IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 25, 2018, accepted January 6, 2019, date of publication January 17, 2019, date of current version March 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2893620

Analyzing the Surface Structure of the Binding Domain on DNA and RNA Binding Proteins

WEI WANG[1,2], KELIANG LI[1], HEHE LV[1], HONGJUN ZHANG[3], SHIGUANG ZHANG[1], YUN ZHOU[1], AND JUNWEI HUANG[1]

[1]Department of Computer Science and Technology, College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China
[2]Laboratory of Computation Intelligence and Information Processing, Engineering Technology Research Center for Computing Intelligence and Data Mining, Xinxiang 453007, China
[3]School of Aviation Engineering, Anyang University, Anyang 455000, China

Corresponding author: Wei Wang (weiwang@htu.edu.cn)

This work was supported in part by the Natural Science Foundation of Henan Province under Grant 182300410368, Grant 182300410130, and Grant 182300410306, in part by the Production and Learning Cooperation and Cooperative Education Project of Ministry of Education of China under Grant 201702115008, in part by the Science and Technology Research Key Project of Educational Department of Henan Province under Grant 16A520016, Grant 17B520002, Grant 17B520036, and Grant 17A520039, in part by the Key Project of Science and Technology Department of Henan Province under Grant 182102210208, Grant 142102210056, and Grant 17B520002, in part by the Ph.D. Research Startup Foundation of Henan Normal University under Grant qd15130, Grant qd15132, and Grant qd15129.

ABSTRACT The study of nucleic acid-binding protein (NBP) has important significance for us to understand critical intracellular activities, such as the transmission of cellular genetic information, cell metabolism, substance transport, and signal transduction. DNA-binding proteins (DBPs) and RNA-binding proteins (RBPs) interact through their diverse binding domains and different types of nucleic acid molecules. In this paper, we used a novel method that combines the CX algorithm and the fractal surfaces algorithm. This method gets the molecular volume and solvent surface area in the local area of the NBPs binding domain residues. Then, based on the algorithm results, the requisite domain residues are divided into three types: peak, flat, and valley. At the same time, we analyzed the solvent accessibility and secondary structural characteristics of the DBPs and RBPs binding domains. Finally, we found that there was an important difference in the distribution of peak residues and valley residues in the two types of NBPs binding domains. Similarly, there were significant differences in the solvent accessibility and secondary structural distribution of the two types of NBPs binding domains. To verify the existence of differences, we constructed SVM classifier to make a distinction between DBPs and RBPs using a 10-fold cross-validation method. Lastly, the SVM classification model achieves AUC of 78%. In summary, we have proposed a new perspective for the study of NBPs binding domains. This method not only calculates the geometric characteristics of the molecule, but also analyzes the protein properties associated with the structure, which will assist in the study of NBPs binding domains.

INDEX TERMS DNA-binding proteins, RNA-binding proteins, solvent accessibility, secondary structure.

I. INTRODUCTION

These NBPs play an important part in the physiological activities of cells, including gene transcription, DNA molecule repair and replication, DNA virus-infected cells, DNA molecule stacking and modification, post-transcriptional regulation of genes, selective cleavage of mRNA, and infection process of RNA viruses [1]–[8]. With the continuous development of structural measurement techniques and high-throughput sequencing technologies, a large amount of NBPs sequence data and structural data have been collected. This provides a solid data foundation for studying NBPs. In recent years, more ideas and methods have been put forward for NBPs research, such as, the identification and prediction of binding sites for binding proteins, building features based on sequence information, establishing classification models and training predicted protein binding sites [9], [10]. Increased protein structure data provides the basis for analyzing the structural features of binding proteins. Extracting effective feature information from structural data and then

The associate editor coordinating the review of this manuscript and approving it for publication was Navanietha Krishnaraj Rathinam.

constructing a classification model for predicting binding sites has yielded significant results [11]–[14]. Existing studies have also proposed new methods for forecasting NBPs, and explored the differences between different NBPs. For example, Janin and Bahadur [15] studied the binding domains of DBPs and RBPs, then analyzed the total area, composition, and some physicochemical properties on the binding domains. Finally, the mechanism of the interaction between nucleic acids and proteins was analyzed [15]. Shao *et al.* [16] processed the sequence data of the NBPs and used a new descriptor called conjoint triad to extract the sequence information of the protein. The descriptor takes into account the nature of amino acids and their neighboring amino acids, and any three consecutive the amino acids are treated as one unit. Finally, two SVM classifiers were constructed to classify DBPs/RBPs and non-nucleic-acid-binding proteins [16]. Yu [17] processed the sequence data of NBPs, extracted the pseudo-amino acid composition of the sequence, and finally used the SVM classifier to perform two-class processing on the three proteins rRNA-, RNA-, and DNA-binding proteins. Our paper presents a new set of methods for describing the structural characteristics of residues in the binding domain. The binding domains of DBPs mainly include: Zinc Finger, Leucine Zipper, Helix-Turn-Helix (HTH), Helix-Loop-Helix (HLH) [18]–[20]. The binding domains of RBPs can be subdivided into RNA recognition motifs, double-stranded RNA-binding domains, K homology domains, arginine-rich motifs [2], [21], [22]. Based on the existing studies on NBPs binding domains, we calculated the characteristics of the local morphological features, secondary structure distribution, and solvent accessibility of residues extracted from the surface of the binding domain based on nucleic acid-protein structure data, and then analyzed the differences between the two proteins. First, we obtained protein structure data from the PDB database, screened out the qualified proteins, and then calculated the local morphological characteristics, solvent accessibility, and secondary structure characteristics of the residues bound to the protein binding domain. The extracted features were analyzed to investigate the differences in structural characteristics between the DBPs and RBPs. SVM classifier was used to demonstrate the differential presence of structural features of NBPs binding domains. It is desired that our research methods and experimental results will contribute to the research and development of binding domains of NBPs.

II. MATERIALS AND METHODS

A. MATERIALS

In our work, structural data of NBPs are downloaded from the PDB database. Until 2018, the 8021 DBPs and the 5660 RBPs have been collected in the PDB database. We removed the low-resolution structure, leaving only X-ray structures with a resolution higher than 3Å and NMR analytical structures. In addition, we used the PISCES program (http://dunbrack.fccc.edu/Guoli/PISCES.php) to remove homology redundancy for the collection of protein data [23]. The structural data with a sequence homology of no more than 30% and a minimum chain length of the 40 residues were chosen. At the same time the nucleotide-free protein data was removed. Finally, we obtained the non-redundant protein nuclear acids complexes (369 DBPs and the 174 RBPs) as training datasets and the remaining complexes (200 DBPs and 103 RBPs) were used as independent test datasets. It is generally considered that the distance between the residues Ca atoms of the binding domain and the nucleic acid molecule should be less than 6 Å, and thus we calculated the binding domain of the NBPs.

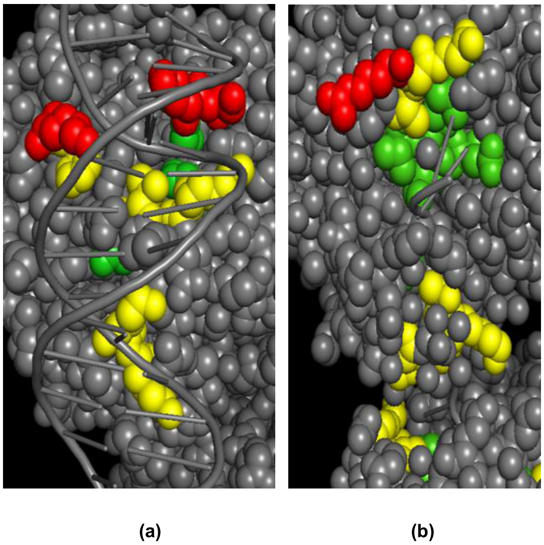


FIGURE 1. Examples of the geometric structure of the binding domain residues of NBPs. Peak residues are shown red. Flat residues are shown yellow, and the valley residues are shown green. (a) The DBP's ID is 3sqi; (b) The RBP's ID is 3b0u.

B. DETERMINATION OF BINDING DOMAIN SHAPE

The surface shape of protein is irregular and contains a variety of forms. There are various sizes of cracks and grooves. There is a difference in the geometry of the binding domain of DBPs and RBPs. Therefore, an algorithm for measuring the local geometry of residues in the binding domain was proposed. We performed statistical analysis on the distribution of geometric features and tried to find structural differences between the binding domains of DBPs and RBPs.

As showed in Fig. 1, we divided the local geometry of the residues in the binding domain into three types: peak, flat, and valley. Pintar *et al.* [24] proposed the CX algorithm, which determined the shape of protrusions and depressions on the protein surface by calculating the ratio of the occupied volume and the free volume of the protein in the sphere. Fractal surfaces can be used to characterize protein surface roughness or irregularity by calculating the relationship between surface area and volume [25]. Based on CX algorithm and fractal surfaces we have designed a new algorithm to measure the local shape of residues. Take the three-dimensional coordinates of the Ca atom of the residues in the binding domain at the sphere center and set a fixed sphere radius *R* (*R* defaults

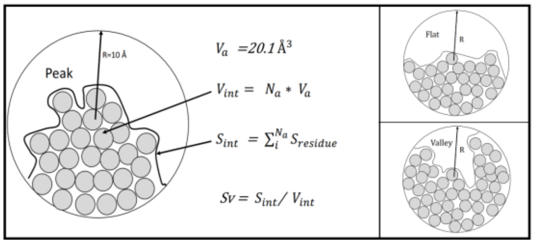


FIGURE 2. Schematic representation of each protein surface identified by improved CX algorithm.

to 10Ař). As shown in Fig. 2, R is the total surface area of the binding domain residues contained within the sphere.S_{int} is calculated by the solvent accessibility of residues. $V_{int}$is the volume of the residue inside the sphere:

$$S_{\text{int}} = \sum_i^N S_{residue} \tag{1}$$

$$V_{\text{int}} = N_a * V_a \tag{2}$$

where N_a is the number of nonhydrogen atoms in the sphere, and V_a is the average volume of a nonhydrogen atom (the value is set to 20.1 Å^3). $S_{residue}$is the solvent accessibility area of the ith residue of the binding domain. The solvent accessibility area can be calculated using DSSP software [26]. The formula for calculating the local geometry of residues in the binding domain is as follows:

$$Sv = S_{\text{int}}/V_{\text{int}} \tag{3}$$

Sv is the ratio of the solvent accessibility area within the sphere to the total volume of all atoms inside the sphere. If the solvent accessibility area of the NBPs binding domain in the sphere is larger and the total volume of the atoms is smaller, the peripheral surface geometry of the residues at the center of the sphere is more inclined to peak shape. If the solvent accessibility area of the NBPs binding domain in the sphere is smaller and the total volume of atoms is larger, the peripheral surface geometry of the residue in the center of the sphere is more inclined to the valley shape. Through experimental analysis, we divide the surface residues into three types: valley ($Sv < 0.3$), plane ($0.3 < Sv < 0.5$), and peak ($Sv > 0.5$).

C. SOLVENT AVAILABILITY ANALYSIS

Based on the relative solvent accessibility values of the surface residues, the degree of exposure to the residues in the solution can be judged, indicating that the solvent accessibility is closely related to the geometry of the protein binding domain [27]. By counting the average relative solvent accessibility values (ASA) and the average solvent accessibility values (RSA) of NBPs surface binding domain surface residues, the degree of exposure of DBP and RBP binding domain residues can be measured, these features can also illustrate the morphological characteristics of the surface of the binding domain from another direction. $S_{binding}$ is the surface area of the binding domain of proteins and nucleotides, R_{accint} is the RSA of the binding region residues, R_{accper} is the average RSA of the binding region residues,S_{per} is the average ASA of the binding region residues, the formula is as follows:

TABLE 1. Abbreviations and Notes For Secondary Structure Categories.

Category	Description of secondary structure types
G	3-turn helix (310helix). Min length 3 residues
H	4-turn helix (α helix). Min length 4 residues
I	5-turn helix (π helix). Min length 5 residues
T	Hydrogen bonded turn (3, 4 or 5 turn)
B	Residue in isolated β-bridge (single pair β-sheet hydrogen bond formation)
S	Bend (the only non-hydrogen-bond based assignment)
C	Coil (residues which are not in any of the above conformations)
E	Extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues

$$S_{binding} = \sum_i^n S_{int} \tag{4}$$

$$S_{per} = S_{binding}/n \tag{5}$$

$$R_{acc_{per}} = \sum_i^n Racc_{\text{int}}/n \tag{6}$$

where n is the number of all interface residues, S_{int} is the ASA of ith residue in the whole interface.

D. SECONDARY STRUCTURE ANALYSIS

DSSP is a standard in the field of secondary structure determination and prediction. According to Pauling's proposed hydrogen bonding pattern to determine which secondary structure belongs, eight kinds of secondary structures are defined as showed in Table 1.

For 8-state prediction, the α-helix is further subdivided into three states: α-helix (H), 310 helix (G), and π-helix (I). The beta chain is subdivided into: beta chain (E) and beta bride (B), and the coil region is subdivided into: high curvature ring (S), beta turn (T') and irregular (L). We constructed an eight-dimensional vector (S_g, S_h, S_i, S_t, S_e, S_b, S_s, S_c) to characterize the secondary structure of residues in the NBPs binding domain, where S_i indicates the frequency of each secondary structure in the binding domain surface.

$$S_i = \frac{n_i}{\sum_{i=1}^{8} n_i} \tag{7}$$

where $n_i$is the number of occurrences of class $i$secondary structure.

E. CLASSIFICATION MODEL AND EVALUATION METHOD

We constructed the SVM classification model based on the previously mentioned features. SVM is a machine learning method based on statistical learning theory [28]. The SVM is performed by the Support Vector Machine scikit-learn v0.19.1 package for python to evaluate the performance of the model [29]. The model was evaluated by 10-fold cross-validation experiments. The overall performance was calculated by averaging the performance of the 10 subsets (at the fold level). Because of the imbalance of datasets, the dataset is selected by down sampling in the training phase. Test indicators used in the classification model are the overall prediction accuracy (ACC), F-measure (F1), the area under the ROC (Receiver Operating Characteristic) curve (AUC), Sensitivity and Specificity. The F-measure can be interpreted as a weighted harmonic mean of the precision and recall. The ROC curve is probably the most robust technique for evaluating classifiers and visualizing their performance. The area under the curve (AUC) is used to measure of the quality of the separation between the examined protein classes. The AUC of 0.5 represents a classification that corresponds to a randomly generated prediction, while the area of 1 corresponds to a perfect classifier.

III. RESULTS AND DISCUSSION

A. DISTRIBUTION OF THREE BINDING DOMAIN RESIDUES

We analyzed the DBPs and RBPs data, based on the "lock and key" paradigm and the morphology of nucleic acid molecule [30]. We found that the interface surface of DBPs and RBPs showed different shape distribution.

As showed in Fig. 3, the peak residues in the DBPs binding domain are distributed more than the peak shape residues of the RBPs binding domain. There is no significant difference in the distribution of the flat shape residues in DBPs binding domain and RBPs binding domain. The valley residues of DBPs binding domain are less distributed than the valley residues of RBPs binding domain. Previous studies have found that zinc finger domains are generally found in the binding domains of DBPs, particularly transcription factors. However, as the research progresses, more and more evidence shows that the zinc finger domain not only exists in the DNA binding domain, but also can be found in the binding domain of RBPs. The zinc finger structure present in the bound protein binding domain highlights the protein surface in the form of a finger. The physical interaction between protein domains is the basis for interactions between proteins. The lock-key structure defines the interaction between proteins that contains complementary domains (locks and keys). Under the "lock and key" paradigm, the interface of DBPs should be significantly different from RBPs. As can be seen from the analysis in Figure 3, there is a significant difference in the distribution of the interface between the DBPs and the RBPs in the peak and valley morphology.

This is in agreement with the conclusion that the distribution of peak and the valley shape residues is more than that of the flat shape residue at the two types of protein binding domains. RBPs usually function during the transcriptional or translational stages, and RNA molecules are more flexible than DNA. Nucleotides in RBPs are often associated with groove regions in the protein, so there are more concave surfaces in the binding domain of RBPs. However, DNA usually infiltrates into the major or minor groove of the binding protein, so that the interfaces of DBPs exhibit more protrusions and the intermediate surface to bind the DNA stably. Fig. 4 shows the state of distribution ratios for residues of three shapes in the binding domain. It can be seen from Fig. 4a that the frequencies of the peak shape residues ARG, GLY, LYS and SER are higher than the others, and the distribution of ARG and LYS shows significant difference. From Fig. 4b and Fig. 4c, it can be observed that the distribution differences of two kinds of protein residues are getting smaller and smaller. For example, the distribution differences of ARG and LYS compared with other residues are the largest at the peak surface, followed by the flat surface, and the valley surface is the smallest. This phenomenon also shows that peak residues are the major part of the interaction between proteins and nucleic acid molecules. As showed in additional S_Fig. 1, there is statistically significant difference in the distribution of the two shapes of the CYS, GLN and THR residues. There is statistically significant difference in the distribution of one shape of the VAL, TRP, MET, LEU, GLU and ASP residues.

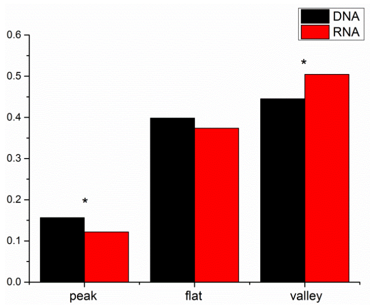
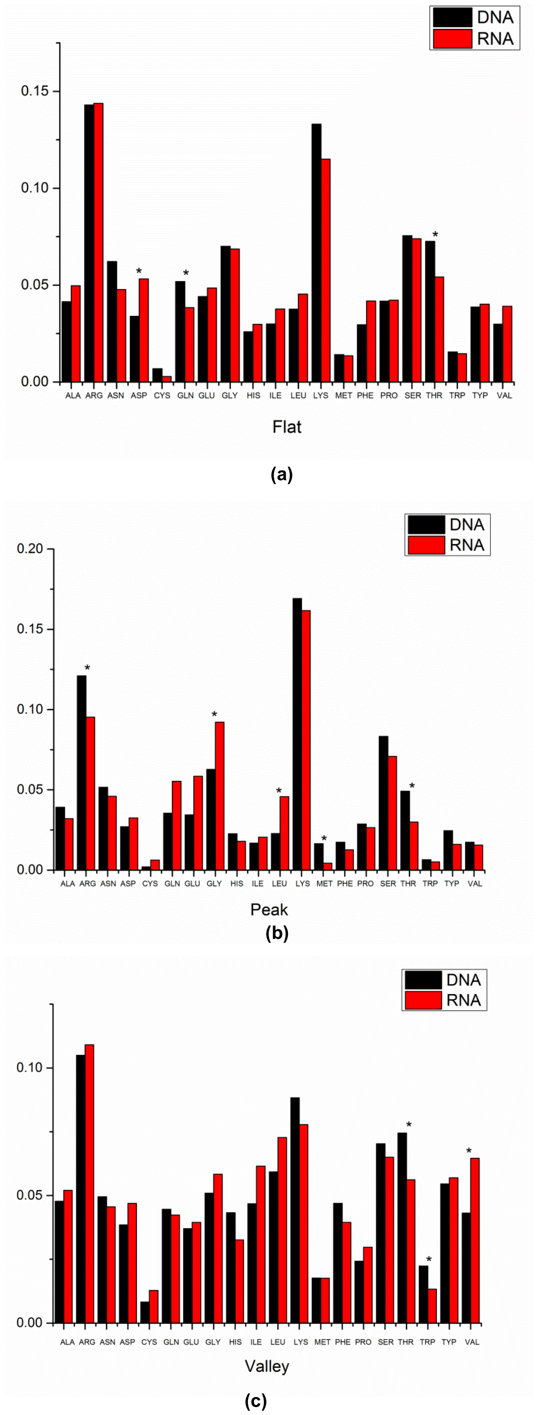


FIGURE 3. Analysis of the residues shapes in DNA and RNA binding domain. The statistical analysis of the distribution was performed using the T-test to measure the significance of the difference. The asterisks marked in the figure indicate significant differences in P-values < 0.05.

B. SOLVENT ACCESSIBILITY

Solvent accessibility is a feature closely related to the surface geometry of a protein. We calculated the average ASA and the average RAS of the residues on the protein binding domain, so as to analyze the solvent accessibility characteristics of DBPs and RBPs surface domain residues. We also analyzed the average ASA and the average RSA of residues in the binding domain.

The distribution status of the solvent accessibility is illustrated in Fig. 5. We found that the ARG and LYS have the higher values than ASA and RAS in the binding domain. In addition, we observed in the previous section that the ARG and LYS tend to show peaks. The binding domain structure can be considered as an external manifestation, and the intrinsic property of the binding protein is an important factor determining the binding process. Among 20 kinds of amino acids, Arg and Lys present positively charged, Asp and Glu present negatively charged, and His has weak positive charge depending on the local environment. Since the acidic quality of the backbone phosphates, the surface of a DNA has

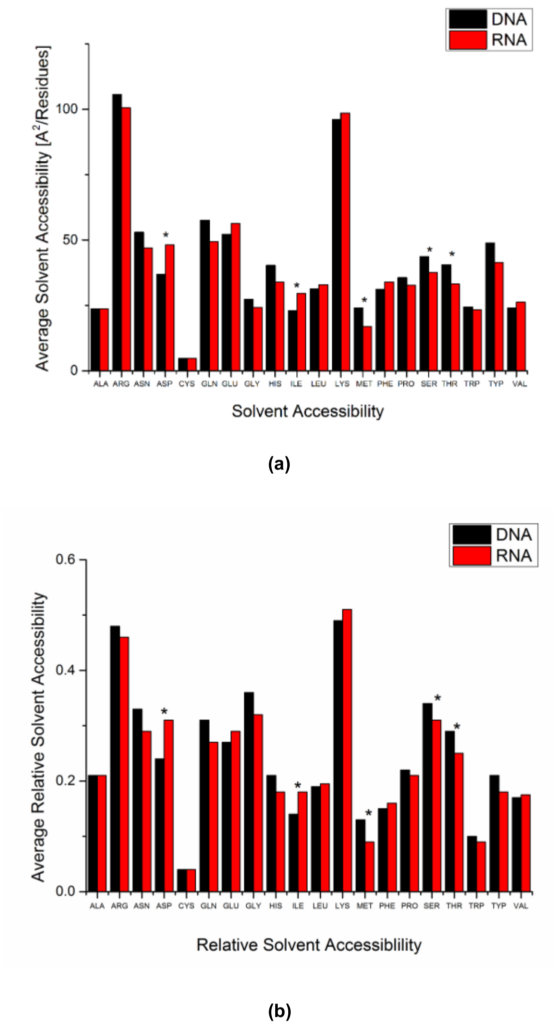


(a)

(b)

(c)

FIGURE 4. **The distribution of peak, flat and valley shape residues. The asterisks marked in the figure indicate significant differences in P-values < 0.05. (a) Distribution of residues in the peak region; (b) Distribution of residues in the flat area; (c) Distribution of residues in the valley area.**



(a)

(b)

FIGURE 5. **The solvent accessibility value and relative solvent accessibility value distribution in the DNA and RNA binding domains of 20 kinds of amino acids. The statistical analysis of the distribution was performed using the T-test to measure the significance of the difference. The asterisks marked in the figure indicate significant differences in P-values < 0.05. (a) Distribution of solvent accessibility values of 20 amino acids. (b) Distribution of relative solvent accessibility values of 20 amino acids.**

a higher negative charge. The positively charged amino acids (ARG and LYS) are more distributed on the binding surface of proteins, which will contribute to bind the protein to the nucleic acid. The SER and THR values shown are higher than others in Fig. 5a, and the residues show statistically significant differences. From the relationship between ASA and RAS, we can infer that the two types of amino acids have a higher exposure ratio on the surface of the NBPs binding domain, and they tend to bind to nucleic acid molecules. This demonstrates that the study of the solvent accessibility of the NBPs binding domain is helpful in constructing a classification model. As showed in additional S_Fig. 2, from the eigenvalues of the solvent accessibility of the 20 kinds of residues and the distribution tendency of the three shapes, it can be observed that there is correlation between the geometry of the residue and the solvent accessibility of the residue. Because the linear trends of the two sets of variables in each subgraph are very similar.

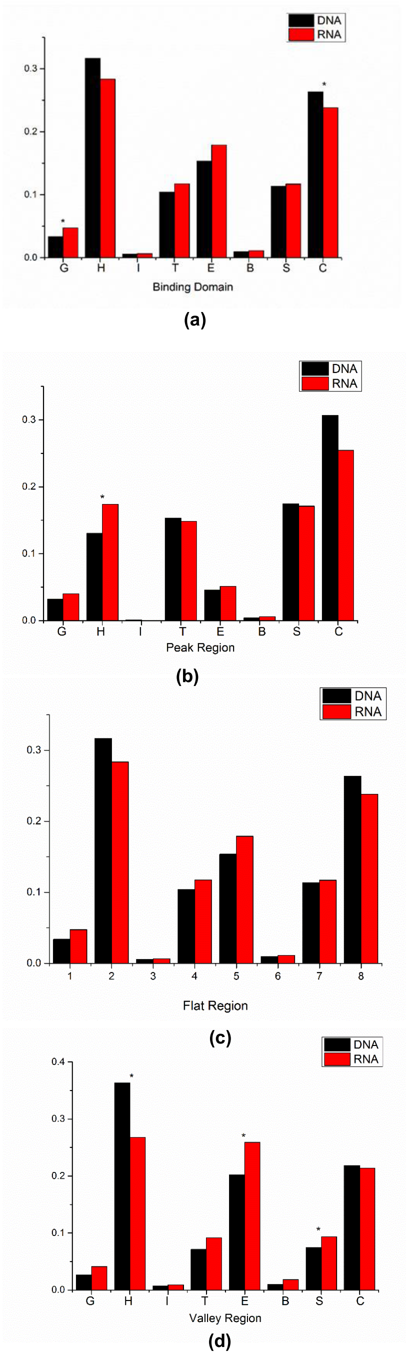
C. SECONDARY STRUCTURE DISTRIBUTION

The distribution of secondary structure in the protein binding domain is illustrated in Fig. 6a. There are no significant differences between DNA and RNA binding proteins for the H-type, T-type, and E-type secondary structures, and the G- and C-type secondary structures present the significant differences. The distribution of the H-type secondary structure in the DBPs and RBPs is the most common, and there are more H-type secondary structures in DBPs than in RBPs. The reason behind this could be that the DBPs binding domains typically consists of zinc finger domains and helix-turn-helices. These domains contain many H-type secondary structures. Similarly, the RNA binding domain has also the zinc finger structure. Fig. 6b shows the secondary structure distribution of peak residues in the NBPs binding domain. It can be seen that the distribution of the H-class secondary structure is significantly different. Fig. 6c shows the secondary structure distribution of flat residues in the NBPs-binding domain. It can be found that there is no significant difference in the distribution of all classes of secondary structure. Fig. 6d shows the secondary structure distribution of flat residues in the NBPs-binding domain. It can be found that there are significant differences in the distribution of H, E, and S secondary structures. These results show that the distribution characteristics of the secondary structure in the binding domain which will conducive to the establishment of a classification model.

D. PREDICTION PERFORMANCE OF CLASSIFICATION MODEL

In the present study, Inbal Paz built a BindUP model based on the electrostatic properties of protein surfaces and other general properties of proteins, classifying and predicting DBPs and RBPs [31]. Jing Yan had constructed a sequence-based DRNApred model, the model could accurately and high-throughputs the prediction and differentiation of DBPs and RBPs binding residues [32]. These studies have some significant results, but all of these are based on sequences or the physicochemical properties of proteins. Shula Shazman proposed an algorithm based on differential geometry. The algorithm can extract the geometric characteristics of the binding domain, and then predict the double-stranded DBPs and the single-stranded RBPs [33]. Different from the previous work, we calculated the geometric characteristics of the NBPs binding domain from the perspective of the binding domain structure.

In the experiment we used a total of three characteristics: First, the distribution of three morphological residues in protein binding domain; second, the volume accessibility characteristics of protein binding domain; third, the secondary structure distribution of residues in protein binding domain. We used the above individual features to build the SVM classifier, and then we gathered all the features together to train the dataset, and finally used the SVM model to perform 10-fold cross validation tests. As shown in Table 2, it can be found that the ACC and AUC of the SVM model only using solvent accessibility are the highest, 0.6381 and 0.7076, respectively. Followed by the shape of the residues, the ACC and AUC results were 0.6095 and 0.6909, respectively. The effect of using only the secondary structure for classification prediction was the worst, with ACC and AUC being 0.5905 and 0.6280, respectively. This is consistent with the previous analysis. For the binding domains of the



(a)

(b)

(c)

(d)

FIGURE 6. The distribution of secondary structure on NBPs binding domains. The statistical analysis of the distribution was performed using the T-test to measure the significance difference. The asterisks marked in the figure indicate significant differences in P-values < 0.05. (a) The distribution of secondary structures in the binding domain. (b) The distribution of secondary structures of the peak shapes in the binding domain. (c) The distribution of secondary structures of the flat shapes in the binding domain. (d) The distribution of secondary structures of the valley shape in the binding domain.

TABLE 2. Evaluation Results of Different Characteristics of SVM Model.

Test Methods	Feature	Acc	Sen	Spe	AUC
10-fold Cross-Validation	Residue Shapes（P<0.05）	0.6095	0.6403	0.5852	0.6909
	Solvent Accessibility（P<0.05）	0.6381	0.5800	0.6304	0.7076
	Secondary Structure（P<0.05）	0.5905	0.5454	0.6250	0.6280
	All Features（P<0.05）	0.7347	0.6000	0.7857	0.7847
Independent Test	All Features（P<0.05）	0.6907	0.5905	0.6425	0.6943

two types of proteins, the attribute values of the secondary structure are the least distinguishable, and the three types of peak, valley and flat are the most different.

This result shows that secondary structure contributes the least to the classification model, because there are many grooves and convex in the binding domain of NBPs. When all features were used to predict, the results were the best. These values of ACC and AUC reach 0.7347 and 0.7847 respectively. These results show that using the three kinds of features are complementary to each other, and the independent dataset has achieved good results. So this results show that these selected features are considered as the optimal feature sets used in our final DBPs and RBPs classification model.

IV. CONCLUSIONS

In the work, we designed a classification model to divide between DBPs and RBPs. The model was built on the structural characteristics of the NBPs binding domain, and finally achieved significant results. The prediction model achieves better prediction rate, and we have further proposed a new method to measure the shape of residues in the binding region. The algorithm uses the binding domain residue Ca atom as the center of the sphere. First, the total volume and surface area of the protein molecules inside the sphere are calculated. Then, the ratio of surface area to volume can represent the local depression and convex shape. In our work, we not only extracted the geometric characteristics of protein, but also calculated secondary structure and solvent accessibility. Finally, the results show that there are significant differences in the residue morphology, secondary structure, and availability of solvent accessibility in the binding domains of DBPs and RBPs. A series of feature extraction methods can also be applied to the prediction of binding protein binding sites, and the prediction of the binding relationship between drug molecules and proteins.

REFERENCES

[1] Y. Yan, D. Zhang, P. Zhou, B. Li, and S.-Y. Huang, "HDOCK: A Web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy," *Nucleic Acids Res*, vol. 45, no. W1, pp. W365–W373, Jul. 2017.

[2] S. Helder, A. J. Blythe, C. S. Bond, and J. P. Mackay, "Determinants of affinity and specificity in RNA-binding proteins," *Current Opinion Structural Biol.*, vol. 38, pp. 83–91, Jun. 2016.

[3] W. Zhang, X. Yue, G. Tang, W. Wu, F. Huang, and X. Zhang, "SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions," *PLoS Comput. Biol.*, vol. 14, no. 12, p. e1006616, 2018.

[4] W. Zhang, Q. Qu, Y. Zhang, and W. Wang, "The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions," *Neurocomputing*, vol. 273, pp. 526–534, Jan. 2018.

[5] W. Zhang, J. Liu, and Y. Niu, "Quantitative prediction of MHC-II peptide binding affinity using relevance vector machine," *Appl. Intell.*, vol. 31, no. 2, pp. 180–187, 2009.

[6] Z. Peng and L. Kurgan, "High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder," *Nucleic Acids Res.*, vol. 43, no. 18, p. e121, Oct. 2015.

[7] W. Zhang, J. Liu, and Y. Niu, "Quantitative prediction of MHC-II binding affinity using particle swarm optimization," *Artif. Intell. Med.*, vol. 50, no. 2, pp. 127–132, 2010.

[8] R. R. Walia *et al.*, "Protein-RNA interface residue prediction using machine learning: An assessment of the state of the art," *BMC Bioinf.*, vol. 13, no. 1, p. 89, 2012.

[9] W. Zhang, J. Liu, Y. Q. Niu, L. Wang, and X. Hu, "A Bayesian regression approach to the prediction of MHC-II binding affinity," *Comput. Methods Programs Biomed.*, vol. 92, no. 1, pp. 1–7, 2008.

[10] W. Wang *et al.*, "Analysis and prediction of single-stranded and double-stranded DNA binding proteins based on protein sequences," *BMC Bioinf.*, vol. 18, no. 1, p. 300, 2017.

[11] Z. P. Liu, L. Y. Wu, Y. Wang, X. S. Zhang, and L. Chen, "Prediction of protein-RNA binding sites by a random forest method with combined features," *Bioinformatics*, vol. 26, no. 13, pp. 1616–1622, 2010.

[12] F. Towfic, C. Caragea, D. C. Gemperline, D. Dobbs, and V. Honavar, "Struct-NB: Predicting protein-RNA binding sites using structural features," *Int. J. Data Mining Bioinf.*, vol. 4, no. 1, pp. 21–43, 2010.

[13] H. Zhao, Y. Yang, and Y. Zhou, "Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets," *Nucleic Acids Res.*, vol. 39, no. 8, pp. 3017–3025, 2011.

[14] W. Wang, J. Liu, and L. Sun, "Cover image, volume 84, issue 7: Analysis of SSBs and DSBs interface," *Proteins Struct. Function Bioinf.*, vol. 84, no. 7, p. C4, 2016.

[15] J. Janin and R. P. Bahadur, "Relating macromolecular function and association: The structural basis of protein–DNA and RNA recognition," *Cellular Mol. Bioeng.*, vol. 1, no. 4, pp. 327–338, 2008.

[16] X. Shao, Y. Tian, L. Wu, Y. Wang, L. Jing, and N. Deng, "Predicting DNA- and RNA-binding proteins from sequences with kernel methods," *J. Theor. Biol.*, vol. 258, no. 2, pp. 289–293, 2009.

[17] X. Yu, J. Cao, Y. Cai, T. Shi, and Y. Li, "Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines," *J. Theor. Biol.*, vol. 240, no. 2, pp. 175–184, 2006.

[18] W. H. Landschulz, P. F. Johnson, and S. L. McKnight, "The leucine zipper: A hypothetical structure common to a new class of DNA binding proteins," *Science*, vol. 240, no. 4860, pp. 1759–1764, 1988.

[19] A. Tripathi and V. A. Bankaitis, "Molecular docking: From lock and key to combination lock," *J. Mol. Med. Clin. Appl.*, vol. 2, no. 1, pp. 1–19, 2017.

[20] S. J. Greive, "DNA recognition for virus assembly through multiple sequence-independent interactions with a helix-turn-helix motif," *Nucleic Acids Res.*, vol. 44, no. 2, pp. 776–789, 2016.

[21] K. Musunuru and R. B. Darnell, "Determination and augmentation of RNA sequence specificity of the Nova K-homology domains," *Nucleic Acids Res.*, vol. 32, no. 16, pp. 4852–4861, 2004.

[22] T. S. Bayer, L. N. Booth, S. M. Knudsen, and A. D. Ellington, "Arginine-rich motifs present multiple interfaces for specific binding by RNA," *RNA*, vol. 11, no. 12, pp. 1848–1857, 2005.

[23] G. Wang and R. L. Dunbrack, Jr., "PISCES: A protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[24] A. Pintar, O. Carugo, and S. Pongor, "CX, an algorithm that identifies protruding atoms in proteins," *Bioinformatics*, vol. 18, no. 7, pp. 980–984, 2002.

[25] M. Lewis and D. C. Rees, "Fractal surfaces of proteins," *Science*, vol. 230, no. 4730, pp. 1163–1165, 1985.

[26] R. P. Joosten *et al.*, "A series of PDB related databases for everyday needs," *Nucleic Acids Res.*, vol. 39, pp. D411–D419, Nov. 2011.

[27] L. Deng, C. Fan, and Z. Zeng, "A sparse autoencoder-based deep neural network for protein solvent accessibility and contact number prediction," *BMC Bioinf.*, vol. 18, no. 16, p. 569, 2017.

[28] M. Kumar, M. M. Gromiha, and G. P. S. Raghava, ‘‘Prediction of RNA binding sites in a protein using SVM and PSSM profile,’’ *Proteins Struct. Function Bioinf.*, vol. 71, no. 1, pp. 189–194, 2008.

[29] F. Pedregosa *et al.*, ‘‘Scikit-learn: Machine learning in Python,’’ *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[30] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert, ‘‘A lock-and-key model for protein-protein interactions,’’ *Bioinformatics*, vol. 22, no. 16, pp. 2012–2019, 2006.

[31] I. Paz, E. Kligun, B. Bengad, and Y. Mandel-Gutfreund, ‘‘BindUP: A Web server for non-homology-based prediction of DNA and RNA binding proteins,’’ *Nucleic Acids Res.*, vol. 44, no. W1, pp. W568–W574, 2016.

[32] J. Yan and L. Kurgan, ‘‘DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues,’’ *Nucleic Acids Res.*, vol. 45, no. 10, p. e84, 2017.

[33] S. Shazman, G. Elber, and Y. Mandel-Gutfreund, ‘‘From face to interface recognition: A differential geometric approach to distinguish DNA from RNA binding surfaces,’’ *Nucleic Acids Res.*, vol. 39, no. 17, pp. 7390–7399, 2011.

Authors’ photographs and biographies not available at the time of publication.