To Improve Prediction of Binding Residues With DNA, RNA, Carbohydrate, and Peptide Via Multi-Task Deep Neural Networks

Zhe Sun, Shuangjia Zheng, Huiying Zhao, Zhangming Niu, Yutong Lu, Yi Pan, and Yuedong Yang

Abstract—Motivation: The interactions of proteins with DNA, RNA, peptide, and carbohydrate play key roles in various biological processes. The studies of uncharacterized protein–molecules interactions could be aided by accurate predictions of residues that bind with partner molecules. However, the existing methods for predicting binding residues on proteins remain of relatively low accuracies due to the limited number of complex structures in databases. As different types of molecules partially share chemical mechanisms, the predictions for each molecular type should benefit from the binding information with other molecule types. Results: In this study, we employed a multi-task deep learning strategy to develop a new sequence-based method for simultaneously predicting binding residues/sites with multiple important molecule types named MTDsite. By combining four training sets for DNA, RNA, peptide, and carbohydrate-binding proteins, our method yielded accurate and robust predictions with AUC values of 0.852, 0.836, 0.758, and 0.776 on their respective independent test sets, which are 0.52 to 6.6% better than other state-of-the-art methods. To my best knowledge, this is the first method using multi-task framework to predict multiple molecular binding sites simultaneously.

Index Terms—Protein binding sites, nucleic acid molecule, carbohydrate, peptide, LSTM, Multi-task, Sequence-based approach

1 INTRODUCTION

Predicting proteins interactions with other molecules is critical for understanding biological processes and discovering drugs. The most important molecules controlling biological activities include DNA, RNA, peptides and carbohydrates (CBH). For example, interactions between protein and nucleic acids are central to many of the vital processes in molecular biology [1] such as transcription, post-transcriptional modification and regulation. Additionally, proteins interacting with their respective binding partner.

This study was supported in part by the National Key R&D Program of China under Grant 2020YFB0204803, National Natural Science Foundation of China under Grants U1611261, 61772566, and 81801332, Guangdong Frontier & Key Tech Innovation Program under Grants 2018B010109006 and 2019B020228001, and Introducing Innovative and Entrepreneurial Teams (2016ZT06D211). (Correspondence author: Yuedong Yang.)

Digital Object Identifier no. 10.1109/TCBB.2021.3118916

RNA-binding proteins can modulate or stabilize RNA structures to make RNA catalytically active [2]. The carbohydrate-binding proteins act as important biomarkers for cell communication, cell adhesion, fertilization, development and differentiation [3]. Peptide interactions with protein domains occur in many cell processes particular signaling pathways [4]. In order to better understand the molecular mechanisms, we need to know the binding residues of these proteins interacting with their respective binding partner. Traditional experimental techniques such as X-ray and NMR experiments are robust but are expensive and time-consuming [5]. With the exponentially increasing protein sequences, it is demanding to make predictions of binding residues from sequences.

In the past decades, many bioinformatics methods have been developed to use physicochemical and evolutionary features extracted from sequences [6]. For DNA-binding residue predictions, representative methods include TargetRNA [7] and HMMBinder [8] based on SVM, CNNSites [9] based on CNN network, DRNApred [10] for accurately predicting and discriminating between DNA- and RNA-binding residues, and SPOT-DNA-Seq [11] based on alignments with known DNA-binding proteins. For RNA-binding residue predictions, methods include PredRBR [12] developed based on the gradient tree boosting, and RNAProSite [13] and others [14]–[16] developed based on machine learning techniques. For predictions of carbohydrate binding residues, the common idea is to find residues frequently observed on the sugar interface [17], [18]. For predictions of peptide-binding residues, several methods have been developed based on machine learning techniques, e.g., PRINT [17] and SPOT-Peptide [19]. Other studies include prediction of
heme binding residues from sequence [20] or predictions from 3D structure [21]. Although many successful methods have been proposed, most of them suffer from low accuracies due to small training sample sizes. The underlying reason is that the complex structures of proteins are difficult to obtain by experiments. To be worse, the predictions of binding residues with different types of molecules were traditionally treated as different problems [22], and the prediction tasks were usually performed separately. Thus, the small sizes of individual binding data sets prevent the applications of deep learning techniques. In fact, most biological molecules are organic molecules, and the similarities in physicochemical properties enable the sharing of interaction patterns. For example, the combined inputs of DNA- and RNA-binding residues into the same learning system have been proven to improve the predictions through the support vector machine techniques [23]. However, these studies have ignored the differences between DNA and RNA molecules, and no study has yet been performed to combine binding information with other molecular types. For this purpose, the multi-task learning provides a promising framework that learns shared information through common networks while retaining the task-related output layers. The shared networks between multiple similar tasks enable a bigger training set that could maintain a larger network.

In this study, we designed a multi-task network architecture (namely MTDsite) to simultaneously predict respective binding residues with DNA, RNA, carbohydrate, and peptide molecules. The shared networks among all tasks can collectively strong abstracting capabilities, and we used the Long Short-Term Memory (LSTM) as our Shared network to collect the information of long-range residues in the protein chain. At the same time, four small specific sub-networks were respectively trained for four individual types to extract individual properties. The benchmark tests indicated the employing of multi-task learning leads to averagely 3.6% improvements over state-of-the-art methods when measured by the area under the receiver operating characteristic curve (AUC).

## 2 METHODS

### 2.1 Benchmark Datasets

We evaluated our method by using the previously curated training and testing datasets. The datasets include the protein binding with DNA, RNA, peptide, and carbohydrate, where a residue was defined as a binding residue if it contains at least one atom within 3.5Å from its binding partner. The sequence identities between the proteins in the training and test sets are less than 30% according to BLASTCLUST [24]. Table 1 lists the datasets, with details as:

The DNA and RNA datasets: The datasets were collected from a recent study [25], where the training and testing datasets are 309 and 47 chains for DNA, and 157 and 17 chains for RNA, respectively.

The peptide dataset: The dataset was downloaded from a recent study [27] that includes 157 chains for training and 17 chains for test. The dataset was originally derived from the PROCARB [28] by keeping only high resolution (<3Å) complex structures.

### 2.1.1 Input Features

We used the same input features for each residue to our previous studies [29] except that we excluded 7 physicochemical features. This is because the addition of physicochemical features didn't bring improvements (results not shown). Finally, the input includes totally 54 features composed of G-PSSM (20 features), G-HHM (20), and G_SPD3 (14) as detailed below.

- **G-PSSM:** Evolutionarily conserved residues are considered to be the same or similar residues maintained between species by natural selection. They have important functional roles like acting as binding sites. In this study, we employed the position specific scoring matrix (PSSM) which is a 20×L dimensional matrix (where L is protein length) generated from PSI-BLAST with E-value threshold of 0.001 in three iterations.
- **G-HHM:** The hidden Markov models (HMMs) have been successfully used in protein structure prediction [30] that assumed a Markov process with unobserved states, and the profile HMMs accomplish the protein structure prediction task well based on the HMM-HMM alignment. The sequence alignment generated by HHblits has been found with higher accuracy than by PSI-BLAST [31]. Here, HMM profile was generated by HHblits that compared to the query sequence with the proteins in the uniprot20_2015_06 database.
- **G-SPD3:** Predicted structural properties by SPIDER3 The structural properties were predicted by SPIDER 3.0 [32], including:
  - **ASA (2):** The accessible surface area (ASA) means the surface area of a biomolecule accessible to a solvent, which
reflected the functional importance of residues. Here, ASA of each residue was obtained by SPIDER 3.0, and the ASA was predicted as relative ASA named as rASA. We also computed the average rASA of the residue and its four adjacent residues.

**Torsional Angles (θ):** The backbone torsional angles are composed of $\Phi$, $\psi$, and $\omega$ that are used to describe local backbone structure of a protein. The $\omega$ was not used here because it is usually at 180° due to the planarity of the peptide bond. The angles between neighboring residues include $\theta$ and $\tau$. Here, we let $\theta$ be the angle between Cα$_{i-1}$ -Cα$_{i}$ -Cα$_{i+1}$, and $\tau$ be the dihedral angle rotated about the Cα$_{i}$ -Cα$_{i+1}$ vector to compute the feature. We extracted the features by using cosine and sine values of the four angles, totally 8 features.

$CN$: CN is the number of residues within a distance cutoff to a given residue in three-dimensional space.

**HSE (θ):** Half-Sphere Exposure (HSE) is an extension of CN, which considers directions of two residues in a top and bottom half of the sphere. Two methods were used to define the plane separating the upper and lower hemisphere, which included HSEo based on the neighboring Ca-Cα directional vector and HSE $\beta$ based on the Ca-Cα direction. In this study, we used the HSE $\alpha$ to describe the feature. For CN and HSE, residue distance is defined as the distance between Ca atoms with a 13 Å cutoff.

**Performance Evaluation.** The performance of methods was measured by the number of correct classified and the number of misclassified instances using the terms below:

**TP:** number of actual binding residues predicted as binding sites

**TN:** number of actual non-binding residues predicted as non-binding sites

**FP:** number of actual non-binding residues incorrectly predicted as binding sites

**FN:** number of actual binding residues incorrectly predicted as non-binding sites

We evaluated the performance of our proposed prediction method in terms of the Matthews’ Correlation Coefficient (MCC), Accuracy, Receiver Operating Characteristic (ROC) curve as:

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

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

$$Sensitive = \frac{TP}{POS}$$

$$Specificity = \frac{TN}{NEG}$$

$$Precision = \frac{TP}{TP + FP}$$

where POS is the number of known binding residues, and NEG is the number of known non-binding residues. MCC varies between 0 and 1, with 1 representing correct prediction for all residues, and 0 by random. Additionally, the Area Under the Curve (AUC) was adopted as our primary evaluation index because of our unbalanced datasets, i.e., the much less numbers of positive than negative samples.

### 2.2 MTDsite Architecture

As shown in Fig. 1, the deep learning network in this method consists of two parts. The first part is a shared Bi-directional Long Short-term Memory (BiLSTM) network called shared network (referred to as BiLSTM), which was used to extract common information from different binding molecules. The second part is four individual networks composed of full connection layer. The predictive fraction of protein-molecules binding residues can be obtained from this part. We didn’t add any shared network in the full connection layers for a balance between the sizes of neural network and the training dataset.

### 2.3 Bi-LSTM Network

Long short-term memory (LSTM) is a recurrent neural network (RNN) architecture [33]. The LSTM network was shown to have better performances than traditional RNN in processing, classifying, and predicting time series when there are indefinitely long separations between important events [34]. This is the main reason why LSTM outperforms alternative Hidden Markov Models and other sequence learning methods in numerous applications [35].

In the prediction of binding residues, extracting information between long range residue pairs is important for constructing an accurate model. Traditional machine learning methods, such as Xgboost and SVM, extract feature information of adjacent residues by creating slide-window. In comparison, LSTM collects information on adjacent residues by establishing various ‘gates’, which saves time on training and adjusting parameters.

BiLSTM is a combination network of BRNN and LSTM by stacking multiple BiLSTM-RNN hidden layers together to build a deep bidirectional LSTM-RNN. The output of one layer is used as the input of the next layer. The hidden state sequence, $h_i^n$ consist of forward and backward sequence $\tilde{h}_i^n$ and $\tilde{h}_i^n$, iteratively computed from $n = 1$ to $N$ and $t = 1$ to $T$ as follows:

$$\tilde{h}_i^n = \sigma(W_{\tilde{h}_{i-1}^{n-1}}\tilde{h}_{i-1}^n + W_{\tilde{h}_i^{n-1}}\tilde{h}_{i-1}^{n-1} + b_\tilde{h}^n)$$

$$h_i^n = \sigma(W_{h_{i-1}^{n-1}}h_i^{n-1} + W_{h_i^{n-1}}h_i^{n-1} + b_h^n)$$

$$y_t = \sigma(W_{h_{N0}}h_i^{N} + W_{h_{h}^{N}}\tilde{h}_i^{N} + b_y)$$

where $y = (y_1, y_2, y_3, \ldots, y_t)$ is the output, $\sigma$ is the activation function.

### 2.4 Multi-Task Networks

The extracted features from BiLSTM were used to predict binding sites through four individual networks specifically for four binding types. Each individual neural network corresponds to the prediction of binding residues for one binding type, which has the same architecture to perform repeated linear and non-linear transformations on the input. Let $x_i$ represent the input of $i$-th layer of the network where $X_0$ is the input feature vector. The transformation performed as:

$$x_{i+1} = \sigma(W_i x_i + b_i)$$
where $W_{i}$ and $b_{i}$ are the weight matrix and bias value for the $i$-th layer respectively, and $\sigma$ is the nonlinear function (ELU for middle layer and Sigmoid for the last output layer). Each chain will be used to train the shared networks and the individual network for its binding type $t$ without affecting other three individual networks by using a loss function:

$$L_{c}(y_{i}, P_{i}) = -\delta_{c,t} \sum_{i=1}^{N_{c}} (y_{i} \log(P_{i}) + (1 - y_{i}) \log(1 - P_{i}))$$

where $y_{i}$ and $P_{i}$ are the binding type (1 for binding and 0 for non-binding) and predicted binding probability for residue $i$ in the input protein chain $c$ with length $N_{c}$. $\delta_{c,t}$ is set 1 if protein chain $c$ is known binding type $t$, otherwise 0.

### 2.5 MTDsite Networks

The networks consist of two parts, the shared network and four specific networks for individual tasks. The input of shared network is a $54 \times L$ (L is the length of a protein and 54 is the number of the features) feature matrix. Through two shared LSTM hidden layers, a $2 \times L$ scoring matrix can be obtained, which is used as input for the next part.

As different tasks have specific properties, the second part is four independent fully connected networks specifically trained for four different tasks. The four specific networks have identical structures (layer and neuron sizes). For each task, only the corresponding specific network will be updated, with the left three specific networks unchanged.

In order to prove the importance of BiLSTM network, we designed a group of comparative experiments, in which the model containing only two full connection layers was used to predict binding residues. AUC values of DNA, RNA, Peptide, and Carbohydrate are far lower than those of MTDsite, which are 0.534, 0.529, 0.517, and 0.522 respectively. The possible reason is that the full connection layer is unable to obtain the information between amino acids in the protein sequence.

### 2.6 Cross-Validation and Independent Test

The cross validations and independent tests were employed to evaluate the robustness and performance of the method. The training data set was evenly divided into ten pieces (folds) at random. In each round, nine folds were employed for training and the remaining fold was used for test. This process was repeated for 10 times so that each fold has been tested once, and all outputs were collected to compute the performances. Based on the optimal hyper-parameters, a model trained by using the whole training set was then tested on the independent test set.
2.7 Model Selection and Parameter’s Optimization

During the optimization of the MTDsite models and hyperparameters, we selected optimal hyperparameters with the highest AUC value through the 10-fold cross validation on the training set. We selected [12,3] for the number of hidden layers, (64, 128, 256) for the number of hidden nodes, and [0.01, 0.001, 0.0001] for the learning rate, and the final optimal values for these are 2, 128, and 0.001, respectively. The specific networks were simply set as the same size, and the optimal setting is single-layer fully connected networks with 64 hidden nodes. We also tried using different size for the four specific networks, but there were no significant changes in the performances (Results not shown). The epoch number is 21. Due to the different lengths of all protein sequences, the batch-size was set as 1 to avoid a decrease in accuracy caused by padding, though this has slowed down the training of the neural network.

For a direct comparison, we trained MTDsite-single models like traditional ways by independently inputting single binding type of training data, and thus four models were obtained for four binding types, respectively. If not specifically mentioned, the model trained on one binding type will be employed to test the corresponding type. The architecture of MTDsite-single is a two-layer BiLSTM network, which is the same as the shared network of MTDsite. The final epoch parameters are 17, 19, 26, and 28 for DNA, RNA, peptide and carbohydrate-binding models.

3 RESULTS

3.1 Performances of MTDsite on the 10-Fold Cross Validations and Independent Tests

Table 2 shows the performances of MTDsite measured by AUC, MCC, sensitivity, and specificity in prediction of DNA-, RNA-, peptide-, and carbohydrate-binding residues. By the 10-fold cross validation, MTDsite achieved AUC values of 0.866, 0.857, 0.760, and 0.779 for the four binding types, respectively. The predictions of DNA- and RNA-binding residues have relatively greater AUC values, likely because DNA and RNA are negatively charged and thus the binding residues are relatively easier to predict. The independent tests have obtained essentially the same AUC values with differences of only 0.002~0.021 for the four tests, indicating the robustness of our models. Fig. 2 shows the ROC curves by the cross validations and independent tests on four binding types, respectively. The performances were also confirmed by the consistent maximum MCC values between the 10-fold cross validations and independent tests, with differences of 0.004~0.021. At the thresholds with the maximum MCC values, our models have high specificities while relatively low sensitivities due to the much greater numbers of non-binding residues (negatives) than the binding residues (positives).

We evaluated the contributions of individual feature group by using only single feature group or excluding one feature group from all features. As shown in Table 3, when individual feature group was used in the prediction, G-PSSM, the evolution features produced by PSIBLAST, yielded the greatest values in regard with the average values of both AUC and MCC values. G-HHM, another feature group produced by HHblits, yielded slightly lower AUC and MCC values. G-SPD3, the structural feature group produced by SPIDER3 package, yielded significantly lower AUC values in average. These results suggest the importance of evolution information for protein binding, consistent with previous findings [17], [29]. When excluding individual feature group, the removal of G-PSSM caused the largest decreases in the average values of both AUC and MCC, again indicating its most important role. Though the removal of the G-SPD3 feature group caused the smallest decrease, the difference is significant (P = 0.001) according to the paired t-test. The decreases are small likely because the G-SPD3 features were derived from the PSSM and HHM profiles, and our neural networks could partly catch the structural information from the two profiles.

3.2 Contributions By the Shared Networks

To evaluate the contributions of the networks shared by different binding types, we trained four MTDsite-single models for comparison, each with single type of binding data. As shown in Fig. 2, the ROC curves of MTDsite-single are consistently below those of MTDsite by independent tests on four binding types. As a result, the AUC values by MTDsite-single are 3.6%, 3.2%, 3.3%, and 4.0% lower than those by MTDsite for DNA, RNA, peptide, and carbohydrate, respectively (Table 4). The difference of ROC curves between MTDsite and MTDsite-single reflect the improvement affected by the data set fusion and multi-task learning. When measured by MCC, the MTDsite-single are 4.7%, 36.7%, 37.2%, and 15.4% worse than the MTDsite.

We further performed cross-type tests by training classifier from one binding type and test it on other binding types. As shown in Table 4, on prediction of DNA and RNA binding residues, similar performances have been achieved by the MTDsite-single (DNA) and MTDsite-single (RNA), indicating a similarity between DNA- and RNA-binding sites. This also explains why predictions could be improved by a simple combination of DNA and RNA binding residues in the previous studies. By comparison, there are significant decreases on other cross-type, among which the MTDsite-single (PEP) made almost random predictions on other three binding types although the peptide-binding dataset used for training has the biggest sample size. Interestingly, the MTDsite-single (DNA) and MTDsite-single (RNA) models produced good predictions on the carbohydrate-binding...
dataset: only 3.9% and 5.2% lower AUC than MTDsite-single (CBH). This is likely because carbohydrate has common properties with the desoxyribose and ribose respectively contained in DNA and RNA molecules. The results indicated that MTDsite was not largely impacted by cross-type prediction. On the other hand, the significant changes of four MTDsite-single models on different types of datasets indicated that they might distinguish the binding sites of different molecules on proteins.

By comparison in the MTDsite, although the specific networks have always produced the best predictions for their respective binding types, other specific networks in the MTDsite could also achieve reasonable predictions. For example, on the prediction of peptide-binding residues, MTDsite (CBH) achieved essentially the same AUC value as the MTDsite (PEP), indicating its potential generality to other binding types.

Fig. 3 shows a direct comparison of MTDsite and MTDsite-single. Among the 237 protein chains from four independent datasets, MTDsite significantly outperforms MTDsite-single. Among the 237 protein chains from four independent datasets, MTDsite significantly outperforms MTDsite-single with P-value of 0.003 according to the paired t-test, where MTDsite have greater AUC value for 162 (68%) chains.

### 3.3 Comparisons With Other Methods

MTDsite was compared with other methods on the independent test sets respectively for four binding types. As

---

**TABLE 3**

| Feature Groups | AUC   | MCC   | Feature Groups | AUC   | MCC   |
|----------------|-------|-------|----------------|-------|-------|
|                |       |       | MTDsite        | 0.806 | 0.328 |
| G-PSSM         | 0.742 | 0.247 | -G-PSSM        | 0.757 | 0.281 |
| G-HHM          | 0.739 | 0.244 | -G-HHM         | 0.767 | 0.302 |
| G-SPD3         | 0.61  | 0.201 | -G-SPD3        | 0.794 | 0.320 |

*Performances based on individual feature groups;  
*Performances by excluding individual feature groups.
shown in Table 5, MTDsite achieved the highest AUC values through all four types of independent datasets, which are 2.2%, 4.4%, 6.6%, and 0.52% higher than other best methods for DNA, RNA, peptide, and carbohydrate datasets, respectively. The improvements are mostly contributed by the multi-task learning as the MTDsite-single models without using multi-task learning yielded close and mostly lower performances than other state-of-the-art methods. Fig. 2 also indicated that all other methods are mostly below the ROC curves by MTDsite in all four tests. The only exception is the SPRINT-CBH method, previously developed by our group for the carbohydrate-binding prediction, which is marginally above the ROC by MTDsite, though MTDsite achieved a higher AUC value. Here, the lower MCC by MTDsite is likely because MTDsite has been optimized for the greatest AUC values while SPRINT-CBH was optimized for MCC values. Additionally, the carbohydrate dataset shows a much larger ratio between the number of negatives and the positives in the test set (26.0 on the carbohydrate compared to 8.5, 6.5, and 16.8 on the DNA, RNA, peptide datasets, respectively), and thus the shared network wasn’t well optimized for the carbohydrate dataset. As a result, the MCC values of MTDsite are 11.8%, 1.4%, and 7.2% percent higher than other methods for DNA, RNA, and peptide, respectively. At the same time, MTDsite achieved the highest accuracy, sensitivity, and precision in

| Table 4 |
|---|
| Comparison of AUC Values of Different MTDsite Models for Different Binding Molecules on Independent Test |
| Models | DNA | RNA | Peptide | CBH |
|---|---|---|---|---|
| MTD-DNA | 0.852 | 0.833 | 0.747 | 0.756 |
| MTD-RNA | 0.829 | 0.836 | 0.738 | 0.742 |
| MTD-PEP | 0.732 | 0.748 | 0.758 | 0.731 |
| MTD-CBH | 0.768 | 0.761 | 0.757 | 0.776 |
| MTD-Single (DNA) | 0.822 | 0.801 | 0.601 | 0.718 |
| MTD-Single (RNA) | 0.801 | 0.810 | 0.600 | 0.709 |
| MTD-Single (PEP) | 0.531 | 0.520 | 0.734 | 0.51 |
| MTD-Single (CBH) | 0.632 | 0.614 | 0.552 | 0.746 |

| Table 5 |
|---|
| Comparison With Other Methods on the Independent Test Sets Based on AUC, ACC, MCC, Sensitivity, and Precision |
| Method | AUC | MCC | ACC | Sen | Pre |
| DNA | | | | | |
| BindN+ | 0.806 | 0.256 | NA | NA | NA |
| SVMnuc | 0.833 | 0.336 | NA | 0.26 | 0.51 |
| NucBind | 0.834 | 0.355 | NA | 0.29 | 0.50 |
| DRNApred | 0.757 | 0.232 | NA | 0.32 | 0.25 |
| COACH-D | 0.810 | 0.266 | NA | NA | 0.70 |
| MTDsite-single | 0.822 | 0.379 | 0.901 | 0.562 | 0.628 |
| MTDsite | 0.852 | 0.397 | 0.916 | 0.684 | 0.734 |
| RNA | | | | | |
| BindN+ | 0.738 | 0.219 | NA | NA | NA |
| SVMnuc | 0.784 | 0.205 | NA | 0.19 | 0.26 |
| NucBind | 0.801 | 0.210 | NA | 0.18 | 0.29 |
| DRNApred | 0.687 | 0.056 | NA | 0.09 | 0.08 |
| COACH-D | 0.711 | 0.356 | NA | NA | 0.45 |
| MTDsite-single | 0.810 | 0.264 | 0.945 | 0.455 | 0.345 |
| MTDsite | 0.836 | 0.361 | 0.931 | 0.612 | 0.351 |
| Peptide | | | | | |
| VisGrid | NA | 0.224 | 0.609 | 0.38 | NA |
| PinUp | NA | 0.007 | 0.547 | 0.207 | NA |
| PepSite | NA | 0.141 | 0.598 | 0.221 | NA |
| Peptimap | NA | 0.262 | 0.629 | 0.436 | NA |
| SPRINT | 0.711 | 0.079 | 0.662 | 0.642 | NA |
| MTDsite-single | 0.734 | 0.218 | 0.947 | 0.32 | 0.29 |
| MTDsite | 0.758 | 0.299 | 0.956 | 0.344 | 0.366 |
| CBH | | | | | |
| SPRINT-CBH | 0.772 | 0.285 | 0.961 | 0.223 | NA |
| StackCBPred | NA | 0.134 | 0.664 | 0.665 | 0.664 |
| MTDsite-single | 0.746 | 0.221 | 0.958 | 0.374 | 0.177 |
| MTDsite | 0.776 | 0.255 | 0.963 | 0.209 | 0.171 |
most cases. It should also be noticed that PepSite and Pepti-
map are structure-based methods. Though MTDsite used
sequence-based information only, our method still achieved
better performances.

### 3.4 Case Study

As an example, we demonstrated the predictions on a
restriction-modification controller DNA-binding protein
(PDBid: 3s8q chain B) consisted of 15 actual binding resi-
dues within totally 77 residues. The running costs by
MTDsite and MTDsite-single are essentially the same (~10
minutes on a CPU server of Intel 2.6GHz) since the comput-
tations were mostly cost to generate the evolution profiles
(G-PSSM and G-HMM). The training costs are also similar
as MTDsite could simultaneously generate four predictors
while training on four training sets. As shown in Fig. 4,
MTDsite predicted 15 binding residues, where 14 residues
are truly binding. In comparison, MTDsite-single predicted
14 binding residues including 13 truly binding residues. We
noticed the residue with the 15th highest predicted score is
non-binding one. Thus, the precisions of two methods are
97.6% and 96.3%, corresponding to MCC of 0.918 and 0.869,
and precision of 0.933 and 0.929, respectively.

### 4 Discussion and Conclusion

We have developed a new architecture to use multiple tasks
learning for predicting binding residues. To our best know-
ledge, this is the first attempt of the strategy for prediction of
binding residues. The sharing between DNA, RNA, pep-
tide, and carbohydrate-binding information was indicated
to improve the predictions of binding residues for all four
types. Our method was indicated robust by the consistent
performances between the cross validations and indepen-
dent tests. The method was proven to outperform state-of-
the-state methods. Such strategy was expected to extend to
other tasks like predictions of ligand- and lipid-binding resi-
dues. The framework is also promising for other similar
tasks, such as protein function prediction, modification sites of
protein or DNA/RNA, and predicting properties for
chemical compounds.

Though our method was shown to outperform other
methods, the method still suffers from low sensitivity and
precision. In future, we may learn high-order structural
information from sequence [36], or directly include protein
structures (from experiments or structural modeling) that
were expected to further advance the predictions as shown
in our previous studies [27], [37]. The current sharing of information was simply based on a
port of shared network. The performances were hurt by
the differences between the binding types, for example, dif-
erent ratios of positives and negatives, the preference of
positive-charge residues in the DNA/RNA binding.
Recently, there are new techniques for the multi-task learn-
ing including the meta-learning [38]. We will employ these
recent techniques in future studies.

### References

[1] M. W. Hentze, A. Castello, T. Schwarzl, and T. Preiss, “A brave
new world of RNA-binding proteins,” Nat. Rev. Mol. Cell Biol.,
vol. 19, no. 5, pp. 327–341, 2018.

[2] S. Gerstberger, M. Hafner, and T. Tuschi, “A census of human
RNA-binding proteins,” Nat. Rev. Genet., vol. 15, no. 12, pp. 829–
845, 2014.

[3] W. Lu and R. J. Pieters, “Carbohydrate-protein interactions and mul-
tivalency: Implications for the inhibition of influenza a virus
infections,” Exp. Opin. Drug Discov., vol. 14, no. 4, pp. 387–395, 2019.

[4] E. Petsalaki and R. B. Russell, “Peptide-mediated interactions in
biological systems: New discoveries and applications,” Curr.
Opin. Biotechnol., vol. 19, no. 4, pp. 344–350, 2008.

[5] N. London, D. Movshovitz-Attias, and O. J. S. Schueler-Furman,
“The structural basis of peptide-protein binding strategies,” Struc-
ture, vol. 18, no. 2, pp. 188–199, 2010.

[6] L. Wang, C. Huang, M. Q. Yang, and J. Y. Yang, “BindN+ for
accurate prediction of DNA and RNA-binding residues from pro-
tein sequence features,” BMC Syst. Biol., vol. 4, 2010, Art. no. 53.

[7] J. Hu, Y. Li, M. Zhang, X. Yang, H. Shen, and D. Yu, “Predicting
protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs,” IEEE/
ACM Trans. Comput. Biol. Bioinform., vol. 14, no. 6, pp. 1389–
1398, Nov./Dec. 2017.

[8] S. Y. C. Rianon Zaman, M. A. Rashid, A. Sharma, A. Dehzangi,
and S. Shatabda “HMMBinder: DNA-binding protein prediction
using HMM profile based features,” BioMed Res. Int., vol. 2017,
2017, Art. no. 459069.

[9] J. Zhou, Q. Lu, R. Xu, L. Gui, and H. Wang, “CNNsite: Prediction
of DNA-binding residues in proteins using convolutional neural
network with sequence features,” Inform. Sci., 2016.

[10] L. Kurgan and J. Yan, “DRNApred, fast sequence-based method
that accurately predicts and discriminates DNA- and RNA-bind-
ing residues,” Nucleic Acids Res., vol. 45, no. 10, 2017, Art. no. e84.

[11] H. Zhao, J. Wang, Y. Zhou, and Y. Yang, “Predicting DNA-bind-
ing proteins and binding residues by complex structure predic-
tion and application to human proteome,” PLoS One, vol. 9, no. 5,
2014, Art. no. e96694.

[12] Y. Tang, D. Liu, Z. Wang, T. Wen, and L. Deng, “A boosting
approach for prediction of protein-RNA binding residues,” BMC
Bioinform., vol. 18, 2017, Art. no. 465.

[13] M. Sun, X. Wang, C. Zou, Z. He, W. Liu, and H. Li, “Accurate
prediction of RNA-binding protein residues with two discriminative struc-
tural descriptors,” BMC Bioinform., vol. 17, 2016, Art. no. 231.

[14] Z.-P. Liu, L.-Y. Wu, Y.-R. Wang, X.-S. Zhang, and L. Chen,
“Prediction of protein-RNA binding sites by a random forest
method with combined features,” Bioinformatics, vol. 26, no. 13,
pp. 1616–1622, Jul. 2010.
Yuedong Yang received the PhD degree in computational biology from the University of Science and Technology of China in 2006. He is currently a professor with the School of Data and Computer Science and National Super Computer Center at Guangzhou, Sun Yat-sen University, China. He has authored or coauthored 90 papers in high-profile journals that have been cited 3500 times, including four ESI highly cited articles. Currently his research group emphasizes on developing HPC and AI algorithms for multi-omics data integration and intelligent drug design. He is also responsible for constructing the HPC platform for biomedical applications based on the Tianhe-2 supercomputer.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.