Received: 21 June 2020 | Revised: 22 December 2020 | Accepted: 23 December 2020

DOI: 10.1002/jmr.2887

RESEARCH ARTICLE

Journal of Molecular Recognition WILEY

SNB-PSSM: A spatial neighbor-based PSSM used for protein–RNA binding site prediction

Yang Liu | Weikang Gong | Zhen Yang | Chunhua Li

Faculty of Environmental and Life Sciences, Beijing University of Technology, Beijing, China

Correspondence
Chunhua Li, Faculty of Environmental and Life Sciences, Beijing University of Technology, Beijing 100124, China.
Email: chunhuali@bjut.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 11474013, 31971180

Abstract

Protein–RNA interactions play essential roles in a wide variety of biological processes. Recognition of RNA-binding residues on proteins has been a challenging problem. Most of methods utilize the position-specific scoring matrix (PSSM). It has been found that considering the evolutionary information of sequence neighboring residues can improve the prediction. In this work, we introduce a novel method SNB-PSSM (spatial neighbor–based PSSM) combined with the structure window scheme where the evolutionary information of spatially neighboring residues is considered. The results show our method consistently outperforms the standard and smoothed PSSM methods. Tested on multiple datasets, this approach shows an encouraging performance compared with RNABindRPlus, BindN+, PPRInt, xypan, Predict_RBP, SpaPF, PRNA, and KYG, although is inferior to RNAProSite, RBscore, and aaRNA. In addition, since our method is not sensitive to protein structure changes, it can be applied well on binding site predictions of modeled structures. Thus, the result also suggests the evolution of binding sites is spatially cooperative. The proposed method as an effective tool of considering evolutionary information can be widely used for the nucleic acid–/protein-binding site prediction and functional motif finding.

KEYWORDS
binding site prediction, position-specific scoring matrix, protein–RNA interfaces, spatial neighbor

1 | INTRODUCTION

Protein–RNA interactions play critical roles in a wide variety of biological processes.[1] Reliable identification of RNA-binding residues on proteins is an important and challenging problem, which is critical for understanding protein–RNA recognition mechanisms, and helpful for complex structure prediction and drug design.

Many predictors based on sequence and structure have been developed and reviewed in literatures over the past several years.[2] The protein sequence evolutionary information is a very effective feature for binding site prediction.[3] The Position-Specific Scoring Matrix (*PSSM*), a common representation of evolutionary features, has been *widely used* in most of the predictors.[2] Murakami et al presented a support vector machine (SVM) classifier (PiRaNhA) that utilizes the standard PSSM combined with physical and chemical properties of residues to predict RNA-binding residues.[4] Since surrounding residues of binding sites usually affect the binding process, it is necessary to incorporate the evolutionary information of the considered site and its surrounding ones. Thus, EL-Manzalawy adopted a sliding sequence window (size of 25) scheme to encode the evolutionary information of the target residue, achieving a higher Matthews correlation coefficient (*MCC*) compared with some other state-of-the-art predictors.[5] Cheng et al made a smooth processing (called smoothed PSSM) before coding the evolutionary information in a sliding window form,[6] in which the evolutionary score for a residue position is replaced by the sum of evolutionary scores of its seven neighboring residues in sequence. This method achieves an evident improvement in binding site prediction. From

the above, most of the methods prefer to utilize a sliding window along sequence to represent the evolutionary characteristic of the central residue. However, we think considering the features of spatial neighbors of a target residue will get a better effect than that of sequence neighbors in binding site prediction.

In previous works, researchers have tried to consider the spatial neighbor features of a target residue in prediction. Wang et al used residue spatial sequence profile to predict binding sites in protein–protein heterocomplexes.[7] Chen et al integrated spatial adjacent residue information and structure information for RNA-binding residue prediction.[8] Tang et al utilized sequence and structure characteristics encoded in a structural window to predict RNA-binding residues.[9] Studies have shown the conserved interface residues often occur clustered together in tertiary structures.[10] This tendency holds true for protein–protein/nucleic acid interactions.[11,12] The higher packing density of conserved residues at interfaces and within enzyme active sites may suggest their cooperativity in function exertion.[10] That the conserved residues form one or more localized clusters within interfaces or tertiary structures will facilitate the formation of some "functional motifs." Thus, based on the above, we think the PSSM profile encoded in a form of spatial neighbors can better reflect the evolutionary characteristics of interface conserved residues than that encoded in a form of sequence neighbors.

In this paper, we propose a new coding scheme of PSSM profile based on the spatially neighboring residues to predict RNA-binding residues. A smooth processing of PSSM profile is introduced into the method. As an application, it is implemented to predict RNA-binding sites on proteins.

2 | MATERIALS AND METHODS

2.1 | Protein–RNA datasets

In this work, we used benchmark dataset RB198[3] as training set and RB44[13] as test set, respectively. In order to compare our method with the other ones, we used benchmark dataset RB111[5] as an independent verification set.

(1) RB198

The data in RB198 were derived from the Protein Data Bank (PDB) by picking up all protein–RNA complexes[14] and then removing the complexes that meet any one of the following: (i) structure resolution worse than 3.5 Å; (ii) protein residues <40 or RNA nucleotides <5; (iii) interface residues <3; and (iv) protein sequence identity >30% with others. The RB198 dataset obtains 134 complexes with 198 protein chains.

(2) RB44

RB44 dataset contains 44 protein chains that have at most 40% sequence identity.[3,13]

(3) RB111

RB 111 has 111 protein chains that share less than 30% sequence similarity with those in RB44.[15]

For all the datasets, an interface residue is defined as the one where at least one atom is closer than 5 Å to any atom of RNA. Table S1 provides the data information.

2.2 | Spatial neighbor–based position-specific scoring matrix (SNB-PSSM)

In order to consider the evolution of surrounding residues around a target one, we propose a new spatial neighbor–based PSSM (SNB-PSSM) method that is different from the smoothed PSSM mentioned above.

In SNB-PSSM, first, for a protein sequence, a standard PSSM profile is generated by PSI-BLAST[16] against the nonredundant (nr) protein sequence database through three iterations with 0.001 as E-value cutoff. For a protein with N residues, the size of PSSM matrix is $20 \times N$ with the evolutionary information for each position encapsulated in a vector of 20 dimensions. Then, the evolutionary score of a target residue is defined as an average value of the evolutionary scores from the standard PSSM over the residues whose Cα atoms are within 7.5 Å from that of the target one. Figure 1 illustrates the definition process.

2.3 | Evolutionary information encoded in a sliding structure window form

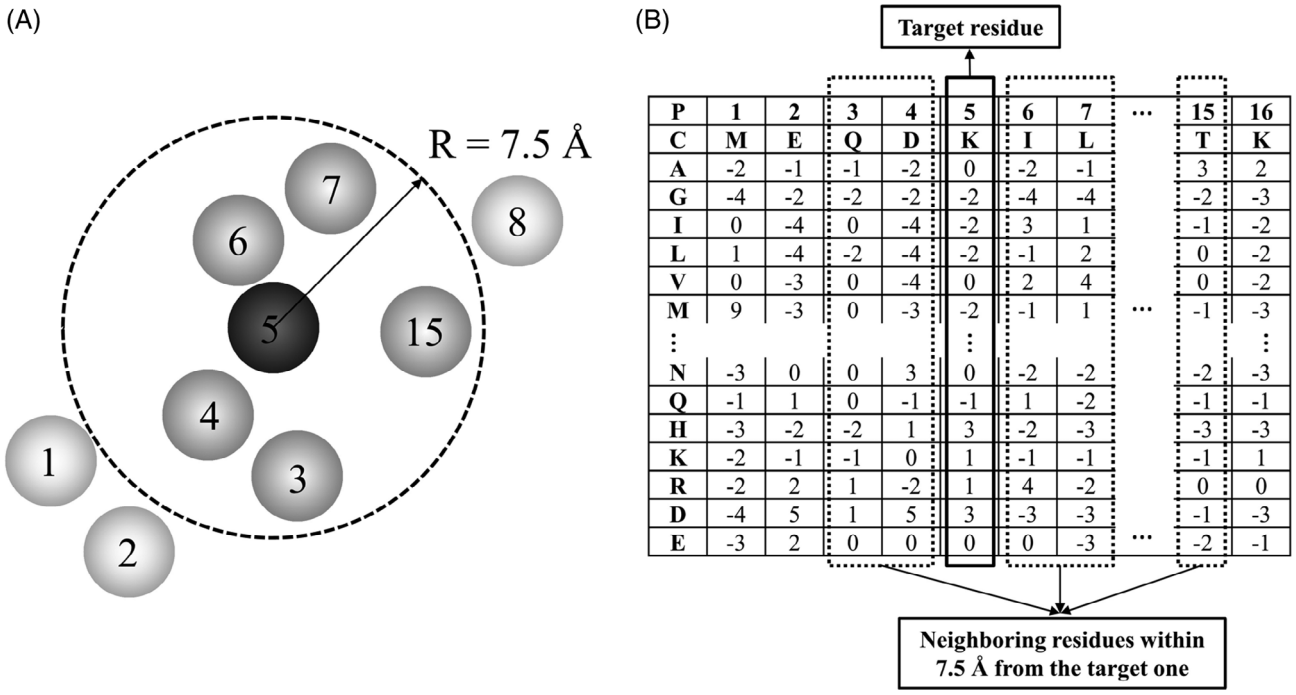
For a protein structure, after its SNB-PSSM profile is obtained, the evolutionary information of a target residue is encoded using the evolutionary scores (from SNB-PSSM) of the spatially nearest w residues (a sliding structure window of size w) to the target one (including the target one). Thus, for a target residue, its evolutionary information is encoded into a $20 \times w$ matrix.

2.4 | Support vector machine (SVM) classifier

Support vector machine[17] is used as the classification method. The package LIBSVM (version 3.0. http://www.csie.ntu.edu.tw/∼cjlin/libsvm/) is used, and the radial basis function (RBF) is chosen as the kernel function. The regularization parameter C and kernel width parameter γ are optimized until an optimal SVM model is obtained.

2.5 | Performance evaluation

Our method is trained through fivefold cross validation on RB198 dataset, tested on RB44 dataset, and compared with other methods on the independent dataset RB111. We assess the performance of classifiers using the overall accuracy (*ACC*), sensitivity (*SN*), specificity (*SP*), and Matthews correlation coefficient (*MCC*) that are defined as follows:



P	1	2	3	4	5	6	7	…	15	16
C	M	E	Q	D	K	I	L		T	K
A	-2	-1	-1	-2	0	-2	-1		3	2
G	-4	-2	-2	-2	-2	-4	-4		-2	-3
I	0	-4	0	-4	-2	3	1		-1	-2
L	1	-4	-2	-4	-2	-1	2		0	-2
V	0	-3	0	-4	0	2	4		0	-2
M	9	-3	0	-3	-2	-1	1	…	-1	-3
⋮					⋮					⋮
N	-3	0	0	3	0	-2	-2		-2	-3
Q	-1	1	0	-1	-1	1	-2		-1	-1
H	-3	-2	-2	1	3	-2	-3		-3	-3
K	-2	-1	-1	0	1	-1	-1		-1	1
R	-2	2	1	-2	1	4	-2		0	0
D	-4	5	1	5	3	-3	-3		-1	-3
E	-3	2	0	0	0	0	-3	…	-2	-1

FIGURE 1 Process of computing the evolutionary score of a target residue in SNB-PSSM method. (A) The residue 5′s spatially neighbors (3, 4, 5, 6, 7, and 15) whose Cα atoms are within 7.5 Å from its Cα atom. (B) The evolutionary score of residue 5 is defined as an average score of evolutionary scores (a vector of 20 dimensions) from the standard PSSM over its spatially neighbors 3, 4, 5, 6, 7, and 15

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

$$SN = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

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

where the true positive (*TP*), false positive (*FP*), true negative (*TN*), and false negative (*FN*) are obtained by comparing the predicted label for each residue with the actual one.

3 | RESULTS AND DISCUSSION

3.1 | Prediction effectiveness of SNB-PSSM

In order to examine the capabilities to identify RNA-binding sites of the three types of PSSMs: standard PSSM, smoothed PSSM, and SNB-PSSM, fivefold cross-validation was used to train them on RB198 dataset (Table S2), and then the tests were performed on RB44 dataset (Table S3). From Table S3, SNB-PSSM attains an overall accuracy of 0.70, higher than 0.67 and 0.62 by the standard and smoothed PSSMs, respectively. Moreover, SNB-PSSM achieves improvements of 39% and 19% in *MCC* compared with the two PSSMs, respectively, and meanwhile, it also acquires the highest specificity although it does not reach the highest sensitivity.

Next, we want to detect the effects of encoding schemes on prediction results. Three encoding ways were constructed: the standard and smoothed PSSMs with a sliding sequence window of size 7, respectively, and the SNB-PSSM with a sliding structure window of the same size. Their performances were tested on RB44 dataset, with the results listed in Table S4. From Table S4, the SNB-PSSM-based way achieves the highest *SP*, *ACC*, and *MCC* with the values 0.75, 0.69, and 0.34, respectively.

In order to optimize the performance of the SNB-PSSM-based structure window method, the size of window needs to be set properly. Through comparing the performances (Tables S5 and S6) of different structure window size sets, 25 was selected as the final size.

The results above indicate it is more effective to consider the evolutionary information in a form of spatial neighbors than to consider it in a form of sequence neighbors for binding site prediction.

3.2 | Comparison with existing protein–RNA interface prediction methods

About the idea of spatial neighbors, it has been provided by previous works. In the work of Wang et al[7]/Chen et al,[8] a profile of a target residue is produced based on the HSSP database[18]/PSSM and is encoded into a vector of 20 × 11/20 × 15 elements where a window of 11/15 spatially neighboring residues is adopted. Chen et al applied their residue profile SpaPF to RNA-binding site prediction. Different

from the above methods, here a spatial neighbor-based smooth processing is performed based on the PSSM before encoding evolutionary information with a structure window scheme, and additionally, the structure window size 25 is adopted through an optimizing process, which both contribute the improvement of our method compared with SpaPF (Table 1). In addition, PSSM (internally computed by PSI-BLAST) is better than HSSP (based on precomputed multiple alignments) mainly because of the former's better quality of the sequence alignments.[19]

Additionally, our method was compared with the four sequence-based prediction servers (FastRNABindR,[5] RNABindR v2,[3] BindN+[20] and PPRInt[21]), and the two structure-based ones (KYG[22] and PRIP[23]) on the independent test set RB111, with the results shown in Table 1. Table S7 gives the descriptions on the servers. From Table 1, our method attains the highest *ACC*, *SP*, and *MCC* of 0.85, 0.88, and 0.26, respectively. Its sensitivity is not as good as some prediction servers. For an imbalanced data, *MCC* is often considered as a more balanced evaluation of performances.[24] Next, we detected our method's performances on four categories of protein chains with different lengths (short, medium, medium long, and long), with the results displayed in Table S8 (Table S9 for details). The results show it consistently outperforms FastRNABindR, RNABindR v2, and SpaPF (available severs or programs) for all categories, and with the increase of sequence length, the improvement has a more evident tendency, especially compared with FastRNABindR, which we think mainly attributes to its consideration of spatially neighbor information as the binding sites of longer chain proteins are more likely composed of the residues far away in sequence.

TABLE 1 Performance comparison of our method with the other seven methods on RB111 dataset

Methods	*SN*	*SP*	*ACC*	*MCC*
Sequence-based methods				
FastRNABindR	0.61	0.76	0.75	0.24
RNABindR v2	**0.63**	0.73	0.72	0.22
BindN+	0.43	0.87	0.84	0.24
PPRInt	0.48	0.79	0.76	0.18
Structure-based methods				
SpaPF	0.42	0.86	0..82	0.21
KYG	0.47	0.80	0.78	0.19
PRIP	0.45	0.78	0.75	0.15
Our method (on experimental structures)	0.44	**0.88**	**0.85**	**0.26**
Our method (on modeled structures)	0.43	0.87	0.85	0.25
TM-score < 0.5	0.43	0.86	0.84	0.23
0.5 ≤ TM-score < 0.7	0.43	0.87	0.85	0.25
TM-score ≥ 0.7	0.45	0.87	0.86	0.26

For better benchmarking, we compared our method's performance with those of the five sequence-based methods RNABindRPlus,[20] BindN+, PPRInt, xypan,[25] and Predict_RBP,[26] and six structure-based ones SpaPF, RBscore,[27] RNAProSite,[28] aaRNA,[29] PRNA[30] and KYG on datasets RNABindR_R106, PRNA_R205, SRCPred_R160, aaRNA_R205, RBscore_R116, RNABindR_R111, and aaRNA_R67 which are some relatively big datasets in NBench website,[31] and the results (*MCC*) are shown in Table 2. Generally, the predictive ability of a method is always assessed by the lowest accuracy on all the datasets rather than the best or average accuracy.[32] Thus, using the lowest *MCC* to compare the programs, from Table 2, our method is better than all of the five sequence-based approaches. Additionally, compared with the six structure-based methods, our method presents a stronger power than SpaPF, PRNA and KYG and is inferior to RNAProSite, RBscore, and aaRNA. From the above, generally our method has a few advantages to some extent, especially compared with the sequence-based methods. In the future, it is desirable

TABLE 2 Prediction performances (*MCC*) of our method and other sequence-based and structure-based ones on seven datasets

Methods	RNABindR_R106	PRNA_R205	SRCPred_R160	aaRNA_R205	RBscore_R116	RNABindR_R111	aaRNA_R67
Sequence-based methods							
RNABindRPlus	0.78	0.66	0.68	0.61	0.42	0.24	0.34
BindN+	0.36	0.32	0.36	0.32	0.23	0.24	0.20
PPRInt	0.51	0.42	0.47	0.38	0.22	0.19	0.28
xypan	0.76	0.78	0.71	0.59	0.31	0.25	0.21
Predict_RBP	0.70	0.63	0.68	0.48	0.18	0.03	0.05
Structure-based methods							
SpaPF	0.69	0.51	0.41	0.39	0.32	0.21	0.23
RBscore	0.44	0.40	0.44	0.44	0.34	0.37	0.48
RNAProSite	0.56	0.52	0.52	0.57	0.43	0.38	0.35
aaRNA	0.50	0.46	0.48	0.48	0.38	0.34	0.42
PRNA	0.74	0.82	0.67	0.50	0.23	0.19	0.18
KYG	0.33	0.29	0.32	0.32	0.25	0.20	0.29
Our method	0.80	0.64	0.48	0.45	0.37	0.26	0.32

to further discuss the combination of SNB-PSSM with sequence-based and structure-based features for protein-RNA binding site prediction.

3.3 | Case studies with our method

Taking two cases from RB111, for example, we give the detailed performance of our method. The first is an RNA-binding protein "cytotoxic domain of colicin E3" (PDB ID: 2XFZ:Y).[33] Our method predicts 27 binding residues correctly with *SN*, *SP*, *ACC*, and *MCC* of 0.79, 0.76, 0.77, and 0.54, respectively (see Figure 2C). As a comparison, we give the prediction by the standard PSSM-based method and the smoothed one (window size of 25). In contrast, 20 and 25 binding residues are identified correctly, respectively, with *SN*, *SP*, *ACC*, *MCC* of 0.59, 0.65, 0.63, 0.23 for the former (see Figure 2A) and 0.74, 0.54, 0.61, 0.26 for the latter (see Figure 2B). The second is a ternary NusB-NusE-BoxA RNA complex.[34] Our method was performed on the RNA-binding protein NusB (PDB ID: 3R2C:A) which has 38 binding residues. Our method predicts 33 correctly (with *SN*, *SP*, *ACC*, *MCC* of 0.87, 0.72, 0.76, 0.53) (see Figure 2F), while it is 25 for the standard PSSM-based method (with 0.66, 0.70, 0.69, 0.33) (see Figure 2D) and 36 for the smoothed one (0.95, 0.51, 0.63, 0.42) (see Figure 2E).

From Figure 2, our method achieves the highest *ACC* and *MCC* values. Also, it can be seen the false-positive residues from our method are more clustered spatially around the predicted true positive sites than those from the other two PSSM-based methods. It is understandable because the spatially neighbors have similar evolutionary information encodings in our method, and therefore they are more likely predicted as the same results, while these residues are probably far away from each other in sequence, and thus their encodings will be largely different in the two methods, leading to different prediction results.

3.4 | Binding site prediction on unbound proteins

In actuality, we need to make predictions on unbound protein structures. Figure 3 shows an example of prediction with our method on the unbound structure (PDB ID: 3SXL:A) and bound one (PDB ID: 1B7F:A) of an RNA-binding protein "*Drosophila melanogaster* sex-lethal protein."[35] From Figure 3, the conformational changes happen mainly on the RNA-binding interface, and the two structures' root mean square deviation (RMSD) of backbone atoms is relatively large with the value 6.7 Å ranked fifth out of 71 structures in benchmark1.0.[36] From Figure 3A, the prediction result on the unbound is satisfactory with SN, SP, ACC, and MCC of 0.42, 0.89, 0.82, and 0.30, respectively, although the evaluation has a drop compared with that on the bound one with SN, SP, ACC, and MCC of 0.75, 0.78, 0.78, and 0.41 (see Figure 3B). Similarly, the predicted false-positive residues of the two structures are all nearby the true binding sites.

3.5 | Binding site prediction on the modeled structures

We want to know our method's performance on modeled structures. We conducted structure modeling with I-TASSER method (threading

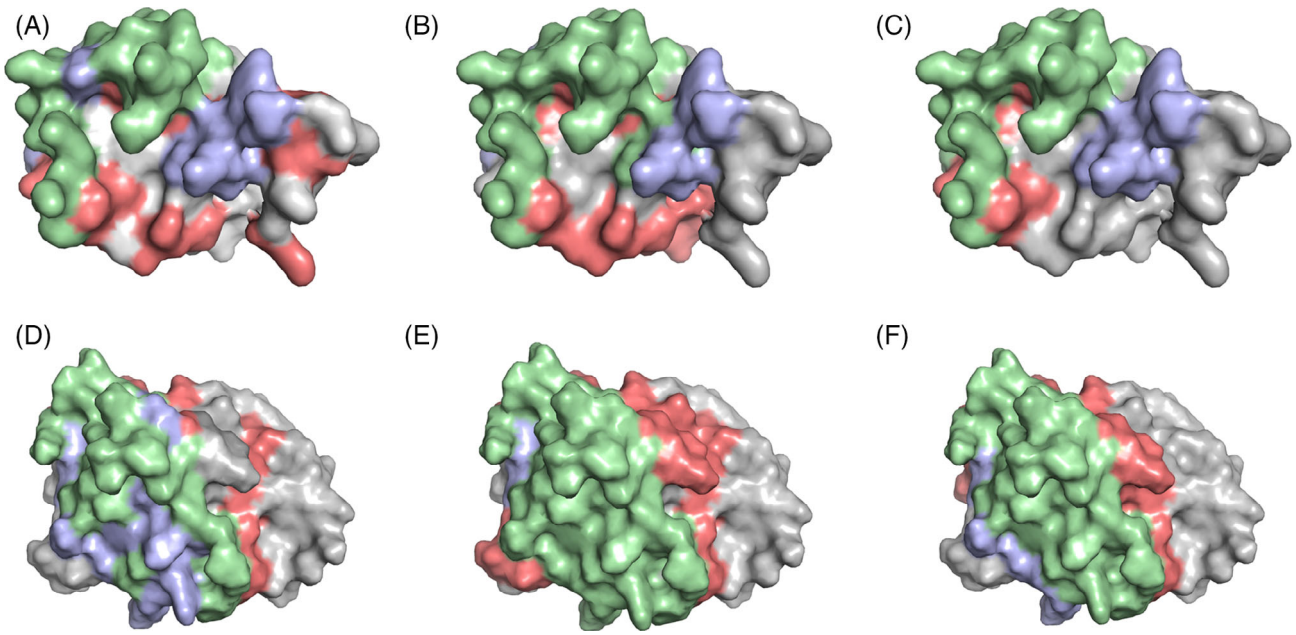


FIGURE 2 Predictions of RNA-binding residues on proteins with the standard PSSM– and smoothed PSSM–based sequence window, and the SNB-PSSM–based structure window methods, respectively. (A) (B) and (C) RNA binding protein "cytotoxic domain of colicin E3" (PDB ID: 2XFZ:Y). (D) (E) and (F) RNA-binding protein NusB in the ternary NusB-NusE-BoxA RNA complex (PDB ID: 3R2C:A). The *TP*, *FP*, and *FN* results are shown in green, red, and marine, respectively

based) developed by Zhang's lab that performs pretty well even on the new fold targets[37] and binding site prediction for all cases in RB111, with the results shown in Figure S1 and Table 1. In modeling, all the templates with sequence identity >30% to the query one are excluded from the template library. From Figure S1, most of the models (84.7%) have a TM-score[38] above 0.5, and the average value

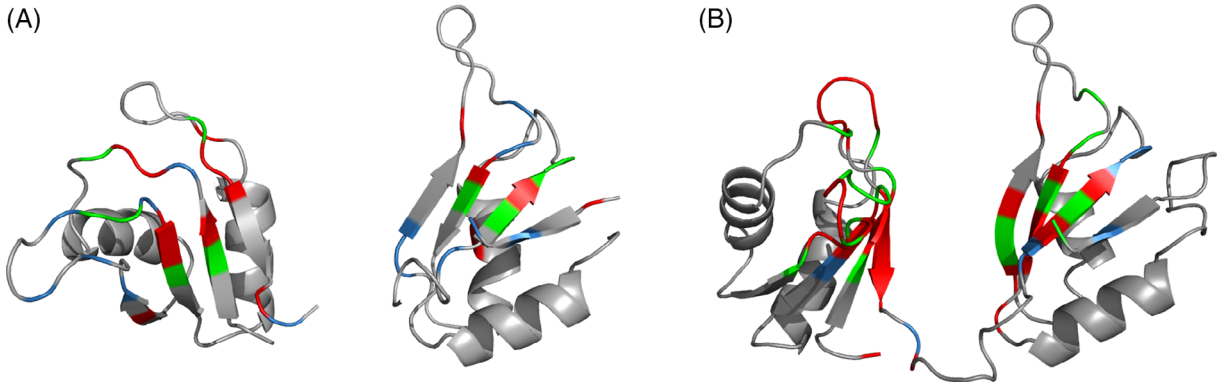


FIGURE 3 Prediction of RNA-binding residues for protein "*Drosophila melanogaster* sex-lethal protein" by our method on the unbound structure (PDB ID: 3SXL:A (A), and on the bound one (PDB ID: 1B7F:A) (B). The *TP*, *FP*, and *FN* results are shown in green, red, and marine, respectively

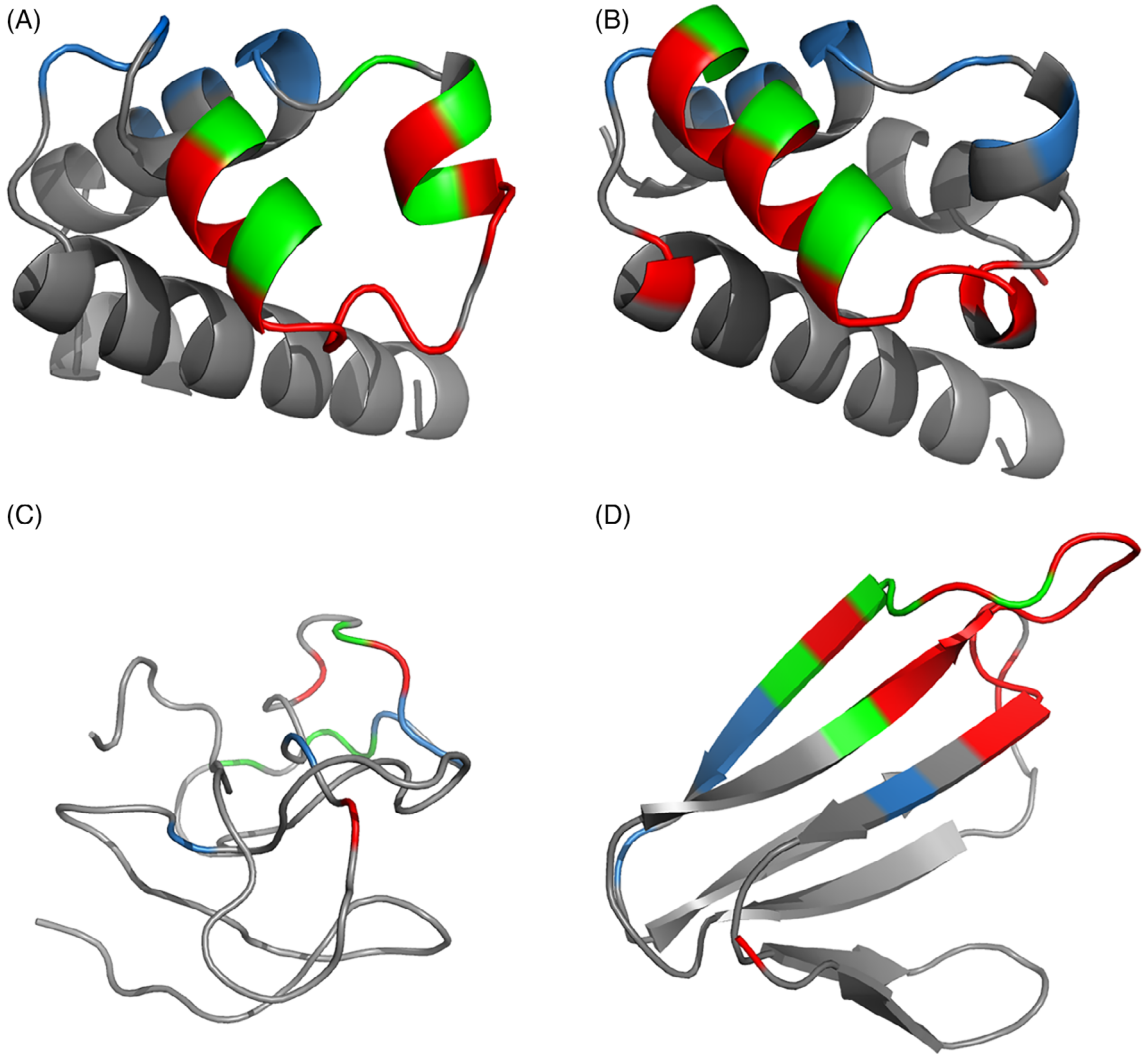


FIGURE 4 Predictions of RNA-binding residues with our method on the modeled structure and experimental one, respectively. (A) and (B) RNA binding protein SRP "the ribonucleoprotein core of the *E. coli* signal recognition particle." (C) and (D) RNA-binding protein "trp RNA-binding attenuation protein." The *TP*, *FP*, and *FN* results are shown in green, red, and marine, respectively

is 0.66. As commonly, the I-TASSER models with TM-score ≥0.5 are considered to be correct folds, we split the models into three groups with TM-score <0.5, 0.5 ≤ TM-score < 0.7, and TM-score ≥ 0.7, respectively. For binding site prediction, only the sequence information is used. From Table 1, as a whole, our method obtains almost the same results on the modeled structures as those on the experimental ones. Additionally, with the elevation of the quality of modeled structures, our method's performance has a slight improvement in terms of *SN* and *MCC* indexes.

In the following, we give our method's performances on the two not well-constructed models. One is the RNA-binding protein SRP "the ribonucleoprotein core of the *E. coli* signal recognition particle" (PDB ID: 1DUL:A).[39] The constructed structure is of TM-score = 0.50, RMSD = 1.8 Å. Compared with the prediction result on the modeled structure with *SN*, *SP*, *ACC*, and *MCC* of 0.46, 0.82, 0.73, and 0.26, respectively (Figure 4A), the prediction on the experimental one has *SN*, *SP*, *ACC*, and *MCC* of 0.38, 0.75, 0.67, and 0.12, respectively (see Figure 4B).

The second is the RNA-binding protein "trp RNA-binding attenuation protein" (PDB ID: 1C9S:S).[40] The constructed structure is of TM-score = 0.35, RMSD = 10.4 Å. Figure 4C shows the prediction result on it with *SN*, *SP*, *ACC*, and *MCC* of 0.44, 0.95, 0.89, and 0.44, respectively, and Figure 4D shows that on the experimental one with *SN*, *SP*, *ACC*, and *MCC* of 0.56, 0.74, 0.71, and 0.21, respectively.

From the above, it can be concluded that the correct folds of modeled structures are helpful to binding site prediction. In addition, our method has a good robustness against the structural variations as long as residue positions are approximately correct, which is mainly because our method is at a coarse-grained level, not very sensitive to the refined three-dimensional structures.

4 | CONCLUSIONS

We propose a new encoding scheme SNB-PSSM to incorporate evolutionary information of spatially neighbors of a target one and apply it to the prediction of RNA-binding sites on proteins. The test on RB44 dataset demonstrates SNB-PSSM method achieves an improvement compared with standard and smoothed PSSMs with *MCC* increasing by 39% and 19% and *ACC* increasing by 4% and 13%, respectively. Using a sliding window encoding scheme, the SNB-PSSM-based structure window method performs better than the standard PSSM– and smoothed PSSM– based sequence window methods, respectively. Additionally, the tests on multiple datasets indicate our method is superior to many classic methods that use PSSM profile, physical and chemical properties, or structure-based features to some extent.

In addition, our method is not sensitive to protein structural changes and has a good robustness against the structural variation, which enables our method to be applied on binding site predictions for modeled structures. This work demonstrates considering evolutionary information of spatially neighboring residues can significantly improve RNA-binding site predictions and suggests binding sites evolve spatially cooperatively to some extent. We believe the proposed method if combined with other sequence-, structure-, and dynamics-derived features can be better used for the predictions of nucleic acid–/protein-binding sites, catalytic sites, and hot spots.

The source code of SNB-PSSM can be freely downloaded at https://github.com/ChunhuaLiLab/SNB_PSSM.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (31971180, 11474013).

CONFLICT OF INTEREST

We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHORS CONTRIBUTIONS

Y.L and C.L performed the research. Y.L and C.L designed the research study. W.G and Z.Y contributed essential reagents or tools. Y.L and W.G analyzed the data. Y.L and C.L wrote the paper.

DATA AVAILABILITY STATEMENT

Data available in article supplementary material

ORCID

Yang Liu https://orcid.org/0000-0002-3332-439X
Weikang Gong https://orcid.org/0000-0001-8797-784X

REFERENCES

1. Keene JD. *RNA regulons: coordination of post-transcriptional events*. Nat Rev Genet. 2007;8(7):533-543.
2. Walia RR, El-Manzalawy Y, Honavar VG, et al. *Sequence-based prediction of RNA-binding residues in proteins*. Methods Mol Biol. 2017; 1484:205-235.
3. Walia RR, Caragea C, Lewis BA, et al. *Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art*. BMC Bioinformatics. 2012;13:89.
4. Murakami Y, Spriggs RV, Nakamura H, Jones S. *PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences*. Nucleic Acids Res. 2010;38(Web Server issue):W412-W416.
5. El-Manzalawy Y, Abbas M, Malluhi Q, et al. *FastRNABindR: fast and accurate prediction of protein-RNA Interface residues*. PLoS One. 2016; 11(7):e158445.
6. Cheng CW, Su EC, Hwang JK, et al. *Predicting RNA-binding sites of proteins using support vector machines and evolutionary information*. BMC Bioinformatics. 2008;9(Suppl 12):S6.
7. Wang B, Chen P, Huang DS, Li JJ, Lok TM, Lyu MR. *Predicting protein interaction sites from residue spatial sequence profile and evolution rate*. FEBS Lett. 2006;580(2):380-384.
8. Chen W, Zhang SW, Cheng YM, Pan Q. *Identification of protein-RNA interaction sites using the information of spatial adjacent residues*. Proteome Sci. 2011;9(Suppl 1):S16.
9. Tang Y, Liu D, Wang Z, Wen T, Deng L. *A boosting approach for prediction of protein-RNA binding residues*. BMC Bioinformatics. 2017;18 (Suppl 13):465.
10. Yang Z, Deng X, Liu Y, Gong W, Li C. *Analyses on clustering of the conserved residues at protein-RNA interfaces and its application in binding site identification*. BMC Bioinformatics. 2020;21(1):57.

11. Guharoy M, Chakrabarti P. *Conserved residue clusters at protein-protein interfaces and their use in binding site identification*. BMC Bioinformatics. 2010;11:286.
12. Ahmad S, Keskin O, Sarai A, Nussinov R. *Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins*. Nucleic Acids Res. 2008;36(18):5922-5932.
13. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. *Computational methods for prediction of protein-RNA interactions*. J Struct Biol. 2012;179(3):261-268.
14. Berman HM, Westbrook J, Feng Z, et al. *The protein data bank*. Nucleic Acids Res. 2000;28(1):235-242.
15. Walia RR, Xue LC, Wilkins K, el-Manzalawy Y, Dobbs D, Honavar V. *RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins*. PLoS One. 2014;9(5):e97725.
16. Altschul SF, Madden TL, Schaffer AA, et al. *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Res. 1997;25(17):3389-3402.
17. Cherkassky V. *The nature of statistical learning theory∼*. IEEE Trans Neural Netw. 1997;8(6):1564.
18. Dodge C, Schneider R, Sander C. *The HSSP database of protein structure-sequence alignments and family profiles*. Nucleic Acids Res. 1998;26(1):313-315.
19. Pietro Di Lena PFLM, Casadio AR. *Divide and conquer strategies for protein structure prediction*. *Mathematical Approaches to Polymer Sequence Analysis and Related Problems*. New York, NY: Springer; 2011:23-46.
20. Wang L, Huang C, Yang MQ, Yang JY. *BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features*. BMC Syst Biol. 2010;4(Suppl 1):S3.
21. Kumar M, Gromiha MM, Raghava GP. *Prediction of RNA binding sites in a protein using SVM and PSSM profile*. Proteins. 2008;71(1):189-194.
22. Kim OT, Yura K, Go N. *Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction*. Nucleic Acids Res. 2006;34(22):6450-6460.
23. Maetschke SR, Yuan Z. *Exploiting structural and topological information to improve prediction of RNA-protein binding sites*. BMC Bioinformatics. 2009;10:341.
24. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. *Assessing the accuracy of prediction algorithms for classification: an overview*. Bioinformatics. 2000;16(5):412-424.
25. Pan X, Zhu L, Fan YX, et al. *Predicting protein-RNA interaction amino acids using random forest based on submodularity subset selection*. Comput Biol Chem. 2014;53PB:324-330.
26. Wang CC, Fang Y, Xiao J, Li M. *Identification of RNA-binding sites in proteins by integrating various sequence information*. Amino Acids. 2011;40(1):239-248.
27. Miao Z, Westhof E. *Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score*. Nucleic Acids Res. 2015;43(11):5340-5351.
28. Sun M, Wang X, Zou C, He Z, Liu W, Li H. *Accurate prediction of RNA-binding protein residues with two discriminative structural descriptors*. BMC Bioinformatics. 2016;17(1):231.
29. Li S, Yamashita K, Amada KM, Standley DM. *Quantifying sequence and structural features of protein-RNA interactions*. Nucleic Acids Res. 2014;42(15):10086-10098.
30. Liu ZP, Wu LY, Wang Y, Zhang XS, Chen L. *Prediction of protein-RNA binding sites by a random forest method with combined features*. Bioinformatics. 2010;26(13):1616-1622.
31. Miao Z, Westhof E. *RBscore&NBench: a high-level web server for nucleic acid binding residues prediction with a large-scale benchmarking database*. Nucleic Acids Res. 2016;44(W1):W562-W567.
32. Miao Z, Westhof E. *A large-scale assessment of nucleic acids binding site prediction programs*. PLoS Comput Biol. 2015;11(12):e1004639.
33. Ng CL, Lang K, Meenan NA, et al. *Structural basis for 16S ribosomal RNA cleavage by the cytotoxic domain of colicin E3*. Nat Struct Mol Biol. 2010;17(10):1241-1246.
34. Stagno JR, Altieri AS, Bubunenko M, et al. *Structural basis for RNA recognition by NusB and NusE in the initiation of transcription antitermination*. Nucleic Acids Res. 2011;39(17):7803-7815.
35. Crowder SM, Kanaar R, Rio DC, Alber T. *Absence of interdomain contacts in the crystal structure of the RNA recognition motifs of sex-lethal*. Proc Natl Acad Sci U S A. 1999;96(9):4892-4897.
36. Perez-Cano L, Jimenez-Garcia B, Fernandez-Recio J. *A protein-RNA docking benchmark (II): extended set from experimental and homology modeling data*. Proteins. 2012;80(7):1872-1882.
37. 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(1):7-8.
38. Xu J, Zhang Y. *How significant is a protein structure similarity with TM-score = 0.5?* Bioinformatics. 2010;26(7):889-895.
39. Batey RT, Rambo RP, Lucast L, Rha B, Doudna JA. *Crystal structure of the ribonucleoprotein core of the signal recognition particle*. Science. 2000;287(5456):1232-1239.
40. Antson AA, Dodson EJ, Dodson G, Greaves RB, Chen XP, Gollnick P. *Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA*. Nature. 1999;401(6750):235-242.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu Y, Gong W, Yang Z, Li C. SNB-PSSM: A spatial neighbor-based PSSM used for protein–RNA binding site prediction. *J Mol Recognit*. 2021;e2887. https://doi.org/10.1002/jmr.2887