OXFORD

Structural bioinformatics

Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods

Hong Su[1], Mengchen Liu[1], Saisai Sun[1], Zhenling Peng[2],* and Jianyi Yang [iD] [1],*

[1]School of Mathematical Sciences, Nankai University, Tianjin 300071, China and [2]Center for Applied Mathematics, Tianjin University, Tianjin 300072, China

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Abstract

Motivation: The interactions between protein and nucleic acids play a key role in various biological processes. Accurate recognition of the residues that bind nucleic acids can facilitate the study of uncharacterized protein–nucleic acids interactions. The accuracy of existing nucleic acids-binding residues prediction methods is relatively low.

Results: In this work, we introduce NucBind, a novel method for the prediction of nucleic acids-binding residues. NucBind combines the predictions from a support vector machine-based *ab-initio* method SVMnuc and a template-based method COACH-D. SVMnuc was trained with features from three complementary sequence profiles. COACH-D predicts the binding residues based on homologous templates identified from a nucleic acids-binding library. The proposed methods were assessed and compared with other peering methods on three benchmark datasets. Experimental results show that NucBind consistently outperforms other state-of-the-art methods. Though with higher accuracy, similar to many other *ab-initio* methods, cross prediction between DNA and RNA-binding residues was also observed in SVMnuc and NucBind. We attribute the success of NucBind to two folds. The first is the utilization of improved features extracted from three complementary sequence profiles in SVMnuc. The second is the combination of two complementary methods: the *ab-initio* method SVMnuc and the template-based method COACH-D.

Availability and implementation: http://yanglab.nankai.edu.cn/NucBind

Contact: zhenling@tju.edu.cn or yangjy@nankai.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Protein–nucleic acids interactions are involved in many biological processes, such as DNA replication, transcription and translation. For example, in the process of gene transcription, the transcription factor (a special kind of protein) binds specific DNA molecules to control the transcription rate of genetic information from DNA to messenger RNA (von Hippel *et al.*, 1984). Many efforts have been done to decipher the interactions between protein and nucleic acids. One of the most direct ways is determining the protein–nucleic acids complex structure by experiments, such as X-ray and/or NMR. For instance, three scientists, Ramakrishnan, Steitz and Yonath, were awarded the Nobel prize in Chemistry in 2009, to recognize their

significant contributions to the determination of the structure and mechanism of ribosome, a complex molecular machine consisting of rich protein–nucleic acids interactions.

It is notoriously difficult and costly to solve the protein–nucleic acids complex structure by experiment. Thus, there is a growing demand for the development of computational algorithms to predict protein–nucleic acids interactions. In fact, a lot of computational studies have been performed, including the recognition of DNA-binding domain/protein (Zhang and Liu, 2017), protein–DNA/RNA docking (Yan *et al.*, 2017), DNA motif pair discovery (Wong, 2017), and so on. Here, we are interested in the prediction of nucleic acids-binding residues in proteins.

A number of methods have been developed for the prediction of nucleic acids-binding residues. For example, DP-Bind (Hwang *et al.*, 2007), DBS_PSSM (Ahmad and Sarai, 2005), MetaDBSite (Si *et al.*, 2011), ProteDNA (Chu *et al.*, 2009), EL_PSSM-RT (Zhou *et al.*, 2017), DR_bind (Chen *et al.*, 2012) and SPOT-Seq (DNA) (Zhao *et al.*, 2014) are specially designed to predict DNA-binding residues. While Pprint (Kumar *et al.*, 2008), RNABindR (Terribilini *et al.*, 2007) and Meta2 (Puton *et al.*, 2012) work for protein–RNA binding residues only. There are a few methods that work for the prediction of both DNA- and RNA-binding residues, such as BindN (Wang and Brown, 2006), BindN+ (Wang *et al.*, 2010) and DRNApred (Yan and Kurgan, 2017). In the work of (Zhang *et al.*, 2017), three hallmarks of DNA-, RNA- and protein-binding residues were analyzed, which should be useful in the development of new prediction methods. Recently a new class of method called DisoRDPbind was developed for the prediction of disordered RNA, DNA and protein binding regions (Peng and Kurgan, 2015; Peng *et al.*, 2017). The Kurgan group pointed out that there are cross predictions between DNA- and RNA-binding residues in existing methods (Yan *et al.*, 2016; Yan and Kurgan, 2017). For a comprehensive review and assessment of existing methods, please refer to the studies in Miao and Westhof (2015), Yan *et al.* (2016), Zhang *et al.* (2017) and Zhao *et al.* (2013), which suggest that predictive accuracy of existing methods is relatively low.

In this work, we will present two new methods to predict DNA-/RNA-binding residues. The first is SVMnuc, an *ab-initio* method using features from three complementary sequence profiles. The second is NucBind, a consensus approach by combing SVMnuc with the template-based approach COACH-D (Wu *et al.*, 2018), to enhance the accuracy and robustness of the prediction.

2 Materials and methods

2.1 Benchmark datasets

Three benchmark datasets are used to assess and compare our methods with others: YFK16, YK17 and MW15, which were collected from the recent studies (Miao and Westhof, 2015; Yan *et al.*, 2016; Yan and Kurgan, 2017). Each dataset consists of 2–8 subsets for protein–DNA and/or protein–RNA binding. The detailed information about these datasets is summarized in Table 1. The structures in these datasets were originally constructed from the Protein Data Bank (PDB) (Rose *et al.*, 2017). A residue is defined as a DNA-/RNA-binding residue if one of the atomic distances between this residue and the DNA/RNA molecule are smaller than a specified distance cutoff. Two cutoffs were used in the above datasets: 3.5 and 5 Å. These datasets are available at http://yanglab.nankai.edu.cn/NucBind/benchmark/.

YFK16. This dataset was from the review (Yan *et al.*, 2016), which was constructed before 2013. Both cutoffs of 3.5 and 5 Å

Table 1. Summary of the benchmark datasets

Dataset	Cutoff (Å)	#Training	#Test	Date of PDB
YFK16_DNA	3.5/5	309/311	47/48	[1976, 2013]
YFK16_RNA	3.5/5	158/158	17/17	[1976, 2013]
YK17_DNA	3.5	339	49	[1976, 2016]
YK17_RNA	3.5	161	33	[1976, 2016]
MW15_DNA	5	NA	31	[2014, 2016]
MW15_RNA	5	NA	15	[2014, 2016]

were considered in this dataset. The structures released before/after 2010 were used for training/test. The sequence identity between the training and test proteins is less than 30%. A unique feature of this dataset is that the binding annotations for the structures in the dataset were enriched by transferring the annotations from other similar proteins in PDB.

YK17. This dataset is an extension of YFK16 (Yan and Kurgan, 2017) by inclusion of more structures released before 2016. A cutoff of 3.5 Å was considered in this dataset. The division of the training and the test sets is similar to the dataset YFK16 and the sequence identity between the training and the test proteins is less than 30% as well.

MW15. This dataset was from the work (Miao and Westhof, 2015), collected after 2014. The sequence identity between this dataset and others used for training the assessed methods is less than 25%. There is no training set in this dataset and it can be used as an independent test dataset, which includes 31 DNA-binding proteins and 15 RNA-binding proteins.

2.2 Overall architecture of the NucBind algorithm

There is no single algorithm to work well for all targets. The combination of complementary algorithms is an effective way to make stable and accurate predictions. We have developed the COACH algorithm for template-based protein–ligand binding residues prediction (Yang *et al.*, 2013a), which was recently improved in COACH-D by the inclusion of molecular docking (Wu *et al.*, 2018). COACH-D has been consistently ranked as the No.1 method in the weekly CAMEO-LB experiments (https://www.cameo3d.org/), in a period of about three years. The success of COACH-D is mainly attributed to the combination of five complementary algorithms. However, COACH-D does not work well in case that no homologous templates are available. To solve this problem, we developed a new *ab-initio* method SVMnuc for DNA-/RNA-binding residues prediction, which was then combined with COACH-D, resulting to another method NucBind.

The overall flowchart of NucBind is shown in Figure 1. The query sequence is submitted to three programs PSI-BLAST (Altschul *et al.*, 1997), HHblits (Remmert *et al.*, 2012) and PSIPRED (Jones, 1999), to generate two sequence profiles and predict its secondary structure profile, respectively (the panel inside the dashed frame). These profiles are used to extract a comprehensive set of features, which are fed into SVMnuc for *ab-initio* prediction. In addition, the query sequence is fed into the I-TASSER Suite (Yang *et al.*, 2015) to generate a structure model (the panel inside the solid frame). The model is submitted to COACH-D to make template-based prediction. The predictions from SVMnuc and COACH-D are combined by NucBind. The NucBind prediction is taken from COACH-D if its confidence score is higher than a dataset-specific cutoff, which was tuned based on the YFK16 training sets. Otherwise, the NucBind prediction is taken from SVMnuc.
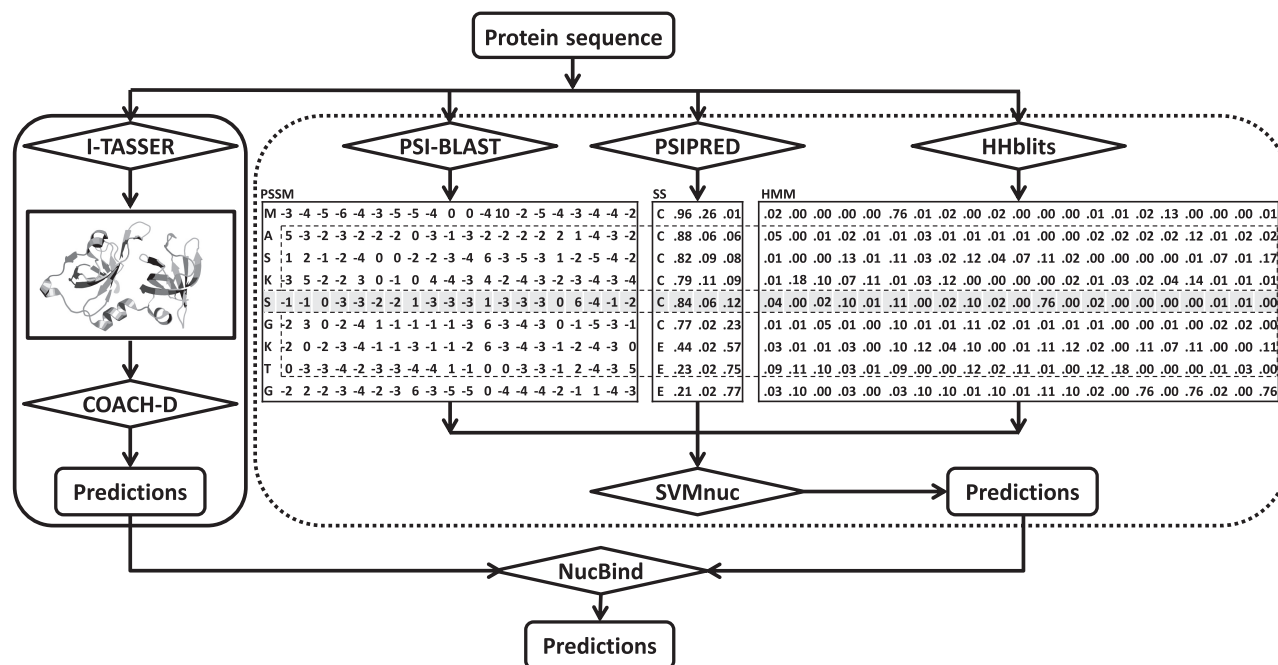
Protein sequence

I-TASSER   **PSI-BLAST**   **PSIPRED**   **HHblits**

PSSM

```
M  -3 -4 -5 -6 -4 -3 -5 -5 -4  0  0 -4 10 -2 -5 -4 -3 -4 -4 -2
A   5 -3 -2 -3 -2 -2 -2  0 -3 -1 -3 -2 -2 -2 -2  2  1 -4 -3 -2
S   1  2 -1 -2 -4  0  0 -2 -2 -3 -4  6 -3 -5 -3  1 -2 -5 -4 -2
K   3  5 -2 -2  3  0 -1  0  4 -4 -3  4 -2 -4 -3 -2 -3 -4 -3 -4
S  -1 -1  0 -3 -2 -2  1 -3 -3 -3  1 -3 -3  0  6 -4 -1 -2
G  -2  3  0 -2 -4  1 -1 -1 -1 -1 -3  6 -3 -4 -3  0 -1 -5 -3 -1
K  -2  0 -2 -3 -4 -1 -1 -3 -1 -1 -2  6 -3 -4 -3 -1 -2 -4 -3  0
T   0 -3 -3 -4 -2 -3 -3 -4 -4  1 -1  0  0 -3 -3 -1  2 -4 -3  5
G  -2  2 -2 -3 -4 -2 -3  6 -3 -5 -5  0 -4 -4 -4 -2 -1  1 -4 -3
```

SS

```
C  .96 .26 .01
C  .88 .06 .06
C  .82 .09 .08
C  .79 .11 .09
C  .84 .06 .12
C  .77 .02 .23
E  .44 .02 .57
E  .23 .02 .75
E  .21 .02 .77
```

HMM

```
.02 .00 .00 .00 .00 .76 .01 .02 .00 .02 .00 .00 .00 .01 .01 .02 .13 .00 .00 .00 .01
.05 .00 .01 .02 .01 .01 .03 .01 .01 .01 .01 .00 .00 .02 .02 .02 .02 .12 .01 .02 .02
.01 .00 .00 .13 .01 .11 .03 .02 .12 .04 .07 .11 .02 .00 .00 .00 .01 .07 .01 .17
.01 .18 .10 .07 .11 .01 .03 .12 .00 .00 .00 .00 .01 .03 .02 .04 .14 .01 .01 .00
.04 .00 .02 .10 .01 .11 .00 .02 .10 .02 .00 .76 .00 .02 .00 .00 .00 .01 .01 .00
.01 .01 .05 .01 .00 .10 .01 .01 .11 .02 .01 .01 .01 .00 .00 .01 .00 .02 .02 .00
.03 .01 .01 .03 .00 .10 .12 .04 .10 .00 .01 .11 .12 .02 .00 .11 .07 .11 .00 .00 .11
.09 .11 .10 .03 .01 .09 .00 .00 .12 .02 .11 .01 .00 .12 .18 .00 .00 .00 .01 .03 .00
.03 .10 .00 .03 .00 .03 .10 .10 .01 .10 .01 .11 .10 .02 .00 .76 .00 .76 .02 .00 .76
```

COACH-D → **Predictions**

SVMnuc → **Predictions**

NucBind → **Predictions**

Fig. 1. The flowchart of the proposed methods. The combination of COACH-D and SVMnuc is done by the ensemble method NucBind

2.3 Template-based prediction by COACH-D

COACH-D is a general-purpose template-based method for protein–ligand binding residues prediction, which combines five individual methods. The prediction is made by transferring the binding residues from homologues ligand-binding templates in the BioLiP database (Yang *et al.*, 2013b). Note that COACH-D does not differentiate between ligand types. To make COACH-D faster and work for DNA-/RNA-binding residues prediction, we restricted its template library to protein–DNA/–RNA complexes. To make fair comparison with other methods, all structure templates and ligand-binding templates with > 30% sequence identity to the query sequence were excluded, in the procedures of both structure modeling and binding residues prediction, respectively.

2.4 Feature design for the *ab-initio* method SVMnuc

The panel inside the dashed frame of Figure 1 is the flowchart of SVMnuc. As the binding residues are evolutionarily more conserved than others, the protein sequence is first submitted to three programs to generate three complementary sequence profiles. Then a comprehensive set of features are extracted from these profiles to encode each residue in a protein. The resulting feature vectors are finally fed into support vector machine (SVM) for the prediction of DNA-/RNA-binding residues. Let L denote the number of residues in a protein.

PSI-BLAST profile. The query sequence is searched by the sequence-profile alignment tool PSI-BLAST (with parameters '-j 3 -h 0.001') through the NCBI non-redundant sequence database, with the sequence profile represented in the form of a position-specific scoring matrix (PSSM) of dimension $L*20$. Each element x in PSSM is converted to the range of (0, 1) by $1/[1 + \exp(-x)]$.

PSIPRED profile. One of the most popular tools PSIPRED was applied to predict the three-state secondary structure (SS) profile. This profile provides the probabilities of each residues folding into one of the three states: alpha, beta and random coil. Thus, the dimension of the SS profile is $L*3$.

HHblits profile. The profile hidden Markov models (HMMs) have been successfully used for protein structure prediction based on the HMM-HMM alignment. It was demonstrated that the alignment generated by HHblits is more accurate than that by PSI-BLAST (Remmert *et al.*, 2012). In this study, the HMM profile was generated by searching the query sequence against the database uniprot20_2015_06 using HHblits. The dimension of HMM profile is $L*30$, but only the first 20 columns are used in this study. The integers in HMM are equal to 1000 times the negative logarithm of the amino acid frequency. Thus, each element x in the HMM profile is converted to a frequency number by the inverse transform $2^{-0.001*x}$.

For each residue in a protein, window-based features are extracted from the three profiles generated above. A residue can be simply represented by the 43 (=20 + 3 + 20) elements in the three profiles. However, the binding residues are not independent with each other because the residues inside a binding pocket have a higher probability to be in contact with DNA/RNA molecules. Thus, a sliding window is used to incorporate the effects of neighboring residues. For a window with size w, the total number of features extracted is $43*w$. The optimal window size remains to be determined by experiments on the training set YFK16_DNA_3.5. The kernel of radial basis function (RBF) is used based on the experimental results in Supplementary Figure S1. The regularization factor C and the kernel parameter γ are optimized based on the 5-fold cross-validation on the training set. The LIBSVM package (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) was used for the implementation of SVM.

2.5 Performance evaluation

The performance of the proposed methods is assessed by six metrics, four for assessing binary prediction, and two for the prediction with propensity score. The first three metrics for the binary prediction are Precision (Pre), Recall (Rec), and Matthews correlation coefficient (MCC).

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (2)$$

where TP (true positive) is the number of correctly predicted binding residues, TN (true negative) is the number of correctly predicted non-binding residues, FP (false positive) is the number of non-binding residues predicted as binding residues, and FN (false negative) is the number of binding residues predicted as non-binding residues. The value of a metric equals zero when the denominator is zero. The higher the above metrics are, the better the prediction is. MCC ranges from -1 to 1 and it is suitable for assessing data with imbalanced distribution.

The fourth metric is the ratio of the cross prediction (Rc) between DNA- and RNA-binding residues, proposed in the works Yan *et al*. (2016) and Yan and Kurgan (2017). It is defined as the fraction of native DNA-binding residues that are predicted as RNA-binding residues or the fraction of native RNA-binding residues that are predicted as DNA-binding residues. Lower value of Rc indicates better prediction.

The last two metrics are for assessing the prediction with propensity score. One is the area under the receiver operating characteristic curve (AUC) and the other is the ratio Rr. The AUC value is between 0 and 1 and the higher AUC value, the better the prediction is. The area under the lower part of the curve (AULC) was introduced in Yan and Kurgan (2017) to measure the performance of a method at the low false positive rate regions. Because the scale of AULC is small, the ratio (Rr) of a method's AULC over the random prediction's AULC is used to reflect the performance. Higher value of Rr indicates more accurate prediction.

Similar to Yan *et al*. (2016) and Yan and Kurgan (2017), for each of the three datasets, the assessments were conducted on a combined set of DNA- and RNA-binding proteins. When assessing for the prediction of DNA-binding (resp. RNA-binding) residues, the native labels for RNA-binding (resp. DNA-binding) proteins are set to 0. The purpose of doing this is to measure the cross predictions between DNA- and RNA-binding residues.

3 Results and discussions

3.1 Optimization of parameters for SVMnuc

All parameters including the window size w and the values for C and γ in SVM, are tuned to maximize the AUC on the training set YFK16_DNA_3.5 based on 5-fold cross validation. For each fixed window size, the default values of C and γ were used. The AUCs and MCCs for different window sizes are shown in Supplementary Figure S2. From the figure, we can see that the optimal window size is 15, at which the AUC and MCC are 0.85 and 0.41, respectively. After the optimal window size is determined, a grid search was performed to optimize C and γ, with more values tested: C in $[2^0-2^7]$ and γ in $[2^{-1}-2^{-8}]$. The final values for C and γ are 2 and 2^{-7}, respectively. For the sake of generality, these values are also used for other datasets without further optimizations.

3.2 Analysis of feature contribution

As the features in SVMnuc are from the PSI-BLAST profile (PSSM), the PSIPRED profile (SS) and the HHblits profile (HMM), their contributions are investigated by dividing the features into seven
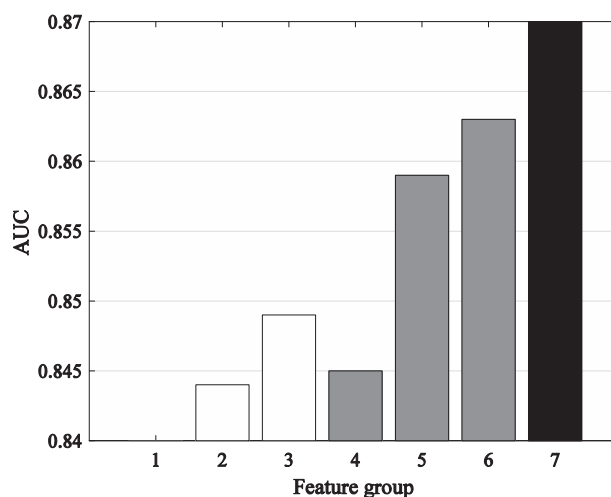

Fig. 2. Predictive performance of SVMnuc on the training set YFK16_DNA_3.5 from different feature groups. The AUC for the feature group 1 is not shown in the figure due to the low value (0.598)

categories: three individual feature groups, (1) SS, (2) HMM and (3) PSSM; combination of two feature groups: (4) SS + HMM, (5) SS + PSSM and (6) HMM + PSSM; and all features (7) ALL (i.e. PSSM+ SS+HMM).

The 5-fold cross validation AUC for each group of features on the training set YFK16_DNA_3.5 is presented in Figure 2. It shows that the AUC for the SS/PSSM feature is the lowest/highest, when using the individual feature group only (white bars in Fig. 2). Improvements were achieved by combing any two feature groups (gray bars in Fig. 2). The highest AUC is achieved when all the features are used together (black bars in Fig. 2). Statistical tests were performed to judge if the AUC improvements by the combined feature groups are significant or not, similar to the procedure used in Meng *et al*. (2018) and Yan and Kurgan (2017). The P-values for the tests are shown in the Supplementary Table S1. It indicates that the improvements by combing more feature groups are all significant at P-value < 0.05 (with an exception of SS + HMM versus HMM). For example, the P-values are 3.6×10^{-7} and 1.0×10^{-11} for the improvement with the combined feature group 6 (i.e. features from PSSM+HMM) over the individual feature group 2 (features from HMM) and 3 (features from PSSM), respectively. These data suggest that the three feature groups are complementary to each other.

3.3 Performance on the training sets of YFK16

The AUCs for COACH-D, SVMnuc and NucBind on the training sets of YFK16 are presented in Supplementary Figure S3. It shows that the template-based method COACH-D has lower AUC than SVMnuc, probably due to the fact that close homologous templates (sequence identity $>30\%$) have been excluded in out experiments for fair comparison. Another reason is that COACH-D was not trained to optimize AUC. The residue-specific propensity scores were calculated based on the majority votes from template-query alignments, making many residues with scores of 0. However, improved AUCs were obtained in NucBind for all datasets by combing the predictions from COACH-D and SVMnuc. The P-values of the AUC differences between these three methods are presented in Supplementary Table S2, which shows that the improvements of

Table 2. Comparison with other methods on the independent test sets based on AUC

Method	Dataset							
	YFK16[a]		YFK16[b]		YK17[a]	YK17[b]	MW15[a]	MW15[b]
Cutoff (Å)	3.5	5	3.5	5	3.5	3.5	5	5
Pprint	NA	NA	0.667	0.629	NA	0.66	NA	NA
RNABindR	NA	NA	0.712	0.694	NA	0.73	NA	NA
DP-Bind(klr)	0.794	0.770	NA	NA	0.76	NA	NA	NA
BindN+	0.806	0.773	0.738	0.704	0.79	0.67	NA	NA
DBS_PSSM	0.810	0.784	NA	NA	0.77	NA	NA	NA
DRNApred	NA	NA	NA	NA	0.77	0.67	NA	NA
DRNApred*	0.757	0.738	0.687	0.666	0.77	0.67	0.725	0.467
COACH-D	0.771	0.764	0.711	0.732	0.69	0.64	0.713	0.579
SVMnuc	0.833	0.821	0.784	0.785	0.80	0.74	0.829	0.789
NucBind	**0.834**	**0.822**	**0.801**	**0.803**	**0.81**	**0.75**	**0.830**	**0.794**

Note: The results for other methods are taken directly from Yan *et al*. (2016) and Yan and Kurgan (2017). The highest values are highlighted in bold type.
[a]DNA-binding;
[b]RNA-binding;
*The server version of DRNApred.

NucBind over COACH-D and SVMnuc are significant at the significance level of 0.05, except on the set YFK16_DNA_5.

3.4 Comparison with other methods
To demonstrate the effectiveness and robustness of the proposed method, we compare NucBind with other state-of-the-art methods on the independent test sets. As DRNApred is one of the latest methods for DNA-/RNA-binding residues prediction and it provides a web server for use (Yan and Kurgan, 2017), we submitted the protein sequences from the test sets to the server and calculated the corresponding metrics for the predictions. We can see the results for the two subsets in the YK17 dataset are identical to the ones reported in Yan and Kurgan (2017). Thus it should be fair to compare our methods with the server version of DRNApred.

The AUCs for all methods are shown in Table 2 and the data for MCC and Rc are available in Supplementary Tables S3 and S4. When measured by AUC, the template-based approach COACH-D does not have advantage over other methods, such as BindN+ and DBS_PSSM, with the above-mentioned reasons. On the contrary, the *ab-initio* method SVMnuc achieves consistently higher AUCs over other methods, which are improved further in NucBind. For example, compared with BindN+, which has competitive accuracy for the prediction of both DNA- and RNA-binding residues, the AUCs of NucBind are 2.5–14.1% higher than BindN+ on the compared datasets. Statistical tests indicate that the AUC improvements of NucBind over BindN+ and other methods are significant at the significance level of 0.05.

We note that the AUC and MCC of the most recent method DRNApred are not the highest. This is because the main purpose of DRNApred is not to predict the binding residues with high accuracy but to reduce the cross predictions between DNA- and RNA-binding residues. The metric ratio (Rc) has been defined to measure the cross predictions. We made a comprehensive comparison between the proposed methods and DRNApred on all datasets based on Rc and other metrics, with results presented in Supplementary Table S5. It shows that the proposed methods have higher MCC, AUC, Pre, Rec and Rr than DRNApred for all datasets. However, DRNApred has the lowest or the second lowest value of Rc for all datasets except MW15_DNA. For this dataset, the cross predictions Rc for all methods are surprisingly high (>0.2), which might be because some proteins in this dataset are able to bind to both DNA

and RNA. For example, the complex structure of the protein (PDB ID: 4OL8) contains both DNA and RNA.

From Supplementary Table S5, we can see that compared with other methods, COACH-D's performance is satisfactory in terms of Rr, Rc and MCC for most datasets. For example, on the dataset YFK16_RNA_3.5, COACH-D's Rr, Rc and MCC are 24.7, 0.01 and 0.36, respectively. In comparison, the corresponding values for DRNApred are 2.8, 0.02 and 0.06 respectively. This good performance of COACH-D can be attributed to the full use of the template information during the inference of both the binding residues and the type of bound ligand. COACH-D predicts the binding residues as DNA (resp. DNA) binding if more than half of the template ligands are DNA (resp. RNA). The usage of template information in COACH-D results to a lower ratio of cross prediction than other *ab-initio* methods, including DRNApred and SVMnuc. The high values of Rr for COACH-D can be attributed to the high values of Pre. Though with such advantages, the AUC for COACH-D is not as high as DRNApred. This suggests that template-based predictions are complementary with *ab-initio* predictions. Thus the combination of COACH-D and SVMnuc leads to the most stable method NucBind. Statistical tests were performed to assess the difference between the proposed methods and DRNApred based on the AUCs presented in Table 2. The *P*-values in Supplementary Table S6 indicate that COACH-D alone does not outperform DRNApred. However SVMnuc and NucBind's AUC improvements over DRNApred are significant at the level of 0.05 for all datasets.

3.5 Case studies
Figure 3 shows two examples that NucBind makes satisfactory prediction for DNA- and RNA-binding residues, respectively. The first example is a DNA-binding protein 'C.Esp1396I bound to a 25 base pair operator site' (PDB ID: 4IWR), from the set YK17_DNA. In this example, there are 17 DNA-binding residues. NucBind predicted 19 binding residues and 16 of them are true positives (Fig. 3B). This leads to a high MCC of 0.857, Pre of 0.842 and Rec of 0.941. On the contrary, DRNApred predicted 27 binding residues and 13 of them are true positives and the remaining 14 are false positives (Fig. 3A). This translates to lower MCC, Pre and Rec of 0.459, 0.481 and 0.765, respectively.

The second example is a RNA-binding protein 'Bacteriophage Qbeta coat protein in complex with RNA operator hairpin' (PDB
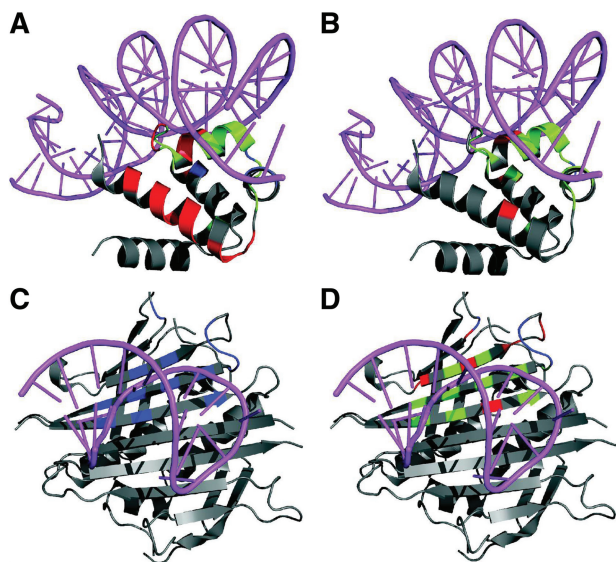
Fig. 3. Two examples of DNA- and RNA-binding residues predictions made by NucBind and DRNApred. (A) and (C) are for the predictions by DRNApred, while (B) and (D) are for NucBind. The protein structure is shown in gray cartoon and DNA/RNA molecules structures are shown in magenta cartoon. TP, FP and FN are shown in green, red and blue cartoon, respectively

Fig. 4. An example to illustrate the improved ranking of the docking models, with the DNA-binding domain structure of the tumor suppressor p53. The protein structure is shown in orange surface. The reference DNA structure, default top model and re-ranked top model are shown in red, green and blue cartoons, respectively

ID: 4L8H), from the set YK17_RNA. There are 14 RNA-binding residues in this protein. A total of 17 residues were predicted as binding residues by NucBind and 10 of them are true positives (Fig. 3D). As can be seen from the figure, other predicted binding residues (false positives) are in fact around the interface of the protein–RNA complex structure. Note that the distance cutoff used for this target is 3.5 Å. These false positives are very possible to become true positives when increasing the distance cutoff in the definition of a binding residue. Nevertheless, the prediction by NucBind for this example is also satisfactory with the MCC, Pre and Rec of 0.597, 0.588 and 0.714 respectively. When increasing the distance cutoff to 5 Å, the number of false positives reduced from 7 to 3. This makes the MCC for the NucBind predictions increase to 0.694. However, DRNApred does not predict any binding residues for this target and the values for all metrics are 0 (Fig. 3C).

3.6 Application of predicted binding residues in molecular docking

As a demonstration for the application of the proposed methods in the field of structure biology, we employed the predicted binding residues to improve the ranking of docking models. The protein–DNA/RNA docking method HDOCK (Yan *et al.*, 2017) was used for this purpose. A similar strategy has been shown to be effective for ranking ligand-binding poses in the COACH-D algorithm (Wu *et al.*, 2018), where the ligands are small molecules rather than DNA and RNA.

The structure of the DNA-binding domain of the well-known tumor suppressor p53 is used as an illustration here. Self docking was not performed as it turned out to be very trivial for HDOCK to find the optimal binding pose. Instead, the unbound structure of p53 (PDB ID: 2OCJ) and the DNA structure from the bound structure (PDB ID: 5LGY) of p53 were submitted to the HDOCK server to perform protein–DNA docking. Here, the bound structure of p53 was obtained from the mTM-align server by searching the unbound structure against the server's structure database, which reports that the root-mean-square deviation (RMSD) between the bound and
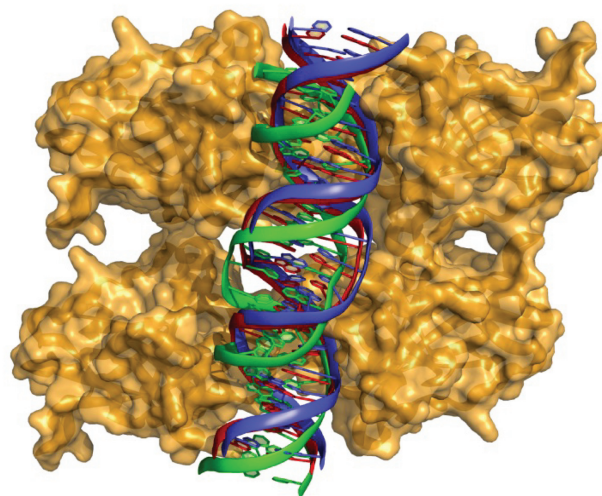
unbound structures is 0.49 Å (Dong *et al.*, 2018). We superimposed the bound structure to the unbound structure to generate a DNA-binding complex structure for the unbound structure. This manually generated DNA structure is used as a reference to assess the accuracy of the docking models.

A total of 12 DNA-binding residues were predicted by NucBind, which were used as reference to rank the docking models as follows. First, the DNA-binding residues in each docking model were calculated at 3.5 Å distance cutoff. Second, the precision of the binding residues from each model was computed. Third, the docking models were ranked based on their precision values. The ligand RMSDs between the docking models and the reference structure were computed. The results are summarized in Figure 4, in which the protein structure is shown in orange surface. It suggests that top model (in blue) by our method is very close to the reference (in red) with 2.32 Å RMSD. In comparison, the default top model (ranked by the HDOCK's docking score, in green) is relatively far away from reference, with 41.19 Å RMSD. This example shows that the predicted binding residues can be useful for docking. However, due to the difficulty in generating complex structure for unbound structures manually, only one example was given here as an illustration.

3.7 Limitations of the proposed methods

Though the proposed methods were demonstrated to have advantages over other state-of-the-art methods, there do exist some limitations. The first is the slow speed and high computational cost. It takes more time and computing resource to generate multiple sequence profiles in SVMnuc and run template-query alignments in COACH-D. As the server is built on a computer cluster with 100 CPU cores, the programs are executed in parallel. In general, it takes about 0.5 hour to return the predictions for each job submission. The second is the relatively high ratio of cross prediction of DNA- and RNA-binding residues in SVMnuc and NucBind. This is mainly caused by the *ab-initio* method SVMnuc, which seems to be a common issue for most *ab-initio* methods, such as BindN+ and DBS_PSSM (please refer to Supplementary Table S4). This problem was pointed out in Yan and Kurgan (2017) and the cross prediction was significantly reduced in the work of DRNApred with an

elaborately designed training strategy. Except DRNApred, the cross prediction in SVMnuc is much smaller than in other *ab-initio* methods, though not intentionally trained. We note that the cross prediction in COACH-D is even smaller than DRNApred for some datasets. However, its combination with SVMnuc did not reduce the cross prediction significantly. This is because the confidence scores for most of the COACH-D predictions are too low to be combined, mainly due to the stringent exclusion of homologous templates at 30% sequence identity. Nevertheless, in most real-world applications, the sequence identities between templates and query proteins are above this cutoff, resulting to predictions with higher confidence scores. It is thus anticipated that the cross prediction could be relieved in real-world applications.

4 Conclusions

Accurate recognition of the residues that bind nucleic acids can facilitate the study of uncharacterized protein–nucleic acids interactions. By utilizing three complementary sequence profiles, we first designed an *ab-initio* approach SVMnuc to predict the nucleic acids-binding residues. Enhanced prediction was further achieved by combining SVMnuc with our previously developed template-based approach COACH-D, which results to the algorithm NucBind. Benchmark tests show that the proposed methods consistently outperform other state-of-the-art methods for the prediction of nucleic acids-binding residues. Though with high prediction accuracy, cross prediction between DNA and RNA-binding residues, a common issue in many methods, was also observed in SVMnuc and NucBind. More efforts are needed to reduce the cross prediction in future work, such as those done in the method DRNApred. The success of NucBind is attributed to the utilization of the improved features from three complementary sequence profiles in SVMnuc and the consensus of complementary methods.

References

Ahmad,S. and Sarai,A. (2005) PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, **6**, 33.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Chen,Y.C. *et al.* (2012) DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res.*, **40**, W249–W256.

Chu,W.Y. *et al.* (2009) ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. *Nucleic Acids Res.*, **37**, W396–W401.

Dong,R. *et al.* (2018) mTM-align: a server for fast protein structure database search and multiple protein structure alignment. *Nucleic Acids Res.*, **46**, W380–W386.

Hwang,S. *et al.* (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, **23**, 634–636.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Kumar,M. *et al.* (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins*, **71**, 189–194.

Meng,Q. *et al.* (2018) CoABind: a novel algorithm for Coenzyme A (CoA)- and CoA derivatives-binding residues prediction. *Bioinformatics*, **34**, 2598–2604.

Miao,Z. and Westhof,E. (2015) A large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comput. Biol.*, **11**, e1004639.

Peng,Z. and Kurgan,L. (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.*, **43**, e121.

Peng,Z. *et al.* (2017) Prediction of disordered RNA, DNA, and protein binding regions using DisoRDPbind. *Methods Mol. Biol.*, **1484**, 187–203.

Puton,T. *et al.* (2012) Computational methods for prediction of protein–RNA interactions. *J. Struct. Biol.*, **179**, 261–268.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Rose,P.W. *et al.* (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.*, **45**, D271–D281.

Si,J. *et al.* (2011) MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. *BMC Syst. Biol.*, **5 (Suppl. 1)**, S7.

Terribilini,M. *et al.* (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.*, **35**, W578–W584.

von Hippel,P.H. *et al.* (1984) Protein–nucleic acid interactions in transcription: a molecular analysis. *Annu. Rev. Biochem.*, **53**, 389–446.

Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.

Wang,L. *et al.* (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.*, **4**, S3.

Wong,K.C. (2017) MotifHyades: expectation maximization for de novo DNA motif pair discovery on paired sequences. *Bioinformatics*, **33**, 3028–3035.

Wu,Q. *et al.* (2018) COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res.*, **46**, W438–W442.

Yan,J. *et al.* (2016) A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief. Bioinform.*, **17**, 88–105.

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

Yan,Y. *et al.* (2017) HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res.*, **45**, W365–W373.

Yang,J. *et al.* (2013a) Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, **29**, 2588–2595.

Yang,J. *et al.* (2013b) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.*, **41**, D1096–D1103.

Yang,J. *et al.* (2015) The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*, **12**, 7–8.

Zhang,J. *et al.* (2017) Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief. Bioinform.* doi: 10.1093/bib/bbx168

Zhang,X. and Liu,S. (2017) RBPPred: predicting RNA-binding proteins from sequence using SVM. *Bioinformatics*, **33**, 854–862.

Zhao,H. *et al.* (2014) Predicting DNA-binding proteins and binding residues by complex structure prediction and application to human proteome. *PLoS One*, **9**, e96694.

Zhao,H. *et al.* (2013) Prediction of RNA binding proteins comes of age from low resolution to high resolution. *Mol. Biosyst.*, **9**, 2417–2425.

Zhou,J. *et al.* (2017) EL_PSSM-RT: dNA-binding residue prediction by integrating ensemble learning with PSSM relation transformation. *BMC Bioinform.*, **18**, 379.