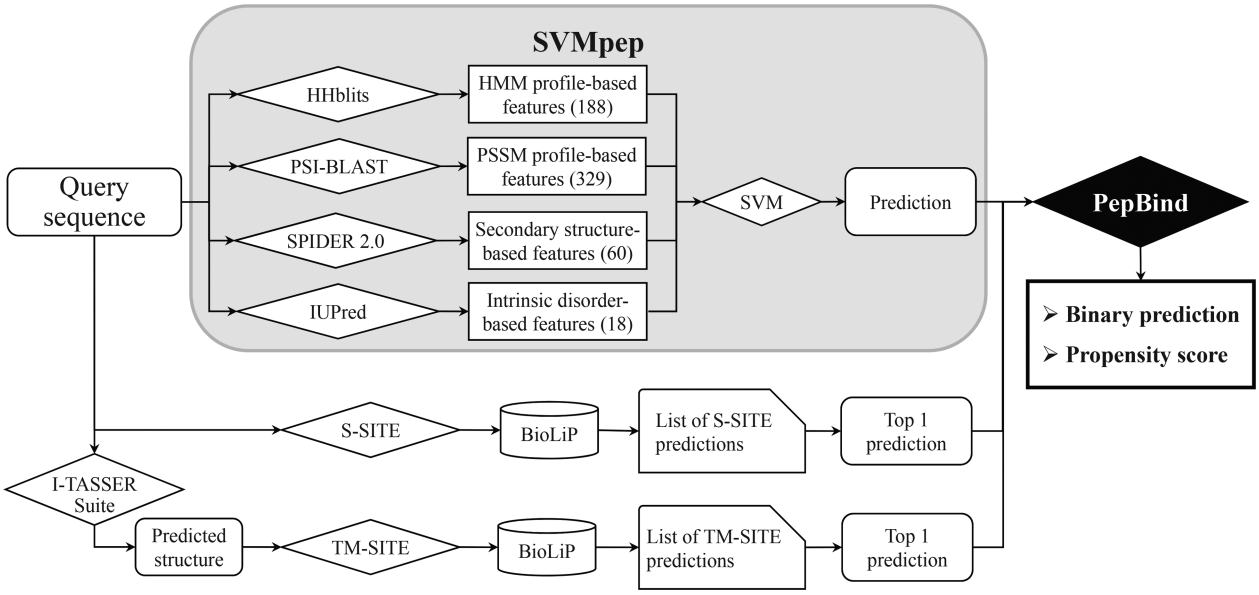


Figure 1. Architecture of the proposed methods SVMpep and PepBind.

Although intrinsic disorder has been reported to participate in protein−peptide interactions,[14] no existing peptide binding predictors considered this valuable sequence characteristics. We therefore proposed a novel *ab initio* model SVMpep by introducing the intrinsic disorder in feature design for the first time. The evaluation shows that the accurate binding prediction of SVMpep benefits from the integration of intrinsic disorder in sequence representation. Moreover, the consensus-based method, which combines several individual predictors together, usually outperforms its components. This fact motivated us to develop the first-of-its-kind consensus-based method PepBind, by combining the *ab initio* model SVMpep with the template-based methods S-SITE and TM-SITE,[15] to further improve the accuracy for the prediction of peptide binding residues. The comprehensive assessments suggest that the consensus PepBind is not only more accurate than all three individual methods but also outperforms the existing sequence-based method SPRINT-Seq and the recent structure-based method SPRINT-Str.

2. METHODS

2.1. Benchmark Data Sets. The 1279 peptide-binding proteins (PBPs) from the work of SPRINT-Seq are used in this study; these contain 16 749 peptide-binding residues and 290 943 nonbinding residues. This data set was initially extracted from the BioLiP database[5] and redundancy with >30% sequence identity was removed. We randomly divided these 1279 PBPs into two subsets of equal size for training and test, respectively. The training set comprises 640 PBPs with 8259 (149 103) binding (nonbinding) residues (denoted by TR640). To save the time used for training, 20% of the proteins in TR640 were randomly selected for a coarse-grained tuning of parameters (*i.e.*, 128 proteins, denoted by TR128). The test set contains 639 PBPs with 8490 (141 840) binding (nonbinding) residues (denoted by TE639). To compare with other methods, we also collected the test data sets of the sequence-based method SPRINT-Seq[10] and the structure-based method SPRINT-Str,[9] which comprise 80 (denoted by TE80) and 125 proteins (denoted by TE125), respectively. Comparisons of these data sets show that TE80 is a subset of TE125 and TE639, while 24 proteins in TE125 are from the training set TR640 and the remaining 101 proteins are from test set TE639.

2.2. *Ab Initio* Method SVMpep. As shown in the gray panel of Figure 1, the query sequence is submitted to the sequence−profile alignment algorithm PSI-BLAST,[16] the profile−profile alignment program HHblits,[17] the secondary structure prediction tool SPIDER 2.0,[18] and the intrinsic disorder predictor IUPred.[19] The derived information is believed to be enriched with evolutionary and structure information, which were widely used in the prediction of protein structure and function.[10,20,21] Notably, it is the first time to introduce the intrinsic disorder information for the prediction of peptide-binding residues. In fact, the peptides involved in this binding event are rich of intrinsic disorder.[14] For each residue, four feature groups are extracted from its neighboring residues embedded in a sliding window. The window size for each feature group was separately selected to optimize the overall performance (AUC, defined later) of 5-fold cross validation on the training set TR128 (Supporting Figure S1).

2.3. Intrinsic Disorder-Based Features. Peptides participated in protein−peptide interactions are characterized with short length, flexible structure, and weak binding affinity. These are consistent with the attributes of short linear motifs, which are mostly located in the intrinsic disordered regions (IDRs).[14] We therefore propose to incorporate the information on intrinsic disorder in the prediction of peptide-binding residues. Here, we employed the method IUPred[19] to predict the intrinsic disorder, including the short and the long IDRs, from the protein sequence. For each type of putative annotations (*i.e.*, short or long IDRs), we derived the following nine features for each query residue. These features indicate the structural flexibility of the query residue and its neighbors and have been successfully applied in the prediction of functions related to intrinsic disorder.[21,22] The first three features are the average, the minimum, and the maximum among the propensity scores of residues in a sliding window of size 19 centered at the query residue. The window size was selected based on optimization using the training set TR128

DOI: 10.1021/acs.jcim.8b00019
J. Chem. Inf. Model. 2018, 58, 1459−1468

(Supporting Figure S1). The fourth feature is the difference between the average of the propensity scores for the residues inside the sliding window (*i.e.*, close neighbors) and the corresponding average of the query residue's remote neighbors (i.e., residues outside the sliding window and 9 residues next to each terminal of this window are used). The fifth and the sixth features are the propensity score and the predicted disorder label (0 or 1) for the query residue, respectively. The last three features are the length of the IDR in which the query residue located, the minimum and the maximum distances of the query residue from both ends of the IDR. In total, we generated 18 (= 9 × 2) features from the IUPred outputs of the short and the long IDRs.

The recent disorder predictor SPOT-disorder[23] was used to replace IUPred to see if SVMpep could be improved further. The 5-fold cross validation on training set TR640 shows that SPOT-disorder-based SVMpep has a very similar value of AUC to the IUPred-based SVMpep (0.73 versus 0.721). Moreover, IUPred was designed for short and long IDRs, while SPOT-disorder does not distinguish between short and long IDRs. Thus, we prefer using the IUPred program for disorder prediction in this work.

2.4. Secondary Structure-Based Features. Secondary structure (SS) comprises three distinct local conformations: α-helix (H), β-strand (E), and γ -coil (C) and is suggested to be useful in predicting peptide-binding residues.[10] For a query sequence, we employed the method SPIDER 2.0[18] to generate the probabilities of the SS types for each residue. The SS type for each residue is set to the one with the highest probability. Utilizing a sliding window of size 7, we first encoded a query residue by a probability matrix (7 × 3 features) and the fraction of each SS type (3 features). We next used a 27-dimensinal vector to represent the secondary structure of the triplet with the query residue and its two nearest neighbors. We also considered two kinds of segment-based features, including the length of the segment enclosing the same SS type of the query residue, and the minimum and the maximum distances of the query residue from both ends of this segment. They were employed previously to predict protein−RNA and protein−peptide binding residues.[10,24] As there are three types of SS, this segment-based property is represented by 9 (= 3 × 1 + 3 × 2) features. In total, 60 features were extracted from the SPIDER 2.0 outputs.

2.5. PSSM Profile-Based Features. The functionally important residues are generally more conserved than others along evolution. The residue conservation can be inferred from a multiple sequence alignment (MSA). We used PSI-BLAST to find homologous sequences for a query sequence from the NCBI's nonredundant data set with three iterations and the *e*-value threshold of 0.001 ("-j 3 -h 0.001"). A position specific scoring matrix (PSSM) and a probability matrix was next derived from the MSA. Based on these two matrices, the conservation of each residue can be represented by a 329-dimensional feature vector, by utilizing a sliding window of size 15. This feature vector includes 300 (= 20 × 15) features extracted from PSSM, 15 relative entropy[25] (RE) values, and 14 near neighbors correlation coefficient (CNCC) from the probability matrix. CNCC was introduced to measure the correlation between neighbor residues in SPRINT-Seq[10]

$$\mathrm{RE}_i = \sum_{k=1}^{20} p_{ik} \log_2 \frac{p_{ik}}{b_k} \tag{1}$$

$$\mathrm{CNCC}_{ij} = \frac{P_i \cdot P_j}{|P_i||P_j|} = \frac{\sum_{k=1}^{20} p_{ik} p_{jk}}{\sqrt{\left(\sum_{k=1}^{20} {p_{ik}}^2\right)\left(\sum_{k=1}^{20} {p_{jk}}^2\right)}}, \quad i \neq j \tag{2}$$

where i is the ith residue in a protein with L amino acids; j is the adjacent residue to the ith residue in a sliding window of size 15; k represents one of the 20 standard residues; b_k is the Robinson background frequency[26] of the residue k; p_{ik} is the probability of the residue k appearing at the ith column (corresponding to the ith residue in the query) of the MSA; and P_i is a 20-dimensional vector, corresponding to the ith row in the probability matrix.

2.6. HMM Profile-Based Features. The profile−profile alignment algorithm HHblits is shown to be faster and more sensitive than PSI-BLAST.[17] We thus used HHblits to generate another profile for the calculation of residue conservations. Specifically, HHblits searches the homologies of a given query from the database uniprot20_2015_06 with the default parameters, i.e., "-n 3 -maxfilt 500000 -diff inf -id 99 -cov 60". Then a hidden Markov model (HMM) profile is obtained. Each line in this profile comprises the emission frequencies (EFs) for the 20 standard amino acids, 7 transition probabilities, and 3 local diversities. The EFs for a given residue in the query are defined by the following equation

$$\mathrm{EF}_{ik} = -1000 \log_2 {}^{p_{ik}} \tag{3}$$

where i is the ith residue; k represents a standard residue. Based on this equation, the EFs are then converted back into probabilities p_{ik}, which equal to 0 when the EF is denoted by a star "*". The recovered probability matrix is then used to measure the residue conservation, by including the 20-dimentional vector from this matrix and the CNCC features as well. Here, we used a sliding window of size 9 to obtain 188 (= 9 × 20 + 8) features in total.

2.7. Support Vector Machine (SVM). A total of 595 (= 18 + 60 + 329 + 188) features have been extracted above and they are fed into SVM for training and classification. SVM has been successfully applied to many different classification problems.[20,24] Due to the better predictive performance, the radial basis functional (RBF) kernel was selected here. Thus, our SVM classifier has two key parameters, the regularization factor C and the kernel parameter γ. Considering C in [0.5, 1, 2, 4, 8, 16] and γ in [0.0625, 0.125, 0.25, 0.5], we performed a grid search to optimize the overall AUC based on a 5-fold cross validation on the training set TR640. The implementation of SVM was based on the LIBSVM package (https://www.csie.ntu.edu.tw/~cjlin/libsvm/). Before training and test, the program "svm-scale" was used to normalize the features into the range of [−1, 1]. A residue is predicted as a binding residue if the probability score from SVM is higher than the threshold of 0.25, to maximize the MCC on the training set TR640 (see Supporting Figure S2).

2.8. Template-Based Methods TM-SITE and S-SITE. Template-based method works by transferring the binding annotations from homologous templates to the query based on query-template alignments. TM-SITE and S-SITE are two typical template-based methods, which make use of structure and sequence profile information, respectively.[15] The protein−ligand complex from BioLiP[5] are used as the template library. In TM-SITE, the alignment was done based structure alignment of ligand-binding specific structural fragments.

DOI: 10.1021/acs.jcim.8b00019
J. Chem. Inf. Model. 2018, 58, 1459−1468

Global structure was not employed to reduce the influence of regions that are not related to ligand binding. For each template of protein−ligand complex, the template fragment corresponds to the subsequence from the first binding residue to the last binding residue (called SSFL). In general, each SSFL consists of at least three residues. For the query structure, its SSFL was extracted from geometry-based binding pocket detection. In S-SITE, the alignment was performed based on sequence profile−profile alignment. It was designed to complement TM-SITE, especially for proteins with no structure information and proteins with low-resolution structures. Scoring functions were designed to judge the reliability of their predictions. For more details about these methods, please refer to the original work.[15] In this work, the template library of TM-SITE and S-SITE was restricted to peptides (*i.e.*, with BioLiP's ligand ID "III") as we aim to peptide-binding residues.

2.9. Consensus-Based Method PepBind. Since the *ab initio* predictor and the template-based method are usually complementary, we next designed a consensus-based method PepBind, by combining SVMpep with two template-based methods S-SITE and TM-SITE.[15] These two methods were developed for the detection of binding residues for ligands of general type, and are available in the I-TASSER Suite.[27] S-SITE utilizes the binding-specific sequence profile−profile alignment and TM-SITE works by the binding-specific substructures matching. The protein structure required by the TM-SITE was generated by the I-TASSER Suite.[27] Note that for both structure modeling and binding residues prediction, all templates with 30% sequence identity to the query sequence were excluded for fair comparison. The outputs of the program S-SITE/TM-SITE comprise a list of predictions, where the predictions are sorted by their confidence scores (c-scores) from high to low. Specifically, the S-SITE (TM-SITE) method suggests that the prediction with c-score ≥ 0.25 (0.35) is reliable. We considered their top 1 predictions (see Figure 1) and the corresponding c-scores to design PepBind. Here, we denote the length of a protein sequence by L; i (i = 1, 2, ..., L) represents the ith residue in this sequence; 0/1 is the binary prediction of a given method, indicating a predicted nonbinding/binding residue; S-SITE/TM-SITE/SVMpep generates both binary $b_i^1/b_i^2/b_i^3$ and propensity score $p_i^1/p_i^2/p_i^3$ for the ith residue. The binary b_i^c and the propensity score p_i^c for the ith residue by PepBind are calculated as follows.

Case 1. If the top 1 predictions of S-SITE and TM-SITE are both reliable, i.e., S-SITE c-score ≥ 0.25 and TM-SITE c-score ≥ 0.35, then these two predictions along with the SVMpep outputs are combined to form the final PepBind outputs, where the binary b_i^c and the propensity score p_i^c are formulated as below.

$$b_i^c = \begin{cases} 1, & \text{if } \sum_{k=1}^{3} b_i^k > 1 \\ 0, & \text{otherwise} \end{cases}, \quad \text{and}$$

$$p_i^c = \begin{cases} \sum_{k=1}^{3} b_i^k p_i^k / \sum_{k=1}^{3} b_i^k, & \text{if } \sum_{k=1}^{3} b_i^k > 1 \\ \sum_{k=1}^{3} (1 - b_i^k) p_i^k / \sum_{k=1}^{3} (1 - b_i^k), & \text{otherwise} \end{cases} \tag{4}$$

Case 2. If only S-SITE or TM-SITE generates a reliable prediction, the corresponding prediction is integrated with the SVMpep prediction, to infer the PepBind outputs by the following formulas.

$$b_i^c = \begin{cases} 1, & \text{if } b_i^k + b_i^3 \geq 1 \\ 0, & \text{otherwise} \end{cases}, \quad \text{and}$$

$$p_i^c = \begin{cases} \max\{p_i^k, p_i^3\}, & \text{if } b_i^k + b_i^3 \geq 1 \\ (p_i^k + p_i^3)/2, & \text{otherwise} \end{cases} \tag{5}$$

where k = 1 or 2, representing the method S-SITE and TM-SITE, respectively.

Case 3. When both S-SITE and TM-SITE do not have reliable predictions, we still employed their top 1 predictions. A residue is regarded as in binding if both methods predict it as a binding residue. The template-based binaries b_i and propensity scores p_i are first refined by

$$b_i = \begin{cases} 1, & \text{if } b_i^1 + b_i^2 = 2 \\ 0, & \text{otherwise} \end{cases}, \quad \text{and}$$

$$p_i = \begin{cases} (p_i^1 + p_i^2)/2, & \text{if } b_i^1 + b_i^2 = 2 \\ \min\{p_i^1, p_i^2\}, & \text{otherwise} \end{cases} \tag{6}$$

We then combined these updated outputs with the SVMpep annotations by the strategy described in case 2.

2.10. Evaluation Criteria. The proposed methods SVMpep and PepBind take the protein sequence as input and provide the binary and the propensity score for each residue, where the propensity score suggests the likelihood of a residue to be in binding. As the nonbinding residues (negatives) are about 17 times more than the binding residues (positives), indicating that the data sets are highly unbalanced, the predictive performance for the binary prediction of our proposed methods are assessed by Precision (Pre), Recall (Rec) and Matthews correlation coefficient (MCC).

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{9}$$

DOI: 10.1021/acs.jcim.8b00019
J. Chem. Inf. Model. 2018, 58, 1459−1468

where TP (true positive) is the number of correctly predicted binding residues, TN (true negative) is the number of correctly predicted nonbinding residues, FP (false positive) is the number of nonbinding residues that are incorrectly predicted as binding residues, and FN (false negative) is the number of binding residues that are incorrectly predicted as nonbinding residues. MCC ranges from −1 to 1. An MCC of zero indicates a random prediction. Higher values of the above measures indicate better binary prediction.

The prediction with the propensity scores is evaluated by the receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC). For each real value p (from 0 to 1), the residues with propensity score $\geq p$ are set as positives (binding residues); otherwise, the residues are set as negatives (nonbinding residues). Therefore, each p corresponds to a point (FP-rate, TP-rate), where TP-rate = TP/(TP + FN), and FP-rate = FP/(FP + TN). By connecting all the points for all values of p between 0 and 1, a ROC curve is generated. The area under the ROC curve (AUC) thus implies the predictive quality of the propensity scores. The AUC value is between 0 and 1 and the higher the better. An AUC of 0.5 indicates a random prediction. The AUC lower than 0.5 indicates a reverse prediction.

3. RESULTS

In this section, we present the performance of the component and the consensus methods on different data sets. The performance of the consensus method PepBind is further analyzed on the proteins with different recognition types and different binding ligands.

3.1. Performance of Component Methods. The results for all component methods are summarized in Table 1 and Figure 2. Based on 5-fold cross validation on the training set TR640, SVMpep achieves the overall AUC, MCC, Rec, and Pre with the values of 0.721, 0.254, 0.106, and 0.680, respectively. This is consistent with the predictive performance on the independent test set TE639, where the value of AUC, MCC, Rec, and Pre are 0.716, 0.253, 0.105, and 0.676, respectively (Table 1). The low Rec value of 0.11 suggests that SVMpep is underpredicted. This could be because the data used for training and testing are extremely unbalanced. But the Pre value reflects that about 68% predicted binding residues are correct. The ROC curve in Figure 2 also shows that the TP-rate of the SVMpep method can reach up to 30% at the low FP-rate of 5%.

Table 1. Overall Performance of Different Methods on the Training Set TR640 and the Independent Test Set TE639[a]

data set	method	AUC	MCC	Rec	Pre
TR640	S-SITE	0.691	0.307	**0.321**	0.364
	TM-SITE	0.603	0.210	0.213	0.284
	SVMpep	0.721	0.254	0.106	**0.680**
	SVMpep + S-SITE	0.747	0.321	0.307	0.445
	SVMpep + TM-SITE	0.717	0.291	0.213	0.466
	PepBind	**0.759**	**0.329**	0.294	0.435
	REpep	0.589	0.071	0.192	0.100
	COACH	0.649	0.253	0.295	0.290
TE639	S-SITE	0.692	0.309	**0.335**	0.360
	TM-SITE	0.611	0.216	0.230	0.284
	SVMpep	0.716	0.253	0.105	**0.676**
	SVMpep + S-SITE	0.751	0.339	0.312	0.440
	SVMpep + TM-SITE	0.731	0.311	0.224	0.505
	PepBind	**0.767**	**0.348**	0.317	0.450
	REpep	0.582	0.073	0.190	0.107
	COACH	0.646	0.245	0.295	0.283

[a]The highest values are highlighted in bold type. The predictive quality on the training set TR640 is based on the overall performance of 5-fold cross validation on the TR640 set. REpep is a baseline method with predictions made based on evolutionary conservation analysis.

Compared to SVMpred, S-SITE and TM-SITE both achieve at least 2 times higher Rec values, but >1.9 times lower Pre values. The overall predictive performances on data sets TR640 and TE639 both indicate that SVMpep can compete with S-SITE, where the former provides better output of propensity scores with 0.02 more AUC and the latter provides more reliable binary prediction by obtaining at least 0.05 higher MCC. The ROC curves in Figure 2 show that S-SITE is better than SVMpep in the region of FP-rate $\leq$ 0.2 but worse when the FP-rate is >0.2. These data indicate that the *ab initio* method SVMpep is complementary to the template-based method S-SITE.

Comparing S-SITE to TM-SITE, we found that the former is surprisingly better than the latter, with at least 0.08, 0.09, 0.11, and 0.08 higher AUC, MCC, Rec, and Pre values, respectively. This is probably caused by the low quality of the structure model, as all templates with ≥30% sequence identity to the query have been removed during structure modeling.

3.2. Performance of the Consensus Method PepBind. Since the template-based methods and the *ab initio* method are complementary to each other, we investigated the predictive quality of the consensus method PepBind. Compared with all its components on the independent test set TE639, PepBind improves the AUC and MCC by at least 7.2 and 12.4%, respectively. Table 1 shows that PepBind can keep the relative high Rec value (∼0.3) from its template-based individual S-SITE and can maintain a Pre value around 0.44. The assessments on the data sets TR640 and TE639 both support that PepBind outperforms the other two consensuses between SVMpep and each template-based method, with at least 1.5 and 2.4% improvement in AUC and MCC, respectively. Figure 2 shows that the ROC curve of PepBind is consistently above the ROC curves for each individual and almost over the curves for the other two combinations. These data suggest that PepBind can improve the prediction with a relatively large margin. This improvement is not only due to the integration of *ab initio* predictor and template-based method but also attributed to the usage of the complementary template-based methods S-SITE and TM-SITE.

3.3. Performance of PepBind on Proteins from Different Families. The protein−peptide interaction is usually mediated by the short linear motifs, thus the peptide recognizable regions are generally short in length. This means that a given sequence interacted with peptides can be locally similar in peptide recognizable regions but dissimilar in the other regions. We therefore investigated the predictive performance of PepBind for the peptide-binding proteins from different protein families. We first mapped the proteins in the test set TE639 into the families of the Pfam database.[28] Then we evaluated the performance of PepBind on 16 families with at least 5 proteins. The results are summarized in Table 2. For most of the listed families, the predictions are in generally accurate with MCC higher than the overall MCC (i.e., 0.348 from Table 1). For example, for the FHA, PDZ, and SH3

DOI: 10.1021/acs.jcim.8b00019
J. Chem. Inf. Model. 2018, 58, 1459−1468

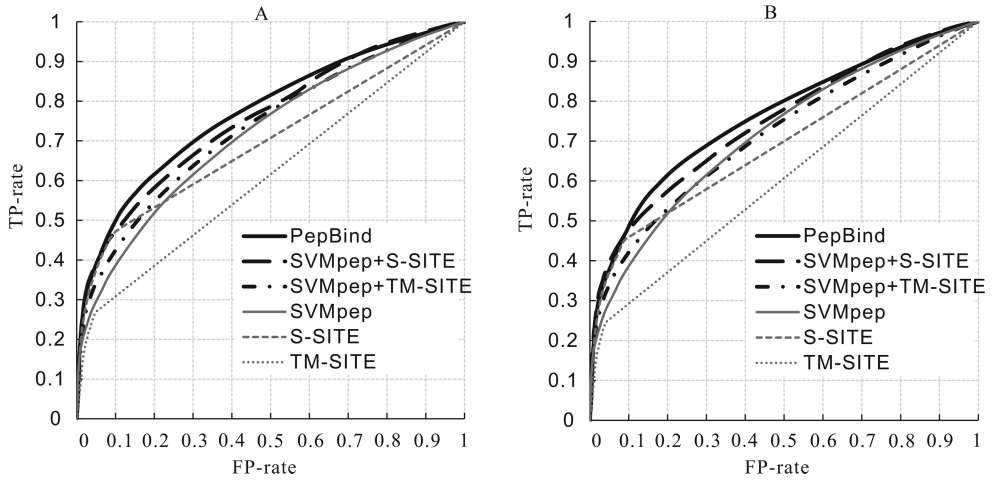


Figure 2. ROC curves for the different methods on the training set TR640 (A) and the independent test set TE639 (B).

Table 2. Predictive Quality of PepBind on Proteins from Different Families[a]

Pfam ID	family name	no. proteins	MCC	Rec	Pre	AUC
PF00498	FHA	6	0.782	0.741	0.851	0.984
PF00595	PDZ	11	0.765	0.741	0.842	0.934
PF00018	SH3	9	0.703	0.825	0.684	0.941
PF00104	Hormone_recep	6	0.684	0.658	0.739	0.934
PF00640	PID	5	0.628	0.710	0.647	0.936
PF00385	Chromo	6	0.626	0.632	0.790	0.867
PF00515	TPR_1	5	0.558	0.616	0.577	0.922
PF00017	SH2	21	0.538	0.576	0.603	0.888
PF00628	PHD	9	0.521	0.533	0.611	0.875
PF00856	SET domain	5	0.508	0.598	0.509	0.900
PF00076	RRM_1	5	0.500	0.519	0.583	0.842
PF00397	WW	9	0.470	0.553	0.517	0.812
PF00073	Rhv	5	0.445	0.536	0.395	0.742
PF00675	Peptidase_M16	5	0.384	0.405	0.400	0.853
PF05193	Peptidase_M16_C	5	0.384	0.405	0.400	0.853
PF00069	Pkinase	12	0.224	0.259	0.263	0.700

[a]The rows are sorted by the MCC values in descending order.

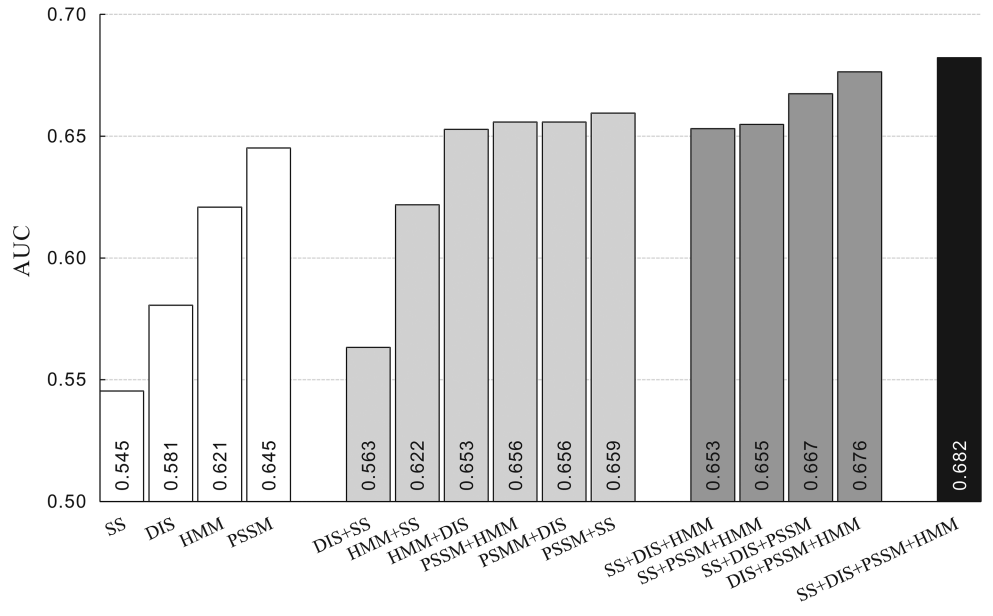


Figure 3. Overall AUC values for the SVM models built with different features. The overall AUC values were calculated by 5-fold cross validation on the training set TR128.

DOI: 10.1021/acs.jcim.8b00019
J. Chem. Inf. Model. 2018, 58, 1459−1468

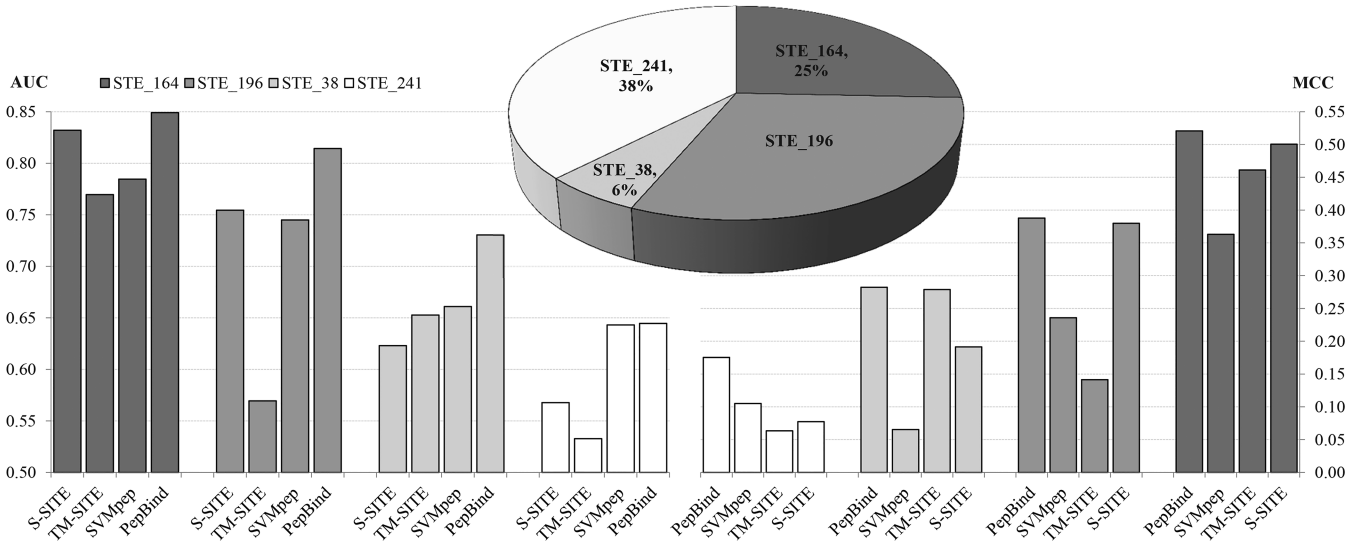


Figure 4. Contribution of the component methods to the consensus-based method PepBind. The four subsets STE_164, STE_196, STE_38, and STE_241, correspond to the 164 proteins with reliable templates derived from both S-SITE and TM-SITE, the 196 (respectively 38) proteins with reliable templates only derived from S-SITE (respectively TM-SITE), and 241 proteins without any reliable templates, respectively. The left and the right panels are the AUC and MCC values for all methods, respectively.

families, the MCC is above 0.7. These data suggest that, though PepBind is not specially trained on proteins of certain families, it could be applied in cross-family predictions with a reasonable accuracy.

3.4. Discriminative Quality for Binding to Other Ligands. Since proteins usually employ different sites to bind different ligands, it is of interest to investigate if the proposed methods are specific to the identification of peptide-binding residues. Here we tested PepBind on proteins binding to other ligands. We randomly extracted 30 carbohydrates-binding[13] (CBH30), 30 DNA-binding,[11] (DNA30) and 30 RNA-binding proteins[11] (RNA30) from previous studies. We ran PepBind on these data sets and collected the predictions. It shows that 0.3%, 1.2%, and 0.5% residues were predicted as peptide-binding residues for the CBH30, DNA30, and RNA30 data sets, respectively. On the other hand, the corresponding ratio for the data set TE80 is as high as 3.8%, which is 11.6, 3.2, and 8.4 times the values for the three different binding data sets. This indicates that the method PepBind is specific for the prediction of peptide-binding residues.

4. DISCUSSION

In this section, we investigate the contributions of the features to the *ab initio* method SVMpep, the contributions of the component methods to the consensus method PepBind. Comparisons with one baseline method, one general-purpose method, and two peptide-binding predictors are presented as well.

4.1. Feature Contributions to the *ab initio* Method SVMpep. The *ab initio* method SVMpep is built with four groups of features with rich evolutionary and structural information. The evolutionary information is inferred from the sequence profile by running the programs PSI-BLAST[16] and HHblits.[17] The structural information comprises the intrinsic disorder derived from the method IUPred,[19] and the secondary structure generated by SPIDER 2.0.[18]

We first investigated the contribution of each feature group for the prediction. In order to save the time used for training, we performed this analysis by 5-fold cross validation on the TR128 set. The results are summarized in Figure 3. We can see that all four feature groups are useful for the identification of peptide-binding residues. The SVM models built upon a single feature group range from 0.545 to 0.645 in AUC, where the profile-based feature group (PSSM and HMM) is more powerful than others (the white bars in Figure 3). Interestingly, the features of intrinsic disorder lead to higher AUC value than the secondary structure.

The AUC was improved by the combinations of different feature groups. Among all the six SVM models implemented by two combining two feature groups, the top two SVM models are both implemented by integration of profile and structure information. The one with the highest AUC combines the features from PSSM profile and secondary structure (AUC = 0.659) and the second utilizes the HMM profile and intrinsic disorder (AUC = 0.656). The incorporation of intrinsic disorder results to an improved AUC. In fact, except for the SVM model built upon secondary structure and intrinsic disorder, the other six SVM models improve their AUC values by 1.2−5.2%, after the inclusion of the intrinsic disorder information (the gray and black bars of Figure 3). In addition, since the sequence profile represented by HMM is different with PSSM, an increment of at least 1.6% in AUC is also obtained by combining them together. As shown in the darker gray and black bars in Figure 3, the SVM model further increase the AUC value to 0.682, by combining all features together. We therefore implement our *ab initio* method SVMpep by employing all the designed features.

4.2. Contributions of the Component Methods to the Consensus PepBind. We investigate the contributions of the component methods S-SITE, TM-SITE, and SVMpep to the consensus method PepBind. Based on if reliable templates (as judged by their respective confidence scores) could be detected by S-SITE and TM-SITE, the test set TE639 was divided into four nonoverlapping subsets (the pie chart in the middle of Figure 4): STE_164 (both are reliable), STE_196 (only S-SITE is reliable), STE_38 (only TM-SITE is reliable), and STE_241 (both are not reliable).

DOI: 10.1021/acs.jcim.8b00019
J. Chem. Inf. Model. 2018, 58, 1459−1468

Table 3. Comparison with Other Methods[a]

data set	method	AUC	MCC	Rec	Pre
Te125	Sprint-Seq	0.680	0.200	0.210	Na
	Sprint-Str	0.780	0.290	0.240	Na
	TM-Site	0.600	0.207	0.212	0.284
	S-Site	0.699	0.310	0.328	0.367
	SVMpep	0.770	0.340	0.196	**0.654**
	PepBind	**0.793**	**0.372**	**0.344**	0.469
Te80	Sprint-Seq*	0.692	0.131	0.639	0.089
	Sprint-Seq	0.680	Na	Na	Na
	TM-Site	0.597	0.195	0.208	0.259
	S-Site	0.660	0.270	0.290	0.320
	SVMpep	0.745	0.298	0.157	**0.626**
	PepBind	**0.758**	**0.337**	**0.316**	0.425
Bte80	Sprint-Seq*	0.694 ± 0.006	0.296 ± 0.012	0.639 ± 0.001	0.653 ± 0.005
	Sprint-Seq	0.711 ± 0.013	0.326 ± 0.005	**0.642 ± 0.015**	Na
	TM-Site	0.595 ± 0.004	0.261 ± 0.004	0.208 ± 0.001	0.850 ± 0.025
	S-Site	0.658 ± 0.003	0.344 ± 0.005	0.29 ± 0.001	0.890 ± 0.008
	SVMpep	0.747 ± 0.005	0.276 ± 0.004	0.157 ± 0.001	**0.964 ± 0.010**
	PepBind	**0.761 ± 0.006**	**0.393 ± 0.008**	0.325 ± 0.001	0.956 ± 0.005

[a]SPRINT-Seq and SPRINT-Str mean the results on the test set were cited from the corresponding publications. SPRINT-Seq* represents the method re-implemented by us.

The performance of all methods on these subsets is presented in Figure 4. When there are reliable templates from both S-SITE and TM-SITE, *i.e.*, on the subset STE_164, PepBind benefits from its individuals, by achieving at least 2.1% and 4.1% more AUC and MCC, respectively. When reliable templates are available for only S-SITE or TM-SITE, the *ab initio* component SVMpep always complements to the corresponding template-based method (S-SITE or TM-SITE) in AUC/MCC. Thus, PepBind was improved by 7.9% (respectively 10.5%) higher AUC and 2.1% (respectively 1.2%) higher MCC on the subset STE_196 (respectively STE_38). When no reliable templates are available for both methods, the *ab initio* predictor SVMpep is more accurate than both S-SITE and TM-SITE on the subset STE_241, with at least 13.3% and 35.6% higher AUC and MCC, respectively. The combination of these three methods contributes to 67.1% improvement in MCC for the consensus method PepBind. These data confirm that the three component methods are complementary to each other. When reliable templates are available, template-based methods S-SITE and TM-SITE have higher contribution to PepBind. On the contrary, when no reliable templates are detected, the *ab initio* method SVMpep contributes more to PepBind.

4.3. Comparison with the Baseline Method REpep and the General-Purpose Method COACH. Evolutionary conservation is usually an indicator for the functionally important residues. We employ the relative entropy (RE) derived from PSSM to measure the conservation and infer the peptide-binding residues based on the RE values for each residue. This baseline method is named as REpep. We also compared our methods with the general-purpose template-based method COACH.[15]

The results are summarized in Table 1. We can see that the prediction by REpep is almost random in terms of MCC value, though its AUC value is about 0.6. Contrarily, the method COACH performs better with ∼0.65 AUC and 0.25 MCC on both the training and the test sets. On the training set TR640, the consensus method PepBind outperforms COACH with 17% and 30% higher AUC and MCC, respectively. On the test set TE639, the improvement increases to 19% and 42% in AUC and MCC, respectively.

4.4. Comparison with Existing Predictors of Peptide-Binding Residues. Most recently, a sequence-based method SPRINT-Seq[10] and a structure-based method SPRINT-Str[9] were developed for the peptide-binding residue prediction. Both of them were shown to outperform other methods. We thus compared them with SVMpep and PepBind on the test data sets TE80 and TE125, where TE80 is a subset of TE125. Since the TE125 set comprises 24 sequences from our training set TR640, we retrained SVMpep and PepBind by replacing the training set TR640 with SPRINT-Str's training set for fair comparison. The updated SVMpep and PepBind methods were then assessed on the test sets TE80 and TE125.

Note that disorder information is used in our methods. It may be unfair to compare with the method SPRINT-Str on the fully disordered proteins (FDPs), which are intrinsically disordered, but with induced folding through binding to peptides. We checked how many FDPs exist in the training and test sets based on the disorder predictions from IUPred. Here a protein is defined as an FDP if more than 90% of its residues are predicted as disordered. It turns out that there are only four FDPs in SPRINT-Str's training set and none in the test sets TE125 and TE80. This is anticipated because all data sets used here were directly obtained from the work of SPRINT-Str and SPRINT-Seq. Thus, it should be fair to make comparisons on the data sets TE125 and TE80.

The results are summarized in Table 3, where the data for other methods were cited from previously published results of SPRINT-Str. It is of interest to compare SVMpep with SPRINT-Seq, as both are sequence-based methods and do not use templates. SPRINT-Seq was trained on a balanced data set, by randomly selecting the same number of nonbinding residues as binding residues. Following this sampling procedure, we first generated a balanced set from TE80 (denoted by BTE80) and next evaluated SVMpep on this set. The results show that SVMpep achieves 7.6% higher AUC but 6.8% lower MCC, compared to the method SPRINT-Seq. The lower binary predictive performance can be explained by the

DOI: 10.1021/acs.jcim.8b00019
J. Chem. Inf. Model. 2018, 58, 1459−1468

design of SVMpep, which was trained on the original unbalanced data set and for real-world application. Actually, the evaluations on TE80 and TE125 both reveal that SVMpep achieves 7.7% higher AUC and 70% higher MCC than SPRINT-Seq. The ROC curve of SVMpep is also over the curve of SPRINT-Seq on the TE125 set (see Supporting Figure S3). These data suggest that SVMpep outperforms SPRINT-Seq for the prediction of natural peptide-binding residues. This benefits from the training strategy with the full set of nonbinding residues and the utilization of intrinsic disorder features. From Table 3, it is interesting to see that the template-program S-SITE outperforms SPRINT-Seq on the TE125 set, by achieving at least 0.02, 0.11, 0.12, and 0.37 higher AUC, MCC, Rec, and Pre, respectively. This further reflects that SPRINT-Seq does not work well on the original unbalanced data set.

The prediction of peptide-binding residues was improved by the recent structure-based method SPRINT-Str. Table 3 shows that SVMpep has 0.01 less AUC but 0.05 more MCC than this method on the independent test set TE125. Supporting Figure S3 shows that the ROC curve of SVMpep is slightly over the curve of SPRINT-Str when the FP-rate is ≤25%. These data suggest that our sequence-based method SVMpep can even compete with the latest structure-based method SPRINT-Str, especially for the low FP-rate region. In addition, the consensus-based method PepBind significantly outperforms the methods SPRINT-Seq and SPRINT-Str on the independent test sets TE80 and TE125. Specifically, Table 3 shows that the AUC and MCC of PepBind is 0.07 higher than and 2.41 times to the corresponding values of SPRINT-Seq on the TE80 set. On the TE125 set, PepBind also performs better than both SPRINT-Seq and SPRINT-Str, with the improvement of AUC, MCC, and Rec of 1.7−16.6%, 28.3−86.0%, and 43.3−63.8%, respectively. Supporting Figure S3 suggests that the ROC curve of PepBind is over the curve of SPRINT-Str with a relatively large margin for the region of FP-rate ≤ 50% and is consistently above the curve of SPRINT-Seq.

To summarize, both the *ab initio* predictor SVMpep and the consensus-based method PepBind provide accurate prediction for the identification of peptide-binding residues, no matter the structure state of the query proteins. Furthermore, the consensus method PepBind outperforms the existing methods SPRINT-Seq and SPRINT-Str with a large margin. The success of the consensus-based method PepBind is attributed to the introduction of intrinsic disorder in its ab initio individual SVMpep, and the integration of this method with the two template-based algorithms S-SITE and TM-SITE.

5. CONCLUSIONS

We developed two sequence-based methods for accurate prediction of peptide-binding residues. The first is a novel *ab initio* method SVMpep by using a comprehensive set of designed features from intrinsic disorder, predicted secondary structure and two sequence profiles. Interestingly, the intrinsic disorder information is a powerful indicator, especially when incorporating it into the profile information. The inclusion of intrinsic disorder in the *ab initio* method SVMpep results to an enhanced accuracy. The second is a consensus method PepBind, which combines SVMpep with two template-based methods S-SITE and TM-SITE. As shown in our assessments, PepBind significantly outperforms both the sequence-based method SPRINT-Seq and the structure-based method SPRINT-Str. The success is attributed to the consideration of intrinsic disorder in SVMpep, and the combination with two complementary template-based methods S-SITE and TM-SITE.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00019.

Data for parameter optimization and ROC comparison with other methods (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: zhenling@tju.edu.cn (Z.P.).
*E-mail: yangjy@nankai.edu.cn (J.Y.).

ORCID

Jianyi Yang: 0000-0003-2912-7737

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work was supported in part by National Natural Science Foundation of China (NSFC 11501407 and 11501306), the Seed Foundation of Tianjin University (1701), Fok Ying-Tong Education Foundation (161003), Fundamental Research Funds for the Central Universities, and the Thousand Youth Talents Plan of China.

■ REFERENCES

(1) Huang, Z. The chemical biology of apoptosis. Exploring protein-protein interactions and the life and death of cells with small molecules. *Chem. Biol.* **2002**, *9*, 1059−1072.

(2) Klug, A. Zinc finger peptides for the regulation of gene expression. *J. Mol. Biol.* **1999**, *293*, 215−218.

(3) Dalrymple, B. P.; Kongsuwan, K.; Wijffels, G.; Dixon, N. E.; Jennings, P. A. A universal protein-protein interaction motif in the eubacterial DNA replication and repair systems. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 11627−11632.

(4) Stanfield, R. L.; Wilson, I. A. Protein-peptide interactions. *Curr. Opin. Struct. Biol.* **1995**, *5*, 103−113.

(5) Yang, J.; Roy, A.; Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **2013**, *41*, D1096−1103.

(6) Petsalaki, E.; Stark, A.; Garcia-Urdiales, E.; Russell, R. B. Accurate prediction of peptide binding sites on protein surfaces. *PLoS Comput. Biol.* **2009**, *5*, e1000335.

(7) Lavi, A.; Ngan, C. H.; Movshovitz-Attias, D.; Bohnuud, T.; Yueh, C.; Beglov, D.; Schueler-Furman, O.; Kozakov, D. Detection of peptide-binding sites on protein surfaces: the first step toward the modeling and targeting of peptide-mediated interactions. *Proteins: Struct., Funct., Genet.* **2013**, *81*, 2096−2105.

(8) Yan, C.; Zou, X. Predicting peptide binding sites on protein surfaces by clustering chemical interactions. *J. Comput. Chem.* **2015**, *36*, 49−61.

(9) Taherzadeh, G.; Zhou, Y. Q.; Liew, A. W. C.; Yang, Y. D. Structure-based prediction of protein−peptide binding regions using Random Forest. *Bioinformatics* **2018**, *34*, 477−484.

(10) Taherzadeh, G.; Yang, Y.; Zhang, T.; Liew, A. W.; Zhou, Y. Sequence-based prediction of protein−peptide binding sites using support vector machine. *J. Comput. Chem.* **2016**, *37*, 1223−1229.

(11) Yan, J.; Kurgan, L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res.* **2017**, *45*, e84.

(12) Meng, Q.; Peng, Z.; Yang, J. CoABind: a novel algorithm for Coenzyme A (CoA)- and CoA derivatives-binding residues

DOI: 10.1021/acs.jcim.8b00019
J. Chem. Inf. Model. 2018, 58, 1459−1468

prediction. *Bioinformatics* **2018**, DOI: 10.1093/bioinformatics/bty162.
(13) Taherzadeh, G.; Zhou, Y.; Liew, A. W.; Yang, Y. Sequence-Based Prediction of Protein-Carbohydrate Binding Sites Using Support Vector Machines. *J. Chem. Inf. Model.* **2016**, *56*, 2115−2122.
(14) Weatheritt, R. J.; Gibson, T. J. Linear motifs: lost in (pre)translation. *Trends Biochem. Sci.* **2012**, *37*, 333−341.
(15) Yang, J.; Roy, A.; Zhang, Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, *29*, 2588−2595.
(16) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(17) Remmert, M.; Biegert, A.; Hauser, A.; Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **2012**, *9*, 173−175.
(18) Heffernan, R.; Paliwal, K.; Lyons, J.; Dehzangi, A.; Sharma, A.; Wang, J.; Sattar, A.; Yang, Y.; Zhou, Y. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep.* **2015**, *5*, 11476.
(19) Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **2005**, *21*, 3433−3434.
(20) Xia, J.; Peng, Z.; Qi, D.; Mu, H.; Yang, J. An ensemble approach to protein fold classification by integration of template-based assignment and support vector machine classifier. *Bioinformatics* **2017**, *33*, 863−887.
(21) Peng, Z.; Kurgan, L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.* **2015**, *43*, e121.
(22) Meng, F.; Uversky, V.; Kurgan, L. Computational Prediction of Intrinsic Disorder in Proteins. *Curr. Protoc. Protein Sci.* **2017**, *88*, 2.16.1−2.16.4.
(23) Hanson, J.; Yang, Y.; Paliwal, K.; Zhou, Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **2017**, *33*, 685−692.
(24) Zhang, T.; Zhang, H.; Chen, K.; Ruan, J. S.; Shen, S. Y.; Kurgan, L. Analysis and Prediction of RNA-Binding Residues Using Sequence, Evolutionary Conservation, and Predicted Secondary Structure and Solvent Accessibility. *Curr. Protein Pept. Sci.* **2010**, *11*, 609−628.
(25) Wang, K.; Samudrala, R. Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinf.* **2006**, *7*, 385.
(26) Robinson, A. B.; Robinson, L. R. Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1991**, *88*, 8880−8884.
(27) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **2015**, *12*, 7−8.
(28) Finn, R. D.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Mistry, J.; Mitchell, A. L.; Potter, S. C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; Salazar, G. A.; Tate, J.; Bateman, A. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*, D279−285.

DOI: 10.1021/acs.jcim.8b00019
J. Chem. Inf. Model. 2018, 58, 1459−1468