Briefings in Bioinformatics, 23(1), 2022, 1–19

https://doi.org/10.1093/bib/bbab411
Problem Solving Protocol

Dissecting and predicting different types of binding sites in nucleic acids based on structural information

Zheng Jiang, Si-Rui Xiao and Rong Liu

Corresponding author: Rong Liu, Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, P. R. China. Tel.: +86-27-87280877; Fax: +86-27-87280877; E-mail: liurong116@mail.hzau.edu.cn

Abstract

The biological functions of DNA and RNA generally depend on their interactions with other molecules, such as small ligands, proteins and nucleic acids. However, our knowledge of the nucleic acid binding sites for different interaction partners is very limited, and identification of these critical binding regions is not a trivial work. Herein, we performed a comprehensive comparison between binding and nonbinding sites and among different categories of binding sites in these two nucleic acid classes. From the structural perspective, RNA may interact with ligands through forming binding pockets and contact proteins and nucleic acids using protruding surfaces, while DNA may adopt regions closer to the middle of the chain to make contacts with other molecules. Based on structural information, we established a feature-based ensemble learning classifier to identify the binding sites by fully using the interplay among different machine learning algorithms, feature spaces and sample spaces. Meanwhile, we designed a template-based classifier by exploiting structural conservation. The complementarity between the two classifiers motivated us to build an integrative framework for improving prediction performance. Moreover, we utilized a post-processing procedure based on the random walk algorithm to further correct the integrative predictions. Our unified prediction framework yielded promising results for different binding sites and outperformed existing methods.

Key words: RNA binding site; DNA binding site; integrative algorithm; ensemble learning; structural homology

Introduction

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are the two main classes of nucleic acids (NAs), which are essential for the continuity of life [1–3]. DNA is the genetic material in all living organisms, while RNA is the key player in the making of proteins under the direction of DNA. The biological functions of DNA and RNA are generally dependent on their interactions with other molecules, such as small ligands, proteins and NAs. For instance, DNA and RNA molecules could serve as potential drug targets for various diseases through interacting with small molecules [4–7]. Protein–NA and NA–NA interactions play indispensable roles in many biological processes, such as transcriptional regulation, protein synthesis and cell development [8–11]. Identification of the binding sites in DNA and RNA is the first step to understanding the mechanisms of these different classes of interactions. Although a diversity of experimental methods, including high-throughput sequencing techniques (e.g. ChIP-seq, CLIP-seq and Hi-C) [12–15], X-ray crystallography and nuclear magnetic resonance, have been commonly utilized to determine the NA-associated interactions and binding regions, the process is still time-consuming and labor-intensive. Accordingly, it is highly desirable to develop computational frameworks to analyze and predict binding nucleotides at a large scale.

Zheng Jiang is a PhD student at the College of Informatics, Huazhong Agricultural University. His research interests focus on structural bioinformatics.
Si-Rui Xiao is a master's student at the College of Informatics, Huazhong Agricultural University. Her research interests focus on structural bioinformatics.
Rong Liu is an associate professor at the College of Informatics, Huazhong Agricultural University. His research interests focus on structural bioinformatics.

Submitted: 7 July 2021; **Received (in revised form):** 26 August 2021

Over the past decade, a series of computational studies have explored the binding sites in DNA and RNA. For instance, intensive efforts have been devoted to predicting the interaction sites of DNA- and/or RNA-binding proteins based on high-throughput sequencing data [16–19]. These algorithms mainly adopted the features extracted from sequence fragments as the input of machine learning or deep learning models to detect the binding signatures in DNA or RNA sequences. However, this category of prediction models is only suitable for specific proteins with abundant experimental data. Moreover, these methods could not explicitly indicate whether a given nucleotide is involved in the binding activity. Based on the RNA tertiary structure, Zeng *et al.* [20] proposed the Rsite algorithm that calculated the distance from each nucleotide to other nucleotides and considered nucleotides corresponding to the extreme points in the distance curve as the functional sites. Furthermore, they found that the tertiary structure-based distance was correlated with the secondary structure-based distance and developed the Rsite2 algorithm by replacing the former with the latter [21]. Nevertheless, their methods were validated using a very limited number of RNA structures and could generate many false positives by considering all extreme points. Wang *et al.* [22] established a network-based approach termed RBind that converted RNA structures into nucleotide networks and used the degree and closeness measures to find ligand- and protein-binding nucleotides. Although RBind could achieve promising precision values, the recall values may be very low due to the high threshold of network features. Recently, Su *et al.* [23] developed the RNAsite algorithm based on the above two network properties and several structural and evolutionary descriptors. This approach clearly improved prediction accuracy, but can only be used to predict ligand-binding regions. He *et al.* [24] built a web server called HNADOCK to model RNA/DNA–RNA/DNA complex structures using RNA structure prediction and double-iterative knowledge-based scoring functions. Although this web server could provide putative binding regions for paired NA structures, it may have little applicability when facing individual structures. Collectively, the aforementioned works significantly prompted the development of NA binding site prediction, but the inherent limitations should be overcome.

Despite the progress achieved, some fundamental issues regarding the binding sites in NAs have yet to be investigated. As mentioned above, RNA and DNA could interact with different types of molecules (e.g. ligands, proteins and NAs). Regarding each type of interaction, the discrepancies between binding and nonbinding nucleotides remain largely unknown. It is also unclear whether there are similarities and differences among the binding sites of different interaction partners in RNA/DNA and between the binding sites in RNA and DNA. Over the past 30 years, in contrast, the different types of binding sites in proteins have been characterized and compared from multifaceted perspectives [25–28]. Therefore, it would be interesting to further ask whether NAs and proteins use a similar or different way to interact with their binding partners. Additionally, although many efforts have been made to identify binding residues in proteins [29–31], few works have explored whether these methodologies could be directly extended to the prediction of binding nucleotides in NAs. Further, it would be worth investigating whether unified or specific prediction models are needed for the two types of NAs and for their different classes of binding sites, but little attention has been given to this point. Because of the structural genomics efforts, an increasing number of the complex structures including NAs have been solved and deposited into the Protein Data Bank [32], which provides the possibility to address the aforementioned issues.

With these problems in mind, we attempted to perform a comprehensive analysis and prediction of binding sites in RNA and DNA. To characterize each nucleotide, we extracted a diversity of descriptors from the sequence, structure and preference aspects, some of which have been successfully applied to the prediction of functional residues in our previous studies [33–35]. Based on the well-collected datasets, we conducted a systematic comparison between binding and nonbinding sites in RNA and DNA and among different categories of binding sites. Unlike the existing algorithms that mainly depended on the feature-based prediction, we proposed an integrative framework that combined machine learning- and template-based models to predict binding nucleotides, because the usefulness of the hybrid strategy has also been shown in our algorithms for identifying binding residues [33–37]. Furthermore, we developed a post-processing step based on random walks to correct the predictions. The final prediction algorithm called NABS (Nucleic Acid Binding Site) showed promising performance on different classes of datasets and significantly outperformed existing methods. The web server and related data are available at http://liulab.hzau.edu.cn/NABS/.

Materials and methods

Data preparation

In this work, the molecules that interact with NAs (RNA/DNA) were classified into three categories: ligands, proteins and NAs. We therefore prepared a total of six datasets, including RNA–ligand (RL), RNA–protein (RP), RNA–NA (RN), DNA–ligand (DL), DNA–protein (DP) and DNA–NA (DN) interactions. The pipeline for collecting these datasets is given below. First, we obtained the complex structures including RNA/DNA using the advanced search interface in the PDB database. Only the crystal structures with a resolution better than 4 Å were reserved. The selected threshold was also used by existing studies [38–40] and could contribute to a greater number of structures for meaningful analyses (Supplementary Figure S1 available online at http://bib.oxfordjournals.org/). Second, we deleted the entries meeting any of the following criteria: (a) only backbone information was provided in the PDB file; (b) the length of NA chains was less than 20 nucleotides or more than 400 nucleotides and (c) the NA structures had less than three binding nucleotides within 4.5 Å of interacting molecules. Third, we categorized the reserved structures into six datasets according to the different types of interacting partners. For each dataset, as suggested by previous RNA-related studies [38–40], we utilized the cd-hit-est program to reduce sequence redundancy at 80% sequence identity and then performed additional filtering at 30% sequence identity using the BLASTClust program [41, 42]. As shown in Supplementary Figure S2A–F available online at http://bib.oxfordjournals.org/, the identities between pairs of NA chains in each dataset were generally less than 60%, showing that the sequence redundancy was effectively reduced. Finally, the RL, RP and RN datasets comprised 74, 170 and 102 RNA structures, respectively, while the DL, DP and DN datasets included 28, 336 and 33 DNA structures, respectively. Note that metal ion-binding sites were excluded from the ligand-binding datasets, because their binding properties may be different from those of nonmetal ligand-binding regions. For instance, metal ions could make contacts with the surface of RNA structures rather than the well-shaped pockets preferred by nonmetal small molecules [23]. Additionally, the

DN dataset only contained DNA–RNA complexes. Since many DNA molecules have the double helix structure and the chains are reversely complemented, we selected the chain with more binding nucleotides as a representative. As shown in Supplementary Figure S2G–I available online at http://bib.oxfordjournals.org/, only one DNA chain in the double helix mainly formed contacts with RNAs, while the two strands possessed comparable numbers of binding nucleotides for ligands and proteins. For each dataset, we used 70% of the structures as the training set (i.e. RL-TR, RP-TR, RN-TR, DL-TR, DP-TR and DN-TR) and the remaining 30% as the testing set (i.e. RL-TS, RP-TS, RN-TS, DL-TS, DP-TS and DN-TS).

Furthermore, we prepared the other three independent sets (i.e. RL-PS, RP-PS and RN-PS) composed of predicted RNA tertiary structures, each of which corresponded to one chain from the aforementioned testing sets of RNA. For each RNA sequence, we used the RNAfold program to predict its secondary structure [43]. In conjunction with the RNA FRABASE resource, the RNAComposer program was adopted to predict its tertiary structure using the corresponding sequence and secondary structure information [44]. A summary of our datasets is provided in Supplementary Figure S3 and Supplementary Table S1 available online at http://bib.oxfordjournals.org/.

Overview of NABS algorithm

As shown in Figure 1, the NABS algorithm was divided into four parts: a template-based module, a feature-based module, an integrative module and a post-processing module. The template-based method could retrieve the optimal reference structure for the query RNA/DNA by structural alignment and identify the binding nucleotides according to the predicted complex structures. However, this method might lose its applicability as the query could not find a reliable template or suffers from intensive conformational changes upon binding. The feature-based method was an ensemble prediction model by considering different machine learning algorithms, feature spaces and sample spaces. In contrast, this predictor may still provide effective predictions in the above scenarios, thereby complementing the limitations of our template method. Through the interplay between these two methods, the predictions from the integrative module were generated and could be further optimized. Due to the spatial clustering of binding nucleotides, the random walk with restart (RWR) algorithm was utilized to correct the results based on nucleotide interaction networks. To predict each type of binding site, the NABS algorithm was implemented based on the corresponding dataset.

Feature extraction

To quantitatively depict each nucleotide, we extracted three groups of descriptors: structural features, preference features and sequence features. These features have been proven to be effective in predicting functional sites in proteins, and most of them could be newly applied to the identification of binding sites in NAs.

Structural features

Relative solvent accessibility. Solvent accessibility refers to the accessible surface area (ASA) of a molecule that is exposed to a solvent. We calculated the ASA of each nucleotide in the NA monomers by the NACCESS program with a probe diameter of 1.5 Å as suggested by existing studies [23, 40, 45]. Further, we computed the relative solvent accessibility (RSA) that was the ratio between the ASA of a given nucleotide and its maximum ASA (i.e. *A*, *G* = 400 Å^2 and *C*, *U*, *T* = 350 Å^2). Sun *et al.* [40] proposed these maximum values based on the assumption that purines could be more exposed than pyrimidines, probably because the former has two carbon rings while the latter has only one [46, 47].

Depth and protrusion indices. Depth index (DPX) and protrusion index (CX) are originally used to describe the local concavity and convexity of proteins [48]. Here, we extended their applications to NAs. The DPX of an atom was defined as the distance between this atom and its nearest solvent accessible atom. The formula is as follows:

$$\mathrm{DPX}_i = \begin{cases} \min \|c_i - c_j\| & \mathrm{asa}_i = 0,\ \ \mathrm{asa}_j > 0 \\ 0 & \mathrm{asa}_i > 0 \end{cases}$$

where c_i and c_j denote the coordinates of atoms i and j, respectively, while asa_i and asa_j denote their solvent accessibilities.

The CX of an atom was defined as the ratio of the free volume within the sphere centered on this atom to the volume occupied by atoms. The formulas are as follows:

$$\mathrm{CX}_i = \frac{V_s - V_o}{V_o}$$

$$V_o = n \times \delta$$

where V_s denotes the volume of the sphere with a radius of 10 Å, and V_o denotes the occupied volume. n is the number of atoms in the sphere and δ is the average volume of atoms and was set to 20.1 Å^3. For each nucleotide, we calculated the average atomic indices as its DPX and CX values.

Laplacian norm. Laplacian norm (LN) has been successfully applied to the prediction of functional residues in our previous studies and could thus be suitable for the study of nucleotides [34, 49]. Note that when we prepared this manuscript, RNAsite used this feature, which was also motivated by our earlier works [23]. The LN measure of each nucleotide was defined as the distance from this nucleotide to the weighted center of its surrounding nucleotides, thereby representing its relative position in a geometric context. A high LN indicates a convex location in the NA structure, whereas a low LN implies a concave position. To generate this measure for each nucleotide, we should calculate the Laplace operator as follows:

$$\Omega_{ij}(\sigma) = \begin{cases} e^{-\frac{\|c_i - c_j\|^2}{\sigma^2}} & \mathit{if}\ \ |i-j| > 1 \\ 0 & \text{otherwise} \end{cases}$$

where c_i and c_j represent the average coordinates of all atoms in nucleotides i and j, respectively. The scale factor σ determines the weights of surrounding nucleotides. A low scale factor represents that adjacent nucleotides are mainly used to construct the geometric context, whereas a high value suggests that more distant nucleotides are used. For each NA, we chose the 0, 1/4, 1/2, 3/4 and 1 quantiles of the distance distribution of all the paired nucleotides as the scale factors. Finally, five weighted distances were calculated for each nucleotide as below and then converted

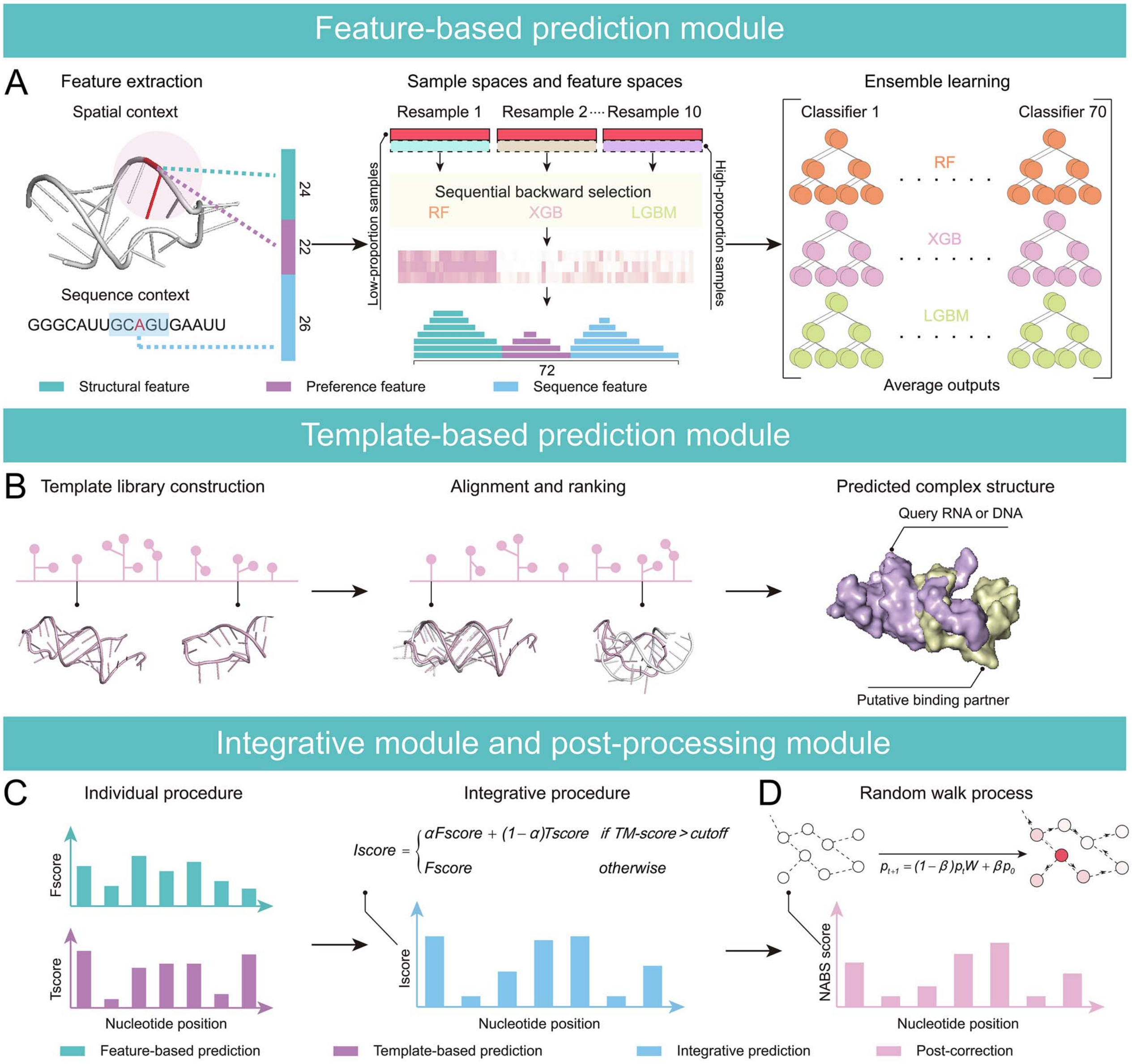

Figure 1. Schematic representation of NABS algorithm. (**A**) Feature-based prediction module. (**B**) Template-based prediction module. (**C**) Integrative prediction module. (**D**) Post-processing module.

into z-scores.

$$\mathrm{LN_i}\left(\sigma\right)=\left\|\mathrm{c_i}-\frac{\sum_j^{|\mathrm{i}-j|>1}\left[\mathrm{c_j}\times\Omega_{\mathrm{ij}}\left(\sigma\right)\right]}{\sum_j^{|\mathrm{i}-j|>1}\Omega_{\mathrm{ij}}\left(\sigma\right)}\right\|$$

Network features. Each NA structure could be converted into an interaction network, in which each node represents a nucleotide, and each edge denotes that there exists a contact between a pair of nucleotides. As suggested by Wang *et al.* [22], a contact was generated if the distance between one heavy atom of a nucleotide and that of the other nucleotide was less than 8 Å. To characterize each nucleotide, we computed four network measures, including degree, closeness, betweenness and clustering coefficient [50]. The resulting features were converted into z-scores.

Preference features

Nucleotide type preference. The binding propensities of amino acids are commonly used features for predicting binding sites in proteins [51]. Here, the numbers of binding and nonbinding nucleotides of NAs involved in a given type of interaction were counted and converted into proportions in terms of different nucleotide types. The preference for a nucleotide type (i.e. the binding propensity) was defined as the ratio between its proportion among binding sites and its proportion among nonbinding sites (Supplementary Figure S4 available online at http://bib.oxfordjournals.org/). This feature shows the binding tendency of different types of nucleotides.

Functional group preference. A nucleotide is composed of three functional groups: a sugar, a nitrogenous base and a phosphate group. The phosphate and sugar groups constitute the backbone of NAs. The statistical analysis of functional groups has been used in the study of protein–RNA interactions [52], and here, we extended this scheme to the other five interaction types. We counted the numbers of different functional groups involved in the binding process for each nucleotide type and calculated their corresponding proportions. This feature shows the binding preference of nucleotides at the level of functional groups.

Secondary structure preference. RNA secondary structures play important roles in the interactions between RNA and other molecules. For each RNA chain, the secondary structure state of a nucleotide was assigned by running the DSSR program [53]. We considered seven categories of secondary structures, including non-loop single-stranded segment, stem, hairpin, bulge, internal loop, multiple loop and pseudoknot. For each nucleotide type, we computed the numbers of binding and nonbinding nucleotides involved in different secondary structure states and converted them into proportions. The proportion matrices are shown in Supplementary Figure S5 available online at http://bib.oxfordjournals.org/. The preference was the ratio between the corresponding elements in the binding and nonbinding matrices. Because the secondary structure of DNA is relatively simple, this feature was not used to predict binding sites in DNA.

Spatial neighboring features

Aside from the native properties of the target nucleotide, its spatial neighboring context could also provide useful information for the identification of this nucleotide. Accordingly, we designed a descriptor, termed the spatial neighboring feature, from the structural and preference perspectives, which was defined as follows:

$$\mathrm{SNF}_i = \frac{\sum_{j \in D} f_j}{\mid D \mid}$$

where D represents the set of neighboring nucleotides physically interacting with the target nucleotide and f_j represents the structural or preference feature of nucleotide j generated in the above process.

Sequence features

Nucleotide composition. Besides the spatial neighborhood, we also considered the sequential neighborhood that was a linear window of consecutive nucleotides centered on the target nucleotide. For each sequence fragment, the compositions of four nucleotide types were calculated as follows:

$$C_r = \frac{n_r}{N} \qquad r \in \{A, G, C, U, T\}$$

where n_r denotes the number of nucleotide r and N is the length of the fragment.

Nucleotide transition. This feature is used to measure the preference for transitions from one nucleotide type to another type in adjacent positions, which is an extension of the composition–transition–distribution features for proteins [54, 55]. The formula is as follows:

$$T_{\mathrm{rs}} = \frac{n_{\mathrm{rs}} + n_{\mathrm{sr}}}{N-1} \qquad r, s \in \{A, G, C, U, T\}$$

where n_{rs} and n_{sr} denote the numbers of dinucleotides rs and sr, respectively. Therefore, each sequence fragment was represented by a 10-dimensional vector.

Nucleotide distribution. To describe the distribution of nucleotide types in the sequential neighborhood, we calculated the total number of nucleotides with a given type and recorded the indices of the first, 50% and 100% of these nucleotides. Based on the selected nucleotides, three descriptors for each nucleotide type were computed as follows:

$$D_r = \frac{\mathrm{Index}\,(r_n)}{N} \qquad r \in \{A, G, C, U, T\}$$

where r_n denotes the nth nucleotide with type r and Index(r_n) is the index of r_n in the fragment. Using the sequence 'GAUUUCAAGAC' as an example, the indices of the first, second (4*50%) and fourth (4*100%) adenines are 2, 7 and 10, respectively, so that the related descriptors are 0.18 (2/11), 0.64 (7/11) and 0.91 (10/11), respectively. As a result, each fragment was denoted by 12 (3*4) distribution descriptors. For these sequence features, the parameter N was determined based on Supplementary Figure S6 available online at http://bib.oxfordjournals.org/.

Feature-based prediction module

The feature-based prediction module was established on the ensemble strategy by fully using the interplay among different machine learning methods, sample spaces and feature spaces (Figure 1A). Here, three machine learning algorithms, including random forest (RF), extreme gradient boosting (XGB) and light gradient boosting machine (LGBM), were used to construct base classifiers. For our datasets, the ratio of binding nucleotides to nonbinding nucleotides was generally imbalanced. For instance, the ratio for the RL dataset was approximately 1:5.6, whereas the ratio for the DP dataset was approximately 1.5:1. The class imbalance problem could impact the performance of machine learning methods. Herein, we used the random undersampling scheme to solve this problem. For a given dataset, the sampling procedure was performed 10 times for the class with a higher proportion. In each iteration, all the low-proportion samples along with the same number of high-proportion samples were adopted. This process thus generated 10 different training sets. Then, the sequential backward selection (SBS) was performed on each training set to generate the optimal feature subset using the RFECV package from Scikit-learn [56]. The SBS method started with all features and iteratively eliminated the least important feature so that the remaining features could improve prediction performance to the greatest extent possible. Using the RF, XGB and LGBM as classifiers, we achieved three optimal feature subsets for each training set. Consequently, a total of 30 optimal feature subsets were generated for the 10 training sets. We constructed different feature spaces by selecting the features with the number of occurrences higher than a certain threshold. Here, the thresholds were set to 0, 5, 10, 15, 20, 25 and 30, respectively, thereby generating seven feature spaces. When the threshold was 0, the space included all features proposed in this work. For each feature space, 10 RF classifiers, 10 XGB classifiers and 10 LGBM classifiers were separately developed based on the 10 training sets. Thus, a total of 210 (3*10*7) classifiers were constructed for each dataset. Finally, for each nucleotide, the average value of the outputs from these classifiers was considered the probability score of being a binding nucleotide (i.e. the probability estimated by our ensemble classifier).

Template-based prediction module

Considering that NAs sharing similar structures could adopt conserved regions to interact with other molecules, we designed a template-based model to predict binding nucleotides (Figure 1B). To this end, we compared the query RNA/DNA with structures from the template library (i.e. training sets) using the structural alignment algorithm RNA-align [57]. The best template together with its binding partner was selected in terms of TM-score. We superimposed the query structure onto the template structure using the rotation matrix generated by RNA-align. Based on the predicted complex structure, one nucleotide in the query was predicted to be a binding nucleotide with a score of 1 if this nucleotide was within 4.5 Å of the interacting partner (i.e. the distance constraint for defining real binding nucleotides in the native structures), otherwise with a score of 0.

Integrative prediction module

To utilize the complementarity between the feature- and template-based modules, we used a piecewise function to integrate their results (Figure 1C). If the query found a high-quality template, the weighted output of these two classifiers was considered the prediction score. Otherwise, the output of the feature-based classifier was adopted. The scoring function is as follows:

$$Iscore = \begin{cases} \alpha Fscore + (1-\alpha)\, Tscore & \text{if } TM - score > \text{cutoff} \\ Fscore & \text{otherwise} \end{cases}$$

where *Fscore* and *Tscore* are the probability scores from the feature- and template-based classifiers, respectively. The parameters α and cutoff were 0.70 and 0.35 for RNA and 0.70 and 0.40 for DNA, respectively. These parameters were selected based on Supplementary Figure S7 available online at http://bib.oxfordjournals.org/.

Post-processing module

Here, we adopted the RWR algorithm to correct the integrative prediction scores (Figure 1D) [58]. As mentioned above, the NA structure can be denoted by a nucleotide network. Starting from a node in the network, the walker has two choices at each step: either moving to a direct neighbor with a probability of $1-\beta$ or going back to the source node with a probability of β. The parameter β is the restart probability and was set to 0.4 in this work. A weight matrix *W* is needed to determine the probabilities of moving to different neighbors and was defined as follows:

$$w_{ij} = \begin{cases} \dfrac{\delta - \text{distance}_{ij}}{\delta} & \textit{if } \text{distance}_{ij} < \delta \\ 0 & \text{otherwise} \end{cases}$$

where w_{ij} denotes the weight between nucleotides *i* and *j*, δ is the distance cutoff for generating nucleotide networks (i.e. 8 Å) and distance$_{ij}$ is the distance between nucleotides *i* and *j*. The column-wise normalization was conducted on this matrix. Let p_0 be the initial probability vector (i.e. the integrative scores for all nucleotides) and p_t be the probability vector at step *t*. The RWR algorithm can be formalized as follows:

$$p_{t+1} = (1-\beta)\, p_t W + \beta p_0$$

In the following process, p_t could be updated gradually until $|p_{t+1} - p_t| < T$, where *T* is a predefined threshold and was set to 10e−6. β and *T* were selected based on Supplementary Figure S8 available online at http://bib.oxfordjournals.org/.

Performance evaluation

To determine the prediction framework, we performed 5-fold cross-validation (CV) on the training sets (i.e. 70% of the whole datasets) of RNA and DNA. For each training set, the RNA/DNA structures were randomly divided into five subgroups with a roughly equal number of structures. Four subgroups were used as the training set of our feature-based model and the template library of our template-based model, while one subgroup was used as the testing set. In addition, the independent test sets (i.e. 30% of the whole datasets) were used to further assess the generalization ability of our algorithm. Notably, when performing 5-fold CV, the knowledge-based preference features were derived from the four subgroups. For independent testing, these features were dependent on the training samples. The area under the curve (AUC) and Matthews correlation coefficient (MCC) were used as the primary measures. Meanwhile, other metrics such as recall, precision and accuracy (ACC) were also computed. These measures were calculated for each NA chain and the average results were reported for each dataset. We also evaluated whether the difference in performance between a pair of prediction models was significant or not. For a given dataset, we randomly selected 50% of NA chains to calculate the average AUC values. This procedure was repeated 10 times and the paired results were applied to statistical tests. The paired *t*-test and Wilcoxon signed-rank test were utilized to evaluate *P*-values.

Results

Analyses of preference features of binding sites in RNA

Based on all the structures collected in this work, the binding preferences of RNA were analyzed in terms of nucleotide types, functional groups and secondary structures, respectively. We found that the RL dataset possessed the lowest proportion of binding nucleotides among the three RNA-related datasets owing to the small sizes of binding partners (i.e. ligands), while the binding proportions were comparable for the RP and RN datasets (Figure 2A–C). For ligand-binding sites, the percentage of guanine (G) was relatively higher than the measures for the remaining nucleotide types, while a lower percentage of uracil (U) was revealed for protein- and NA-binding sites. According to the propensity scores, the RL dataset displayed the strongest binding preference for G, while U and C achieved the highest measures in the RP and RN datasets, respectively (Figure 2D–F). Regarding the functional groups, the binding nucleotides preferred to adopt nucleobases to contact ligands and NAs (Figure 2G–I). The similar tendency was observed in the study of RL complexes conducted by Kligun and Mandel-Gutfreund [59]. For the RP dataset, in contrast, the proportion of phosphate groups was remarkably elevated, probably due to the fact that the positively charged residues in proteins can interact with the negatively charged phosphate groups in NAs [60]. In Supplementary Figure S9 available online at http://bib.oxfordjournals.org/, from the physicochemical viewpoint [61], our analysis indicated that the four types of nucleotides tended to use their functional groups to make contacts with the sidechains of positively charged polar amino acids (e.g. R and K). Moreover, we found that RNAs may not only prefer to adopt major

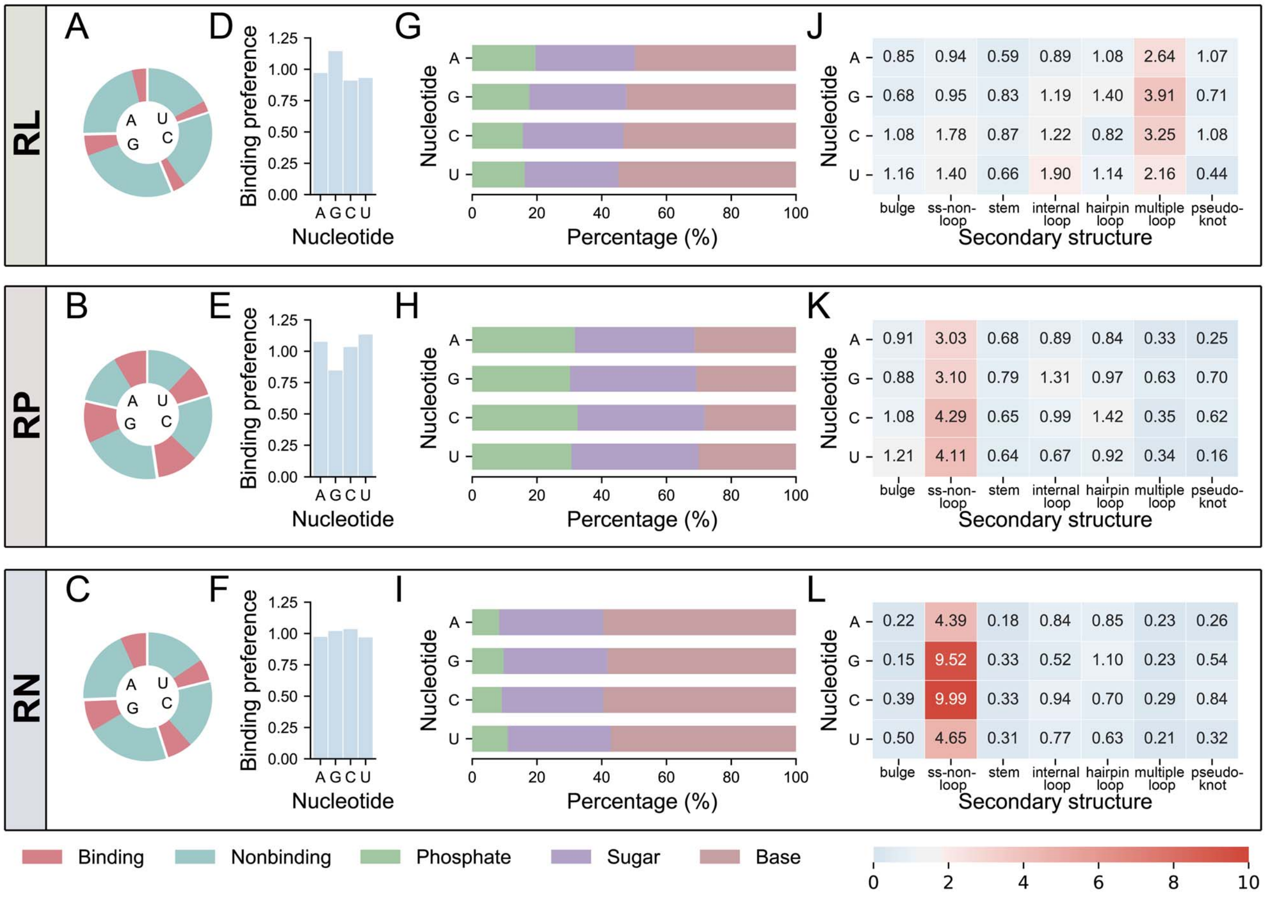

Figure 2. Analyses of preference features of binding sites in RNA. (**A–C**) Proportions of binding and nonbinding sites. (**D–F**) Binding preferences for nucleotide types. (**G–I**) Binding preferences for functional groups. (**J–L**) Binding preferences for secondary structure states. The ss-non-loop in (J-L) denotes the non-loop single-stranded segment.

grooves to interact with arginines but also frequently use non-loop single-stranded regions and stems to contact positively charged residues. As illustrated in Figure 2J–L, the most favorable secondary structures of ligand-binding regions were multiple loops followed by other unpaired elements. Existing studies have reported that ligand-binding sites were enriched in loops having noncanonical pairs [62]. For the other two datasets, RNA preferentially used single-stranded segments to contact proteins and NAs. This preference was especially obvious for the latter, probably because the RN dataset contained a group of RNAs without a stem-loop structure (e.g. RNAs involved in the RNA–DNA helices).

Analyses of structural features of binding sites in RNA

We then systematically compared the binding and nonbinding sites in RNA-related datasets based on structural properties. As shown in Figure 3A–C, the positive samples in the RL dataset had lower RSA and CX measures compared to the negative samples. Conversely, the positive samples in the RP and RN datasets possessed higher RSA and CX values. These indicated that RNA may use hydrophobic and concave locations to contact small ligands but could utilize exposed and locally convex regions to interact with proteins and NAs. For the LN features of nucleotides from local to global scales, RNA binding sites for ligands achieved lower values than nonbinding sites, while those for proteins obtained higher values (Figure 3D–H), further suggesting their preferences for relatively concave and convex regions in RNA structures, respectively. In Figure 3I–L, ligand-binding sites achieved greater values for the network centrality measures (i.e. degree, closeness and betweenness), which was in agreement with the observation from Su *et al.*'s work [23]. The opposite tendency was revealed for the degree and closeness measures of protein- and NA-binding sites. These suggested that the binding regions associated with ligands (e.g. binding pockets) were more likely to be close to the center of RNA structures, whereas the protruding surfaces that were far from the center could be recognized by proteins and NAs. Notably, the betweenness of positive samples in the RN dataset was higher than that of negative samples, probably because the binding nucleotides in the RNA–DNA helices played an indispensable role as a bridge in the corresponding nucleotide networks. Furthermore, we examined the above features for different types of RNAs. For the local structural features (i.e. RSA, DPX and CX), the preferences of various RNA groups were generally consistent with those of the whole dataset (Supplementary Figures S10–S12 available online at http://bib.oxfordjournals.org/). Nevertheless, some groups showed specific preferences for Laplacian features and network features. For instance, the opposite trend was observed for messenger RNAs and viral RNAs in the RP dataset when the results derived from all structures were adopted as the reference. Collectively, RNA may interact with ligands through

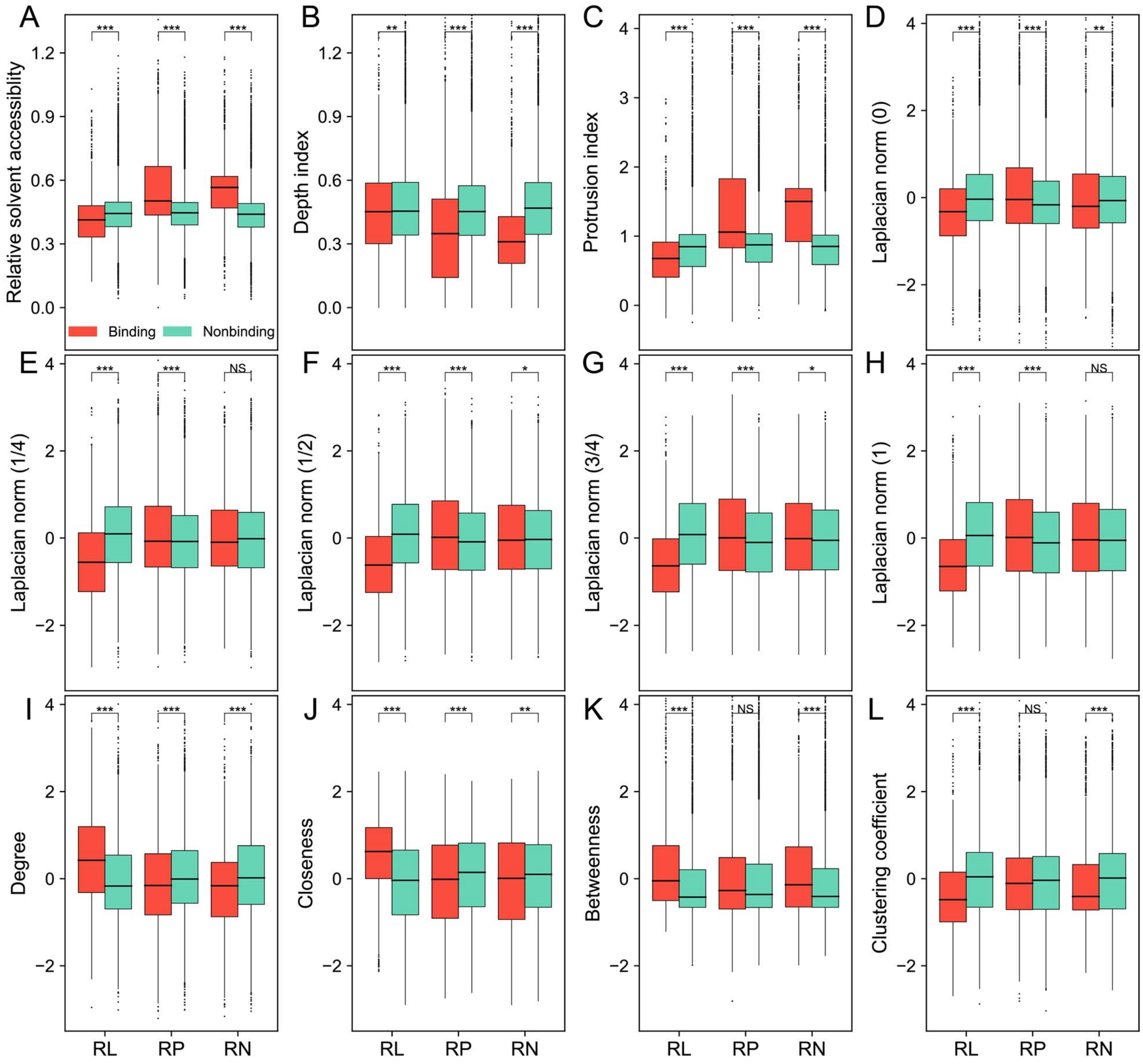

Figure 3. Analyses of structural features of binding sites in RNA. (**A–C**) Local structural descriptors. (**D–H**) Laplacian descriptors. (**I–L**) Network descriptors. The statistical significance is evaluated using the Wilcoxon rank sum test. $^{***}P < 0.001$, ** $0.001 \leq P < 0.01$, * $0.01 \leq P < 0.05$ and ns: $P \geq 0.05$.

forming binding pockets and contact proteins and NAs using convex surfaces. Our previous works and other studies reported the similar binding locations in proteins [33, 35, 63]. Accordingly, RNAs and proteins might adopt a similar way to interact with their binding partners. Compared to RNAs, however, proteins could be more likely to select the locations that are closer to the center of molecular structures or topological networks when they are involved in interactions with NAs [33, 35].

Performance evaluation on RNA-related training sets

Based on the above features, we constructed a group of classifiers using the three machine learning methods (i.e. RF, XGB and LGBM) combined with random undersampling and evaluated their performance by conducting 5-fold CV on training sets. As shown in Table 1, when we used all features and performed the resampling process 10 times, the classifiers achieved AUCs of approximately 0.68, 0.66 and 0.74 for ligand-, protein- and NA-binding sites, respectively. Notably, none of these machine learning algorithms consistently showed the best performance on the three datasets. The support vector machine was also tested in this work, but the performance was generally worse than that of bagging- and boosting-based algorithms (Supplementary Table S2 available online at http://bib.oxfordjournals.org/). The correlations among different features are presented in Supplementary Figure S13 available online at http://bib.oxfordjournals.org/. Furthermore, we examined the prediction ability of different groups of features in a similar way. From Supplementary Table S3 available online at http://bib.oxfordjournals.org/, we observed that the structural features performed significantly better than the sequence and preference properties. When the spatial neighboring features were added into the structure- and preference-based models, the performance was improved, especially for the latter. The results of structure-based classifiers were even more favorable than the performance of the classifiers based on all features, demonstrating that redundant

Table 1. Performance of different classifiers on RNA-related training sets

Classifier	RL-TR		RP-TR		RN-TR	
	MCC	AUC	MCC	AUC	MCC	AUC
NABS-RF	0.189	0.686	0.213	0.669	0.277	0.757
NABS-XGB	0.180	0.677	0.201	0.658	0.272	0.737
NABS-LGBM	0.196	0.682	0.222	0.673	0.261	0.740
NABS-FEA	0.257	0.701	0.228	0.676	0.270	0.757
NABS-TEM	0.221	NA	0.179	NA	0.193	NA
NABS-INT	0.294	0.727	0.234	0.698	0.314	0.753
NABS	0.300	0.736	0.235	0.702	0.318	0.760

NABS-RF, NABS-XGB and NABS-LGBM denote the RF-, XGB- and LGBM-based classifiers based on all features and 10 sample spaces. NABS-FEA denotes the feature-based ensemble learning classifier, and NABS-TEM denotes the template-based classifier. NABS-INT denotes the integrative classifier by combining NABS-FEA and NABS-TEM, and NABS denotes the final classifier by incorporating a post-processing procedure into NABS-INT. The annotations for different classifiers in this table are identical to those in the following tables.

information affected the performance. Herein, we conducted the feature selection procedure as suggested in the Methods section. As expected, the structure-related features were most frequently observed in the 30 optimal feature subsets (Supplementary Figure S14 available online at http://bib.oxfordjournals.org/). By the number of occurrences, we built seven feature spaces and assessed their performance (Supplementary Table S4 available online at http://bib.oxfordjournals.org/). Compared to the original feature space (i.e. all features), the other six subsets generated better results. Finally, we established 210 classifiers by exploiting the diversity of machine learning algorithms, feature spaces and sample spaces and adopted the average of their outputs as the prediction score. The correlation analysis for the outputs of the 210 classifiers is shown in Supplementary Figure S15 available online at http://bib.oxfordjournals.org/. Our ensemble classifiers yielded AUCs of 0.701, 0.676 and 0.757 for the training sets, which were comparable to the optimal performance shown in Supplementary Table S4 available online at http://bib.oxfordjournals.org/. Furthermore, the ensemble strategy could enhance the robustness of our machine learning module.

Aside from the machine leaning-based module, we also applied the template-based module to the training sets by using RNA-align as the structural alignment engine. For each query, the best template was reserved. As shown in Supplementary Figure S16 available online at http://bib.oxfordjournals.org/, most structures in the three RNA-related datasets can obtain a template with a TM-score greater than 0.25. Especially, the RNA chains binding to proteins were more likely to find structures with higher similarities (e.g. TM-score >0.55) compared with those interacting with ligands and NAs, maybe because the current RP dataset was more complete than the other two datasets. Moreover, the sequence identities between the template and query structures were typically in the range of 10–50%. Based on the predicted complex structure, we generated the putative binding nucleotides in each query RNA using the distance constraint. As shown in Table 1, the template based-module achieved MCCs of 0.221, 0.179 and 0.193 for the RL-TR, RP-TR and RN-TR dataset, respectively. The performance was not as good as the results from the feature-based module. In terms of the quality of templates, moreover, the template method can achieve more favorable results for the structures having higher TM-scores (Supplementary Table S5 available online at http://bib.oxfordjournals.org/).

After obtaining the predictions from the feature- and template-based modules, we used an integrative way to elevate accuracy based on their complementary relationship. For each dataset, we separated the RNA structures into two groups as shown in Supplementary Table S6 available online at http://bib.oxfordjournals.org/. When the query structures retrieved a template with a TM-score greater than 0.35, the feature- and template-based classifiers can achieve comparable measures. Moreover, if the integrative strategy was applied to these subgroups, the AUCs for the RL and RP datasets and the MCC for the RN dataset were clearly improved. If the queries cannot find a good template (e.g. TM-score <0.35), the prediction ability of template methods degenerated significantly, especially for ligand- and protein-binding nucleotides, while the machine learning classifiers still provided effective predictions for these structures. Because we used the interplay of individual modules for RNA structures having reliable templates and disregarded the predictions of template-based models for RNA structures without good templates, the measures for the whole datasets were reasonably improved (Table 1).

Finally, we applied the RWR algorithm to further correct the integrative prediction scores. As shown in Table 1, through the post-processing procedure, the AUCs for the three training sets were increased from 0.727, 0.698 and 0.753 to 0.736, 0.702 and 0.760, respectively. The final MCC values were generally lower, probably because of the class imbalance in the datasets (Supplementary Table S1 available online at http://bib.oxfordjournals.org/). Supplementary Figure S17A–C available online at http://bib.oxfordjournals.org/ illustrates that the changes in prediction scores were not so large but were enough to have a certain influence on the overall performance. The corrected predictions slightly increased the number of false negatives but remarkably decreased the number of false positives (Supplementary Figure S17D available online at http://bib.oxfordjournals.org/). This suggested that the isolated binding predictions were effectively removed by RWR, therefore causing a moderate improvement in performance. Based on the final predictions, we further exhibited that NABS might have preferences for binding nucleotides in different secondary structure states (Supplementary Table S7 available online at http://bib.oxfordjournals.org/) and for different categories of RNAs (Supplementary Figure S18 available online at http://bib.oxfordjournals.org/), suggesting that the difficult samples (e.g. the bulge group in the RN dataset and the ribosomal RNAs in the RL dataset) should be given more concern in the future. Moreover, we investigated the overlapping binding sites among the three training sets. A certain number of RNAs could use the same binding regions to contact proteins and NAs, and our classifiers can identify the majority of overlapping nucleotides (Supplementary Figure S19 available online at http://bib.oxfordjournals.org/). In addition to 5-fold CV, the leave-one-chain-out validation and 10-fold CV were also utilized to

Table 2. Performance of different classifiers on RNA-related testing sets

Classifier	RL-TS		RP-TS		RN-TS	
	MCC	AUC	MCC	AUC	MCC	AUC
NABS-RF	0.227	0.717	0.227	0.690	0.263	0.754
NABS-XGB	0.240	0.693	0.232	0.690	0.279	0.747
NABS-LGBM	0.225	0.680	0.240	0.698	0.302	0.745
NABS-FEA	0.264	0.703	0.236	0.696	0.319	0.756
NABS-TEM	0.194	NA	0.249	NA	0.231	NA
NABS-INT	0.296	0.713	0.307	0.734	0.311	0.749
NABS	0.280	0.724	0.304	0.748	0.331	0.752

evaluate our classifiers (Supplementary Table S8 available online at http://bib.oxfordjournals.org/). The measures were relatively better than those of 5-fold CV, probably because more structures were included in the training set and template library.

Performance evaluation on RNA-related testing sets

We further evaluated different classifiers using the testing sets (Table 2). Compared with the results for 5-fold CV, the performance of the feature-based classifiers based on various machine learning methods and ensemble learning was comparable or better. In contrast, the template method exhibited more remarkable fluctuations in performance. Through combining the individual modules, the superior performance was obtained for the RL-TS and RP-TS datasets. By the post-correction process, the AUCs for the three datasets were 0.724, 0.748 and 0.752, respectively. The *P*-values of statistical significance tests are shown in Supplementary Table S9 available online at http://bib.oxfordjournals.org/. Moreover, the independent testing procedure was repeated 100 times by selecting different testing structures. The average performance and standard deviations suggested that our models may avoid the problem of overfitting (Supplementary Figure S20 available online at http://bib.oxfordjournals.org/). According to Supplementary Figure S1 available online at http://bib.oxfordjournals.org/, in addition, we used the structures with a resolution better than 3 Å as training sets and the remaining structures as testing sets. The results were comparable to those derived from 5-fold CV, suggesting that our method could effectively identify the binding nucleotides in relatively low-resolution structures (Supplementary Table S10 available online at http://bib.oxfordjournals.org/). We also performed cross-site-type predictions on the three testing sets. As shown in Supplementary Table S11 available online at http://bib.oxfordjournals.org/, only the performance across the protein- and NA-binding sets was better than expected by chance, indicating that the specific classifiers tended to produce the different predicted sites for RNAs.

If the proposed method was only suitable for experimentally determined structures, the application scope of NABS would be severely restricted. Accordingly, we generated the predicted structure for each RNA chain in the testing sets. As shown in Table 3, the performance of all classifiers decreased significantly compared with the results in Table 2. NABS achieved AUCs of approximately 0.61 for these datasets. This may be due to the discrepancies between the native and predicted structures, which can be measured by the root-mean-square deviation (RMSD). Meanwhile, we calculated the difference in the structural features between the native and predicted structures. The RMSD was highly correlated with the difference resulting from structural features (Supplementary Figure S21A–C available online at http://bib.oxfordjournals.org/), implying that the structural deviation resulted in the alterations in features and thus induced negative impacts on the feature-based prediction. Moreover, we compared the template quality between the paired structures. A smaller number of predicted structures achieved a reliable template (Supplementary Figure S21D–I available online at http://bib.oxfordjournals.org/), therefore leading to the difficulty in the template-based prediction. Additionally, we set a series of thresholds in terms of RMSD. Supplementary Table S12 available online at http://bib.oxfordjournals.org/ shows that the performance of NABS was gradually improved by decreasing the cutoff value, suggesting that our algorithm could be applied to the predicted RNA structures with high confidence (e.g. RMSD <15 Å).

Comparison with other RNA binding site prediction methods

In this section, we compared NABS with other algorithms, including Rsite, Rsite2, RBind and RNAsite [20–23]. Rsite and Rsite2 predicted functional sites in RNA based on the tertiary structure- and secondary structure-derived distance features, respectively. RBind adopted the degree and closeness measures to detect ligand- and protein-binding nucleotides, while RNAsite (published very recently) combined the above two network features with several structural and sequence attributes to identify ligand-binding sites. Besides the datasets prepared in this work, the native and predicted structures collected by RBind were also used for assessment. For the first three competing algorithms (i.e. Rsite, Rsite2 and RBind), we adopted their standalone programs to generate prediction results. Figure 4A–I displays the head-to-head comparison between NABS and these methods in terms of the MCC values of RNA chains from different datasets. The vast majority of the dots were scattered in the lower triangles, suggesting that our method obtained better results for most structures. Figure 4M–O reveals the average MCC value of each dataset using different methods. For both the native and predicted structures, NABS obviously outperformed the three methods, among which only RBind showed certain prediction ability for ligand-binding sites. Moreover, both NABS and RBind performed more favorably on the ligand-binding datasets of RBind, probably because most of these RNA structures have the canonical binding pockets to accommodate small molecules and the binding sites could thus be easily identified [23]. Although Wang *et al.* suggested that RBind could achieve promising precision measures [22], their method missed many real binding sites and therefore obtained lower recall and MCC measures (Supplementary Table S13 available online at http://bib.oxfordjournals.org/). To compare NABS with RNAsite, we implemented this algorithm based on the above datasets.

Table 3. Performance of different classifiers on predicted RNA structures

Classifier	RL-PS		RP-PS		RN-PS	
	MCC	AUC	MCC	AUC	MCC	AUC
NABS-RF	0.142	0.607	0.062	0.565	0.098	0.604
NABS-XGB	0.143	0.621	0.092	0.573	0.094	0.588
NABS-LGBM	0.113	0.603	0.062	0.589	0.092	0.602
NABS-FEA	0.124	0.617	0.074	0.577	0.074	0.591
NABS-TEM	0.033	NA	0.061	NA	0.043	NA
NABS-INT	0.124	0.616	0.105	0.595	0.089	0.611
NABS	0.156	0.617	0.134	0.601	0.098	0.607

As shown in Figure 4J–O, RNAsite could be adopted to predict protein- and NA-binding sites as well as ligand-binding sites in RNA. Although several identical features were shared by NABS and RNAsite, our algorithm obtained generally superior performance on various datasets. In addition, we compared these two methods using the three datasets of RNAsite. Note that because the structural similarities were reduced in these datasets, our template method was not used in this scenario. Even so, NABS still achieved higher AUCs and comparable MCCs (Supplementary Table S14 available online at http://bib.oxfordjournals.org/). *P*-values of statistical significance tests are shown in Supplementary Tables S15 and S16 available online at http://bib.oxfordjournals.org/. Finally, we compared NABS with a sequence-based predictor called RBPbinding, which was trained by CLIP data to predict protein-binding regions [64]. NABS showed more favorable performance than RBPbinding on the RP-TS dataset (Supplementary Table S17 available online at http://bib.oxfordjournals.org/). The advantages of our algorithm could be due to the following reasons: (a) we extracted multifaceted structural features (including local and global descriptors) in conjunction with sequence and propensity features to characterize nucleotides; (b) we built a robust feature-based module based on ensemble learning and then combined this module with a template-based module to elevate performance and (c) we designed an effective post-processing procedure to correct the prediction results (especially false positives) through the RWR algorithm.

Analyses of preference and structural features of binding sites in DNA

To compare DNA and RNA, we performed the same analyses for the binding sites in DNA. As shown in Figure 5A–C, the binding proportion was lowest in the DL dataset, and there were comparable proportions in the other two datasets, which was consistent with the results from RNA-related datasets. However, the percentages of protein- and NA-binding nucleotides were greater than the percentage of nonbinding nucleotides, which was contrary to the observations for RNA. In Figure 5D–F, the ligand- and NA-binding nucleotides showed a strong preference for G and T, respectively, whereas the difference between the preferred nucleotide types (i.e. G and T) was very small in the DP dataset. According to Figure 5G–I, the functional group preferences of DNA were similar to those of RNA, but a significantly higher percentage of nucleobases together with an extremely low proportion of phosphate groups was observed for NA-binding sites. This implied that the contacts between DNA and other NA chains were dominated by base-pairing interactions. Regarding the DP dataset, the increase in the percentage of phosphate groups might also be due to their electrostatic interactions with positively charged amino acids (Supplementary Figure S9 available online at http://bib.oxfordjournals.org/).

We also compared the structural properties between binding and nonbinding sites in the DL, DP and DN datasets. As shown in Figure 6A–C, compared to RNA-related datasets, these three datasets did not demonstrate very significant differences in solvent accessibilities and geometric attributes, probably because DNA exists mainly as base-paired helices and most nucleotides thus possess similar local features. From Figure 6D–H, we can find that the three datasets shared an identical pattern. Namely, the LN measures of positive samples were generally smaller than those of negative samples. This phenomenon was also observed in the RL dataset. Unlike RNA, DNA may rarely have binding pockets due to its helical conformation. We therefore proposed that the positions in DNA bound by other molecules (i.e. ligands, proteins and NAs) would be adjacent to the middle of DNA chains. As illustrated in Figure 6I–L, the closeness and betweenness values were higher for the three categories of binding sites. The results further supported that the middle regions of DNA played important roles in interacting with different partners. It should be noted that DNA chains solved in PDB structures were generally incomplete, so we performed an additional analysis using the DP dataset. In Supplementary Figure S22 available online at http://bib.oxfordjournals.org/, the differences in structural features were consistent for the groups with different chain lengths, implying that the longer DNAs could exhibit similar properties. Compared with RNAs that provide binding pockets for ligands and protruding surfaces for proteins and NAs, we suggested that DNAs adopt structurally different ways to interact with other molecules.

Performance evaluation on DNA-related datasets

Furthermore, we applied the proposed algorithm to identify binding nucleotides in DNA. As shown in Table 4 and Supplementary Table S6 available online at http://bib.oxfordjournals.org/, the utilities of different modules were generally observed by performing 5-fold CV on training sets, and NABS achieved AUCs of 0.661, 0.764 and 0.811 for ligand-, protein- and NA-binding sites, respectively. DNA-related datasets showed more remarkable discrepancies in prediction results than RNA-related datasets (AUCs: 0.70–0.76). The relatively worse measure for the DL-TR dataset could mainly result from the feature-based predictions. First, compared with the other types of binding sites in RNA and DNA (Figures 3 and 6), the ligand-binding regions in DNA did not show preferences for some structural properties, especially the local descriptors (e.g. RSA, CX, DPX and degree), when using nonbinding nucleotides as the reference. Second, the sequence features could play important roles in increasing the performance discrepancies for DNA. In

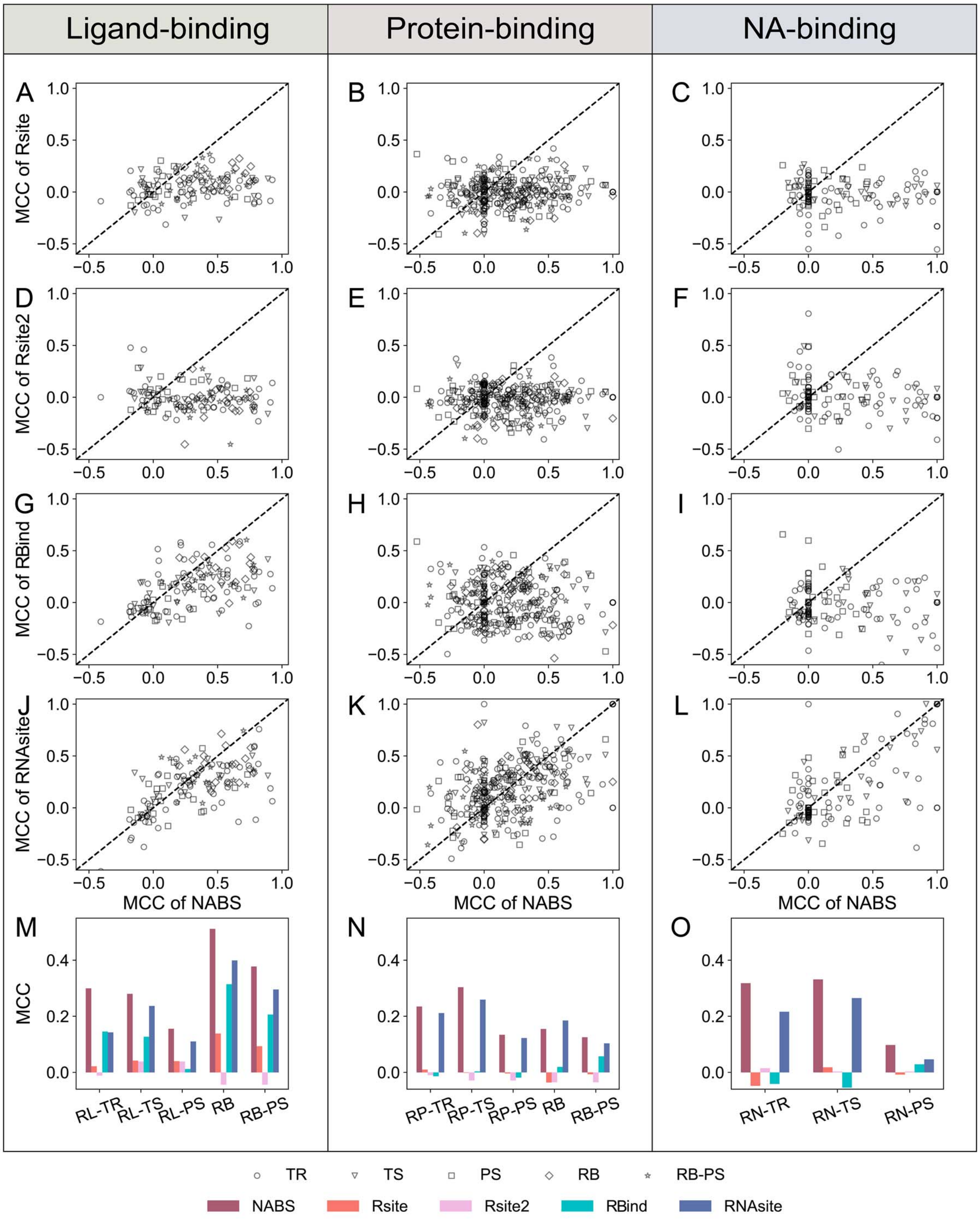

Figure 4. Comparison between NABS and other methods based on MCC measures. (**A–C**) Rsite versus NABS. (**D–F**) Rsite2 versus NABS. (**G–I**) RBind versus NABS. (**J–L**) RNAsite versus NABS. (**M–O**) Average MCC measures of all structures in various datasets based on different methods. In (**A–L**), each dot represents an RNA structure. TR, TS and PS denote the training, testing and predicted structures prepared in this work. RB and RB-PS denote the native and predicted structures prepared by RBind.

Supplementary Figure S6 available online at http://bib.oxfordjournals.org/, based on the optimal sequential neighborhood, the AUC for DL-TR was 0.550, while the measures for DP-TR and DN-TR were 0.668 and 0.695, respectively, which were even greater than the AUC of NABS for DL-TR. Supplementary Figure S14 available online at http://bib.oxfordjournals.org/ also displays that sequence descriptors were generally preferred by protein- and NA-binding sites in the feature selection process. In contrast, the contributions of sequence features were not so significant for RNA-related datasets (AUCs: 0.52–0.60). On the other hand, we found that the template method achieved surprising performance with an MCC of 0.702 for NA-binding sites, which contributed greatly to the final result of NABS for DN-TR. This implied that a number of DNA chains had highly similar

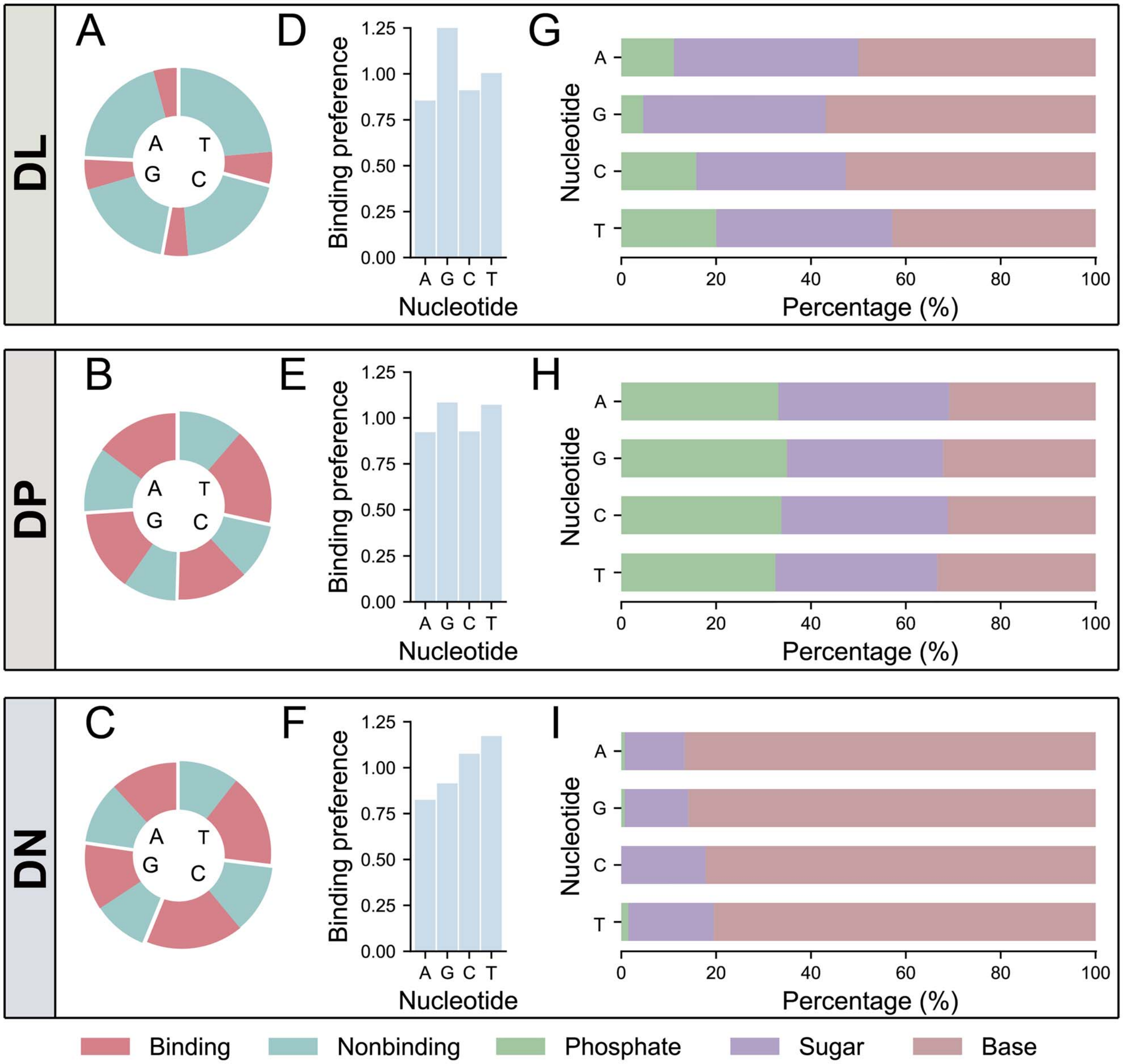

Figure 5. Analyses of preference features of binding sites in DNA. (**A–C**) Proportions of binding and nonbinding sites. (**D–F**) Binding preferences for nucleotide types. (**G–I**) Binding preferences for functional groups.

structures in this dataset. Herein, we extracted the header information of PDB files and categorized DNA chains into different groups based on their functions. As revealed in Supplementary Figure S23 available online at http://bib.oxfordjournals.org/, by the number of DNA chains, the top two groups were associated with hydrolysis and transcription, respectively. We observed that the chains in each group were indeed similar by checking their structures with PyMOL. Supplementary Figure S23 available online at http://bib.oxfordjournals.org/ shows that the results of chains in specific functional groups were generally superior to those of functionally isolated chains. Moreover, we investigated the sharing binding sites among the training sets. We found that 13 (12) DNAs adopted the same regions to interact with proteins and NAs (ligands). The majority of overlapping nucleotides can also be identified by our predictors (Supplementary Figure S19 available online at http://bib.oxfordjournals.org/).

Besides, the three DNA-related testing sets were used to evaluate the performance of NABS. As shown in Table 5 and Supplementary Figure S20 available online at http://bib.oxfordjournals.org/, compared to the results for training sets, the performance of feature- and template-based classifiers was moderately reduced. Using the hybrid strategy and the post-processing procedure, the final model yielded AUCs of 0.640, 0.753 and 0.822 for the DL-TS, DP-TS and DN-TS datasets, respectively. *P*-values of statistical significance tests are shown in Supplementary Table S9 available online at http://bib.oxfordjournals.org/. We also performed cross-site-type predictions on the testing sets. The cross predictions were largely effective (Supplementary Table S11 available online at http://bib.oxfordjournals.org/), probably because DNAs could use a similar way to contact the three types of interacting partners. Accordingly, the different classifiers may provide the same predicted sites

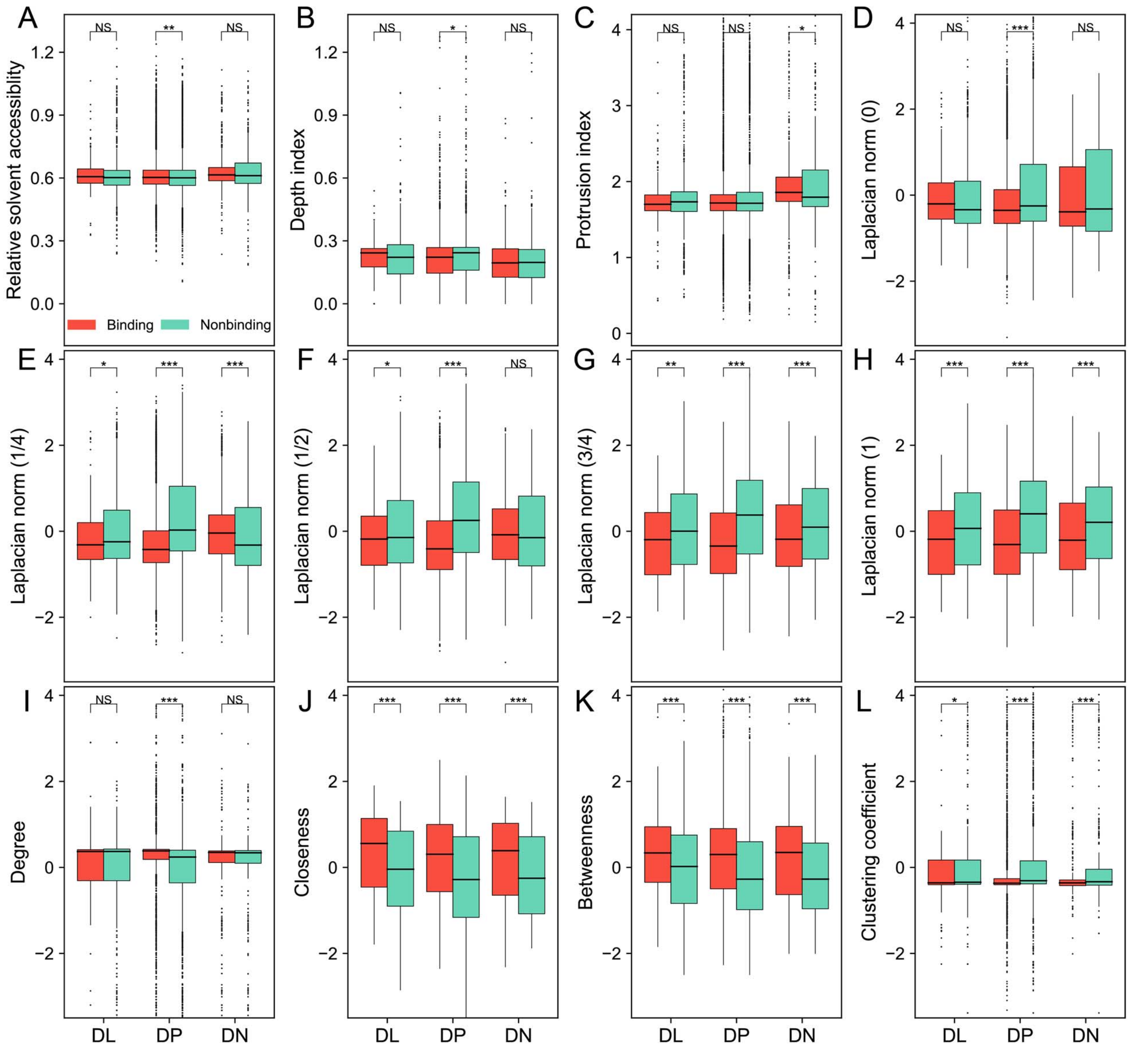

Figure 6. Analyses of structural features of binding sites in DNA. (**A–C**) Local structural descriptors. (**D–H**) Laplacian descriptors. (**I–L**) Network descriptors. The statistical significance is evaluated using the Wilcoxon rank sum test. ***$P < 0.001$, ** $0.001 \le P < 0.01$, * $0.01 \le P < 0.05$ and ns: $P \ge 0.05$.

Table 4. Performance of different classifiers on DNA-related training sets

Classifier	DL-TR		DP-TR		DN-TR	
	MCC	AUC	MCC	AUC	MCC	AUC
NABS-RF	0.131	0.602	0.342	0.744	0.494	0.798
NABS-XGB	0.100	0.609	0.337	0.744	0.421	0.760
NABS-LGBM	0.063	0.582	0.332	0.746	0.467	0.785
NABS-FEA	0.159	0.652	0.335	0.750	0.478	0.786
NABS-TEM	0.122	NA	0.263	NA	0.702	NA
NABS-INT	0.163	0.655	0.381	0.761	0.519	0.815
NABS	0.208	0.661	0.380	0.764	0.549	0.811

for DNAs. Finally, we compare our algorithm with two genomic sequence-based methods, namely DeepSNR and D-AEDNet, which can identify transcription factor binding sites at the base-pair level using deep learning techniques [65, 66]. Among the eight proteins used in these two studies, four proteins were involved in protein–DNA complexes in which 33 DNA chains could be used for comparison (Supplementary Table S17 available online at http://bib.oxfordjournals.org/). Our method

Table 5. Performance of different classifiers on DNA-related testing sets

Classifier	DL-TS		DP-TS		DN-TS	
	MCC	AUC	MCC	AUC	MCC	AUC
NABS-RF	0.147	0.633	0.305	0.725	0.488	0.772
NABS-XGB	0.106	0.620	0.301	0.712	0.493	0.785
NABS-LGBM	0.123	0.619	0.297	0.716	0.498	0.813
NABS-FEA	0.128	0.622	0.313	0.724	0.500	0.806
NABS-TEM	0.046	NA	0.211	NA	0.522	NA
NABS-INT	0.158	0.633	0.341	0.747	0.552	0.829
NABS	0.200	0.640	0.336	0.753	0.571	0.822

generally outperformed both DeepSNR and D-AEDNet, indicating the superiority of NABS over the existing sequence-based methods. Altogether, our methodology could be applicable to the binding sites in DNA, and the advantages of various modules were also helpful in improving performance.

Case studies

Since the above analyses and predictions were performed at the macroscopic level, we chose several representative structures to further validate the reliability of our results. Figure 7A displays that the 2′-deoxyguanosine riboswitch adopts a binding pocket to accommodate the ligand (PDB ID: 3SKI) [67]. The feature-based classifier accurately predicted most binding nucleotides aside from two positions but generated many false positives (Figure 7B). In contrast, the template-based model did not achieve any false positives but missed four binding nucleotides. Through the model integration and post-correction, NABS successfully predicted 9 out of 11 binding nucleotides together with two false positives. Figure 7C shows that a small regulatory RNA RydC interacts with a bacterial Hfq protein using the protruding region of its 3′-end poly-U tail (PDB ID: 4V2S) [68]. This complex can bind to target messenger RNAs for sRNA-based regulation. The third example is the structure of the CRISPR RNA and target RNA duplex, in which nucleotides contact each other by base-pairing interactions (Figure 7E, PDB ID: 5XWP) [69]. As revealed in Figure 7D and F, similar to the classifier for the first example, the feature-based method yielded a greater number of false positives, while the template method obtained higher precision measures together with lower recall measures. The following two steps remarkably decreased the number of false positives but maintained the number of true positives. The AUCs of these two examples were 0.96 and 0.99, respectively.

We also selected three DNA structures in this section. The first DNA molecule that was extracted from the human topoisomerase I-DNA covalent complex can be targeted by anticancer drugs (Figure 7G, PDB ID: 1SEU) [70]. It is clear that this chain uses its middle part to contact the small molecule. As shown in Figure 7H, although this query obtained a template with a high TM-score, the predicted complex did not provide useful information for ligand-binding regions (there were no true positives). The integrative predictions thus depended on the outputs of our feature method. After performing the RWR algorithm, the false positives remain unchanged. This was probably because these positions were assigned to the very high binding scores by both the feature and template modules. The second example is the homing endonuclease I-CreI in complex with its specific DNA sequence (Figure 7I, PDB ID: 6FB5) [71]. The central region of the target DNA is essential for the indirect readout of this interaction. From the aspect of true positives, the feature- and template-based models favorably complemented each other, which led to the increase in the number of correctly predicted binding nucleotides when using the integrative module (Figure 7J). The post-processing procedure converted the remaining two false negatives into true positives but yielded an additional false negative. Figure 7K reveals that the DNA chain from an elongation complex binds to a 10-mer RNA sequence (PDB ID: 4BY1) [72]. As shown in Figure 7L, the false negative offered by the template method was eliminated through the integrative strategy, while one of the false positives resulting from the feature method was deleted by the post-correction step. As a result, NABS yielded an AUC of 0.99 for this structure. Furthermore, we visualized the other six structures whose AUCs were similar to the overall performance on the testing datasets (Supplementary Figure S24 available online at http://bib.oxfordjournals.org/). Collectively, these examples not only confirmed the reasonability of our results but also showed the merits and weaknesses of various modules in our algorithm.

Discussion

The existing works have performed comprehensive investigations and established a variety of prediction methods for binding sites in proteins, but the corresponding study on NAs is still in its infancy. To address this problem, we first constructed six datasets that comprised RNA or DNA structures in complex with different classes of binding partners (i.e. ligands, proteins or NAs). For each interaction type, we compared binding regions in NAs with nonbinding regions based on binding preferences and structural features. Protein- and NA-binding sites in both RNA and DNA were remarkably larger than ligand-binding sites. NA molecules tended to adopt nucleobases to interact with small ligands and other NAs and preferentially used a greater proportion of phosphate groups to contact proteins. Moreover, RNA was more likely to use unpaired nucleotides to contact other molecules from the secondary structure aspect. According to structural analyses, RNA could interact with ligands through forming binding pockets and contact proteins and NAs using protruding surfaces, while DNA may adopt regions closer to the middle of the chain to make contacts with other molecules. In conjunction with the existing knowledge about protein binding sites, we suggested that RNAs and proteins could use structurally similar strategies to interact with other molecules, but DNAs may choose a different way probably due to their low structural complexity. The results demonstrated the differences between NA binding sites and nonbinding sites and among the various types of NA binding regions.

Based on the above biological insights, we constructed a feature-based ensemble learning classifier that made full use

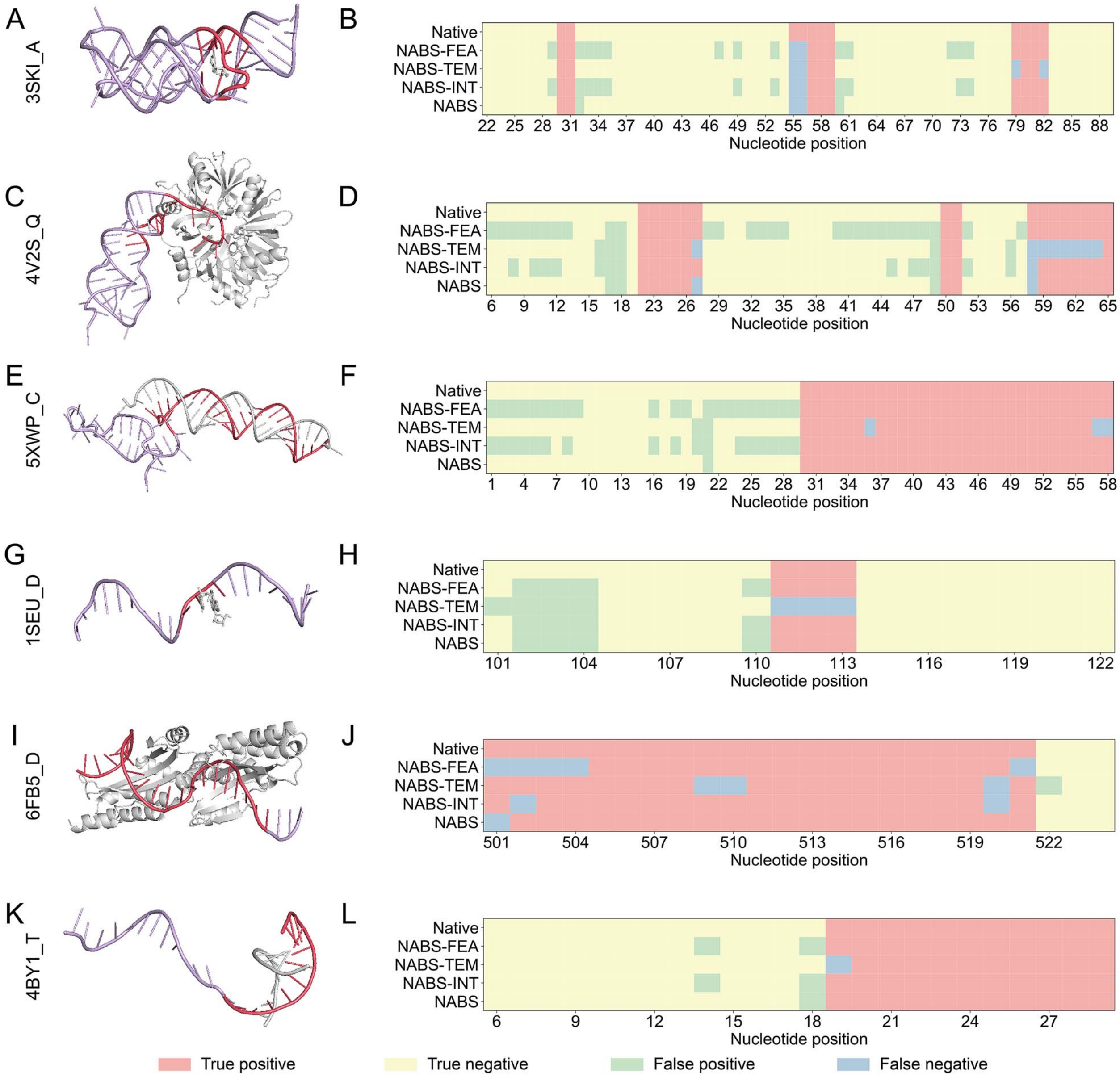

Figure 7. Tertiary structures and prediction results of representative RNAs and DNAs. (**A–F**) Three representative RNAs. (**G–L**) Three representative DNAs. For each example, the tertiary structure is shown in the left figure, and the prediction results are shown in the right figure. NABS-FEA denotes the feature-based ensemble learning classifier, and NABS-TEM denotes the template-based classifier. NABS-INT denotes the integrative classifier by combining NABS-FEA and NABS-TEM, and NABS denotes the final classifier by incorporating a post-processing procedure into NABS-INT.

of the interplay among different machine learning methods, feature spaces and sample spaces. This module was suitable for the structures in diversified datasets. The feature importance analysis demonstrated that the structural features were essential for the prediction of binding sites in RNA, while the sequence context, as well as structural information, played important roles in the identification of binding sites in DNA. As an alternative, a template-based classifier was also established by exploiting homology information. This classifier performed favorably on RNA and DNA chains having high-quality templates but lost its prediction ability when reliable templates were unavailable. For RNA-related datasets, at least half of the structures can obtain a good template, while less than one quarter of the structures from DNA-related datasets can find such a reference. For the whole datasets, the feature-based classifier showed more favorable performance than the template-based classifier. Utilizing the complementarity between these two classifiers, an integrative approach was established to elevate the accuracy. Further, we designed a post-correction process by performing the RWR algorithm on nucleotide interaction networks, which effectively deleted false positives derived from the integrative model. The promising results for various datasets suggested that our unified framework could be used to predict different types of binding nucleotides. In particular, NABS can be applied to predicted RNA structures as well as their native structures provided that the modeling results were acceptable. Finally, we implemented NABS as a user-friendly web server, which could yield putative binding nucleotides for different partners within several minutes (Supplementary Figure S25 available online at http://bib.oxfordjournals.org/).

Despite the progress achieved here, some problems are worth further studies in the future. First, the major binding strategies proposed for the two categories of NAs were summarized based on the statistical analyses of available structures. Inevitably, there are some exceptional cases and the recognition of binding sites in these structures would be more challenging. Therefore, more attention should be given to these cases. Second, although most features adopted in this work can be effectively used to identify binding sites in NAs, the quantitative representation of nucleotides could be further improved. For instance, for RNA or DNA chains in the helical structures, all nucleotides may be involved in base-pairing interactions and have highly similar local structural features (e.g. RSA, DPX and CX). We thus need to design finer local descriptors (e.g. the angle between adjacent nucleotides) to uncover the inherent difference between binding and nonbinding sites in this scenario. For the sequence context, furthermore, we only used the composition–transition–distribution features. In fact, we also checked the sequence-derived structural features provided by existing software (e.g. SPOT-RNA, RNAsnap, DNAshape and DynaSeq) [38, 73–75], but no remarkable improvement was observed (Supplementary Table S18 available online at http://bib.oxfordjournals.org/). Currently, some existing programs could readily yield diversified sequence descriptors [76, 77], which could be further selected to achieve better performance. Third, we used conventional machine learning algorithms to implement feature-based classifiers due to relatively small datasets. Recently, deep learning frameworks have been commonly applied to the recognition of protein functional residues [78, 79] and the prediction of structural properties for RNA [38, 74], so these methodologies could be used to predict binding nucleotides in NAs when more structures become available. Fourth, the current template-based model used RNA-align as the search engine and only depended on the results from the best template. We could test other RNA structure alignment tools (e.g. RMalign, STAR3D and Rclick) and integrate these programs to improve the selection of optimal reference structures [80–82]. Moreover, combining the results from multiple templates could be valuable to the increase in prediction accuracy. Fifth, the subgroups of nucleotides in RNA based on secondary structures and the subgroups of RNAs and DNAs based on functions had varying degrees of accuracy so that we could establish specific predictors for these subgroups to improve performance. Sixth, this study and previous works have predicted the binding sites in different types of biomacromolecules (e.g. DNA, RNA and proteins). In the future, novel algorithms could be developed to identify the physical contacts between biomacromolecules, such as residue–nucleotide contacts across protein–NA interfaces and nucleotide–nucleotide contacts across NA–NA interfaces, which could provide useful insights into the determination of complex structures. Altogether, this work not only offers an overview of specific characteristics of different binding sites in DNA and RNA but also provides an effective and efficient tool to predict these critical regions, which may help deepen our understanding of the mechanisms underlying the interactions between NAs and other molecules.

Key Points

- We characterize the binding sites of small ligands, proteins and nucleic acids in RNA and DNA from multifaceted viewpoints.
- We conduct a systematic comparison between binding and nonbinding sites in RNA and DNA and among different categories of binding sites.
- We develop an algorithm that combines feature- and template-based strategies to predict different categories of binding sites in RNA and DNA.

Funding

National Natural Science Foundation of China (32071249).

References

1. Gilbert W. Origin of life: the RNA world. *Nature* 1986;**319**:618.
2. Birney E, Stamatoyannopoulos JA, Dutta A, *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;**447**:799–816.
3. Morris KV, Mattick JS. The rise of regulatory RNA. *Nat Rev Genet* 2014;**15**:423–37.
4. Dervan PB. Molecular recognition of DNA by small molecules. *Bioorg Med Chem* 2001;**9**:2215–35.
5. Thomas JR, Hergenrother PJ. Targeting RNA with small molecules. *Chem Rev* 2008;**108**:1171–224.
6. Blount KF, Breaker RR. Riboswitches as antibacterial drug targets. *Nat Biotechnol* 2006;**24**:1558–64.
7. Philips A, Milanowska K, Lach G, *et al.* LigandRNA: computational predictor of RNA-ligand interactions. *RNA* 2013;**19**:1605–16.
8. Enright AJ, John B, Gaul U, *et al.* MicroRNA targets in Drosophila. *Genome Biol* 2003;**5**:R1.
9. Farh KK, Grimson A, Jan C, *et al.* The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* 2005;**310**:1817–21.
10. Wang X, Arai S, Song X, *et al.* Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* 2008;**454**:126–30.
11. Engreitz JM, Sirokman K, McDonel P, *et al.* RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell* 2014;**159**:188–99.
12. Licatalosi DD, Mele A, Fak JJ, *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 2008;**456**:464–9.
13. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;**10**:669–80.
14. Darnell RB. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA* 2010;**1**:266–86.
15. Eagen KP. Principles of chromosome architecture revealed by Hi-C. *Trends Biochem Sci* 2018;**43**:469–78.
16. Zhou T, Shen N, Yang L, *et al.* Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci U S A* 2015;**112**:4654–9.
17. Zhang S, Zhou J, Hu H, *et al.* A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res* 2016;**44**:e32.
18. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 2016;**44**:e107.
19. Alipanahi B, Delong A, Weirauch MT, *et al.* Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8.

20. Zeng P, Li J, Ma W, *et al*. Rsite: a computational method to identify the functional sites of noncoding RNAs. *Sci Rep* 2015;**5**:9179.
21. Zeng P, Cui Q. Rsite2: an efficient computational method to predict the functional sites of noncoding RNAs. *Sci Rep* 2016;**6**:19016.
22. Wang K, Jian Y, Wang H, *et al*. RBind: computational network method to predict RNA binding sites. *Bioinformatics* 2018;**34**:3131–6.
23. Su H, Peng Z, Yang J. Recognition of small molecule-RNA binding sites using RNA sequence and structure. *Bioinformatics* 2021;**37**:36–42.
24. He J, Wang J, Tao H, *et al*. HNADOCK: a nucleic acid docking server for modeling RNA/DNA-RNA/DNA 3D complex structures. *Nucleic Acids Res* 2019;**47**:W35–42.
25. Shazman S, Elber G, Mandel-Gutfreund Y. From face to interface recognition: a differential geometric approach to distinguish DNA from RNA binding surfaces. *Nucleic Acids Res* 2011;**39**:7390–9.
26. Bahadur RP, Zacharias M, Janin J. Dissecting protein-RNA recognition sites. *Nucleic Acids Res* 2008;**36**:2705–16.
27. Jones S, van Heyningen P, Berman HM, *et al*. Protein-DNA interactions: a structural analysis. *J Mol Biol* 1999;**287**:877–96.
28. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* 1996;**93**:13–20.
29. Naderi M, Lemoine JM, Govindaraj RG, *et al*. Binding site matching in rational drug design: algorithms and applications. *Brief Bioinform* 2019;**20**:2167–84.
30. Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform* 2016;**17**:88–105.
31. Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief Bioinform* 2019;**20**:1250–68.
32. Burley SK, Berman HM, Bhikadiya C, *et al*. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* 2019;**47**:D464–74.
33. Liu R, Hu J. DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches. *Proteins* 2013;**81**:1885–99.
34. Sun J, Wang J, Xiong D, *et al*. CRHunter: integrating multifaceted information to predict catalytic residues in enzymes. *Sci Rep* 2016;**6**:34044.
35. Yang XX, Deng ZL, Liu R. RBRDetector: improved prediction of binding residues on RNA-binding protein structures using complementary feature- and template-based strategies. *Proteins* 2014;**82**:2455–71.
36. Fan BL, Jiang Z, Sun J, *et al*. Systematic characterization and prediction of coenzyme A-associated proteins using sequence and network information. *Brief Bioinform* 2021;**22**:bbaa308.
37. Yang X, Wang J, Sun J, *et al*. SNBRFinder: a sequence-based hybrid algorithm for enhanced prediction of nucleic acid-binding residues. *PLoS One* 2015;**10**:e0133260.
38. Hanumanthappa AK, Singh J, Paliwal K, *et al*. Single-sequence and profile-based prediction of RNA solvent accessibility using dilated convolutional neural network. *Bioinformatics* 2020;**36**:5169–76.
39. Sun S, Wang W, Peng Z, *et al*. RNA inter-nucleotide 3D closeness prediction by deep residual neural networks. *Bioinformatics* 2021;**37**:1093–8.
40. Sun S, Wu Q, Peng Z, *et al*. Enhanced prediction of RNA solvent accessibility with long short-term memory neural networks and improved sequence profiles. *Bioinformatics* 2019;**35**:1686–91.
41. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.
42. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
43. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, *et al*. ViennaRNA package 2.0. *Algorithms Mol Biol* 2011;**6**:26.
44. Popenda M, Szachniuk M, Antczak M, *et al*. Automated 3D structure composition for large RNAs. *Nucleic Acids Res* 2012;**40**:e112.
45. Hubbard SJ, Thornton JM. *NACCESS, Computer Program*. Department of Biochemistry and Molecular Biology, University College London, London, 1993.
46. Ahmad S. Sequence-dependence and prediction of nucleotide solvent accessibility in double stranded DNA. *Gene* 2009;**428**:25–30.
47. Singh YH, Andrabi M, Kahali B, *et al*. On nucleotide solvent accessibility in RNA structure. *Gene* 2010;**463**:41–8.
48. Ligeti B, Vera R, Juhasz J, *et al*. CX, DPX, and PCW: web servers for the visualization of interior and protruding regions of protein structures in 3D and 1D. *Methods Mol Biol* 2017;**1484**:301–9.
49. Liu HF, Liu R. Structure-based prediction of post-translational modification cross-talk within proteins using complementary residue- and residue pair-based features. *Brief Bioinform* 2020;**21**:609–20.
50. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
51. Liang S, Zhang C, Liu S, *et al*. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res* 2006;**34**:3698–707.
52. Krüger DM, Neubacher S, Grossmann TN. Protein-RNA interactions: structural characteristics and hotspot amino acids. *RNA* 2018;**24**:1457–65.
53. Lu XJ, Bussemaker HJ, Olson WK. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res* 2015;**43**:e142.
54. Dubchak I, Muchnik I, Holbrook SR, *et al*. Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci U S A* 1995;**92**:8700–4.
55. Dubchak I, Muchnik I, Mayor C, *et al*. Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *Proteins* 1999;**35**:401–7.
56. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
57. Gong S, Zhang C, Zhang Y. RNA-align: quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinformatics* 2019;**35**:4459–61.
58. Lovasz L. Random walks on graphs: a survey. *Combinatorics* 1996;**2**:353–98.
59. Kligun E, Mandel-Gutfreund Y. Conformational readout of RNA by small ligands. *RNA Biol* 2013;**10**:982–9.
60. Iwakiri J, Tateishi H, Chakraborty A, *et al*. Dissecting the protein-RNA interface: the role of protein surface shapes and RNA secondary structures in protein-RNA recognition. *Nucleic Acids Res* 2012;**40**:3299–306.

61. Andrabi M, Mizuguchi K, Sarai A, *et al*. Prediction of mono- and di-nucleotide-specific DNA-binding sites in proteins using neural networks. *BMC Struct Biol* 2009;**9**:30.
62. Oliver C, Mallet V, Gendron RS, *et al*. Augmented base pairing networks encode RNA-small molecule binding preferences. *Nucleic Acids Res* 2020;**48**:7690–9.
63. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol* 1997;**272**:121–32.
64. Choi D, Park B, Chae H, *et al*. Predicting protein-binding regions in RNA using nucleotide profiles and compositions. *BMC Syst Biol* 2017;**11**:16.
65. Salekin S, Zhang JM, Huang Y. Base-pair resolution detection of transcription factor binding site by deep deconvolutional network. *Bioinformatics* 2018;**34**:3446–53.
66. Zhang Y, Wang Z, Zeng Y, *et al*. High-resolution transcription factor binding sites prediction improved performance and interpretability by deep learning method. *Brief Bioinform* 2021. doi: 10.1093/bib/bbab273.
67. Pikovskaya O, Polonskaia A, Patel DJ, *et al*. Structural principles of nucleoside selectivity in a 2′-deoxyguanosine riboswitch. *Nat Chem Biol* 2011;**7**:748–55.
68. Dimastrogiovanni D, Fröhlich KS, Bandyra KJ, *et al*. Recognition of the small regulatory RNA RydC by the bacterial Hfq protein. *Elife* 2014;**3**:e05375.
69. Liu L, Li X, Ma J, *et al*. The molecular architecture for RNA-guided RNA cleavage by Cas13a. *Cell* 2017;**170**:714–26.
70. Staker BL, Feese MD, Cushman M, *et al*. Structures of three classes of anticancer agents bound to the human topoisomerase I-DNA covalent complex. *J Med Chem* 2005;**48**:2336–45.
71. Prieto J, Redondo P, López-Méndez B, *et al*. Understanding the indirect DNA read-out specificity of I-CreI Meganuclease. *Sci Rep* 2018;**8**:10286.
72. Kinkelin K, Wozniak GG, Rothbart SB, *et al*. Structures of RNA polymerase II complexes with Bye1, a chromatin-binding PHF3/DIDO homologue. *Proc Natl Acad Sci U S A* 2013;**110**:15277–82.
73. Zhou T, Yang L, Lu Y, *et al*. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* 2013;**41**:W56–62.
74. Singh J, Hanson J, Paliwal K, *et al*. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat Commun* 2019;**10**:5407.
75. Andrabi M, Hutchins AP, Miranda-Saavedra D, *et al*. Predicting conformational ensembles and genome-wide transcription factor binding sites from DNA sequences. *Sci Rep* 2017;**7**:4071.
76. Chen Z, Zhao P, Li F, *et al*. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform* 2020;**21**:1047–57.
77. Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res* 2019;**47**: e127.
78. Lam JH, Li Y, Zhu L, *et al*. A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nat Commun* 2019;**10**:4941.
79. Alley EC, Khimulya G, Biswas S, *et al*. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 2019;**16**:1315–22.
80. Zheng J, Xie J, Hong X, *et al*. RMalign: an RNA structural alignment tool based on a novel scoring function RMscore. *BMC Genomics* 2019;**20**:276.
81. Ge P, Zhang S. STAR3D: a stack-based RNA 3D structural alignment tool. *Nucleic Acids Res* 2015;**43**:e137.
82. Nguyen MN, Verma C. Rclick: a web server for comparison of RNA 3D structures. *Bioinformatics* 2015;**31**:966–8.