Work Report

贾宁欣

04-25-2021

OXFORD

Briefings in Bioinformatics, 00(00), 2021, 1–12

doi: 10.1093/bib/bbaa397
Problem Solving Protocol

NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning

Jun Zhang, Qingcai Chen and Bin Liu

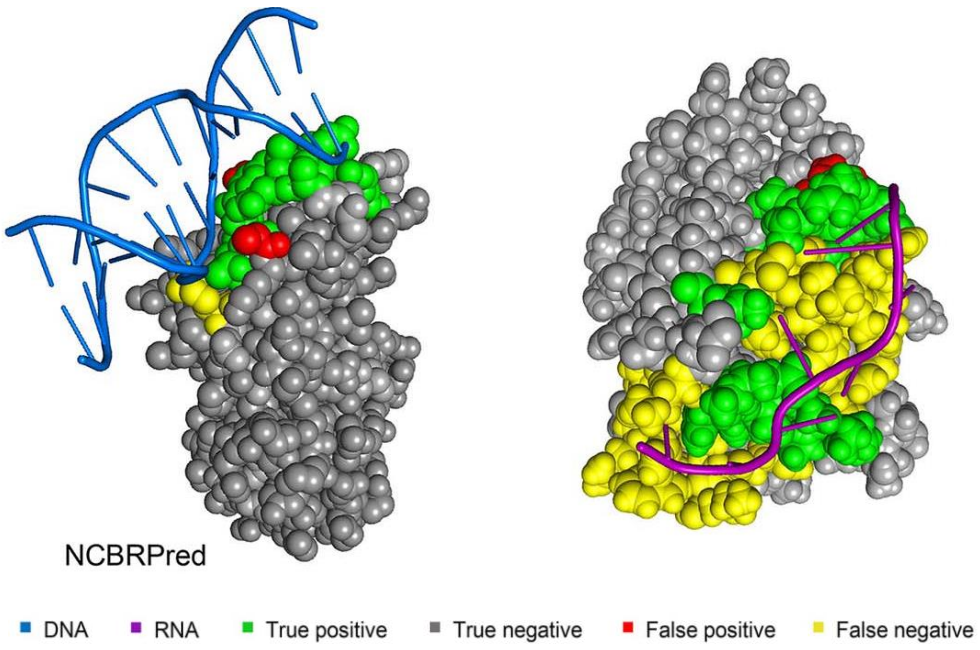
Corresponding author: Bin Liu, Harbin Institute of Technology, HIT Campus Shenzhen University Town, Xili, Shenzhen 518055, China, and School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China. Tel.: (+86) 010-68911310; E-mail: bliu@bliulab.net

Abstract

The interactions between proteins and nucleic acid sequences play many important roles in gene expression and some cellular activities. Accurate prediction of the nucleic acid binding residues in proteins will facilitate the research of the protein functions, gene expression, drug design, etc. In this regard, several computational methods have been proposed to predict the nucleic acid binding residues in proteins. However, these methods cannot satisfactorily measure the global interactions among the residues along protein. Furthermore, these methods are suffering cross-prediction problem, new strategies should be explored to solve this problem. In this study, a new computational method called NCBRPred was proposed to predict the nucleic acid binding residues based on the multilabel sequence labeling model. NCBRPred used the bidirectional Gated Recurrent Units (BiGRUs) to capture the global interactions among the residues, and treats this task as a multilabel learning task. Experimental results on three widely used benchmark datasets and an independent dataset showed that NCBRPred achieved higher predictive results with lower cross-prediction, outperforming 10 existing state-of-the-art predictors. The web-server and a stand-alone package of NCBRPred are freely available at http://bliulab.net/NCBRPred. It is anticipated that NCBRPred will become a very useful tool for identifying nucleic acid binding residues.

Key words: nucleic acid binding residue prediction; cross-prediction problem; multilabel learning; sequence labeling model

NCBRPred



Problem:

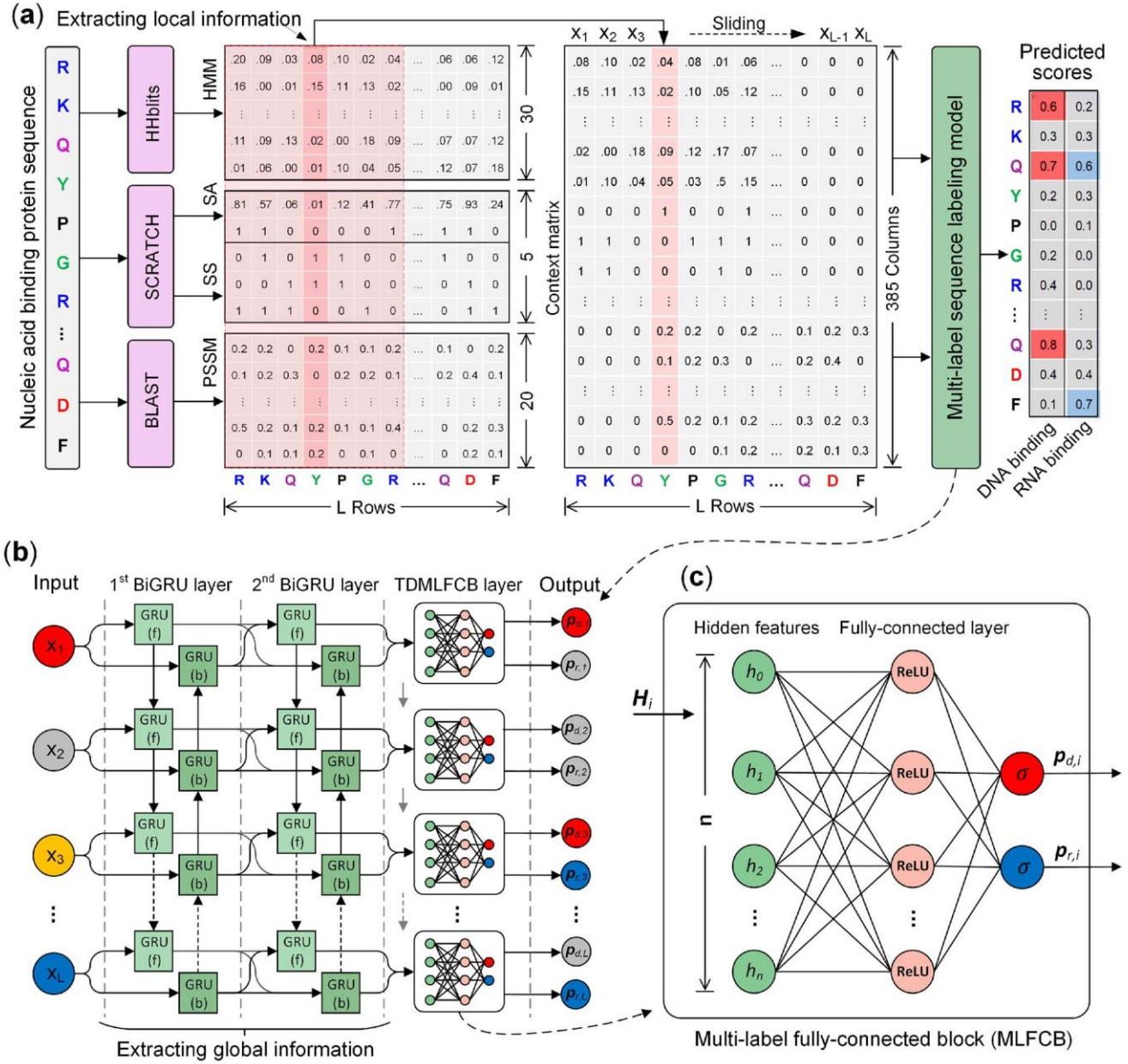
cross-prediction

(If a DNA-binding residue predictor is only trained with DNA-binding proteins and does not consider the RNA-binding proteins, it can accurately predict the DNA-binding residues but also prefers to identify the RNA-binding residues as DNA-binding residues.)

Reason:

DNA-binding residues and RNA-binding residues share some similar characteristics.

NCBRPred


(a) Extracting local information
Nucleic acid binding protein sequence
HHblits
SCRATCH
BLAST
HMM
SA
SS
PSSM
30
5
20
L Rows
Context matrix
Sliding
385 Columns
Multi-label sequence labeling model
Predicted scores
DNA binding
RNA binding
(b)
Input
1st BiGRU layer
2nd BiGRU layer
TDMLFCB layer
Output
GRU (f)
GRU (b)
Extracting global information
(c)
Hidden features
Fully-connected layer
ReLU
Multi-label fully-connected block (MLFCB)

NCBRPred

GRU (gate recurrent unite)

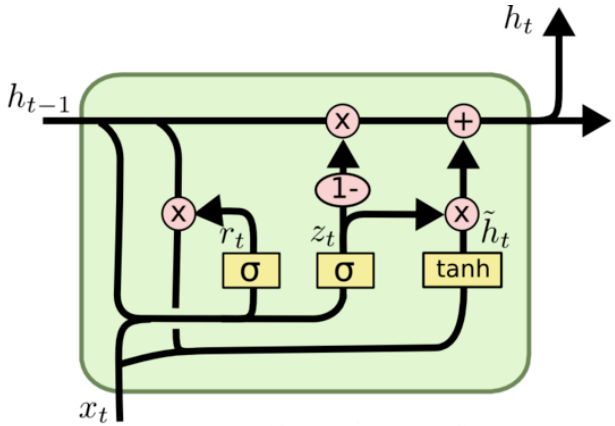
$$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$$

$$\Gamma_u = \sigma(W_u[\, c^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_r = \sigma(W_r[\, c^{<t-1>}, x^{<t>}] + b_r)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$$

$$a^{<t>} = c^{<t>}$$



LSTM (long short term memory)

$$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$$

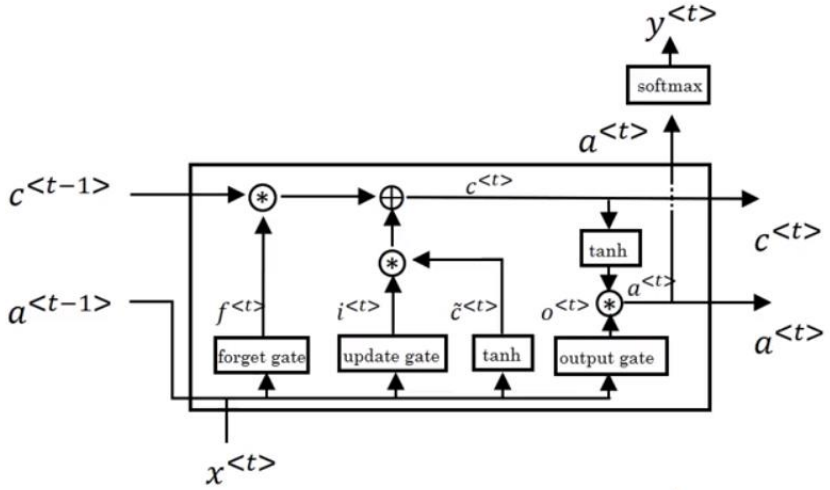
$$\Gamma_u = \sigma(W_u[\, a^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_f = \sigma(W_f[\, a^{<t-1>}, x^{<t>}] + b_f)$$

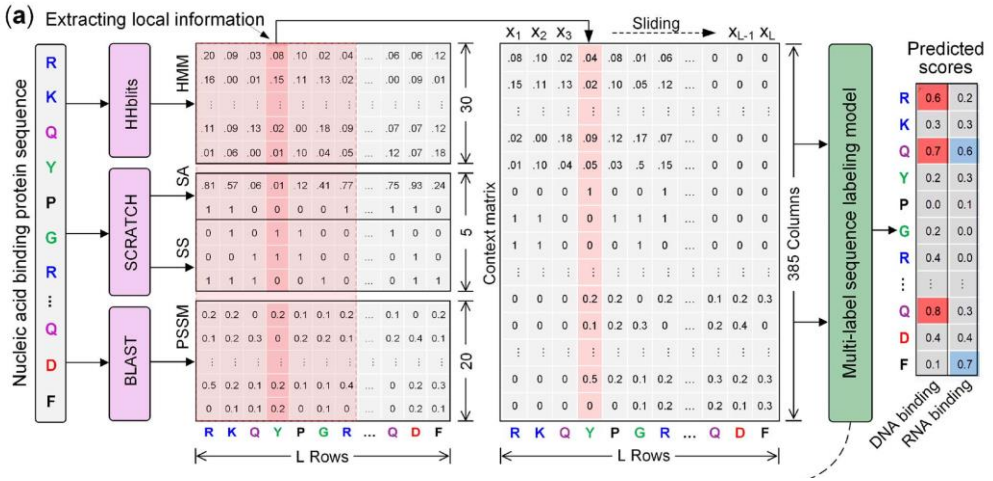
$$\Gamma_o = \sigma(W_o[\, a^{<t-1>}, x^{<t>}] + b_o)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$$

$$a^{<t>} = \Gamma_o * \tanh c^{<t>}$$



NCBRPred



$$loss = \frac{1}{N}\sum_{i=1}^{2}\sum_{j=1}^{N} l_{i,j}$$

$$l_{i,j} = \begin{cases} y_{i,j} \times \log\left(p_{i,j}\right) + \left(1 - y_{i,j}\right) \times \log\left(1 - p_{i,j}\right) & y_{i,j} \geq 0 \\ 0 & y_{i,j} < 0 \end{cases}$$

where N represents the total number of residues in the training data; $y_{i,j}$ is the true label for the i-th label of the j-th residue; $p_{i,j}$ is the predicted score for the i-th label of the j-th residue. For the disorder residues in training datasets, the true labels were set as −1.

1	0	DNA-binding residue
0	1	RNA-binding residue
1	1	both
-1	-1	disordered residue

Table 1. The statistical information of the four datasets used in this study

Dataset	Training set				Test set			
	Protein chains	Residues			Protein chains	Residues		
		D[a]	R[b]	N[c]		D[a]	R[b]	N[c]
YK17 [8]	488	7764	4684	90 594	82	955	807	17 119
YFK16-3.5 [5]	467	6932	4647	88 508	64	875	409	13 679
YFK16–5 [5]	469	10 848	6941	82 811	65	1452	648	12 927
MW15 [21]	NA	NA	NA	NA	46	760	368	9447

[a]DNA-binding residues.
[b]RNA-binding residues.
[c]Residues neither bind to nucleic acid residue nor are disordered residue.

NCBRPred

Abstract

Introduction

Methods

- Datasets
- Protein representation
- Architecture of NCBRPred
- Performance evaluation

Result and discussion

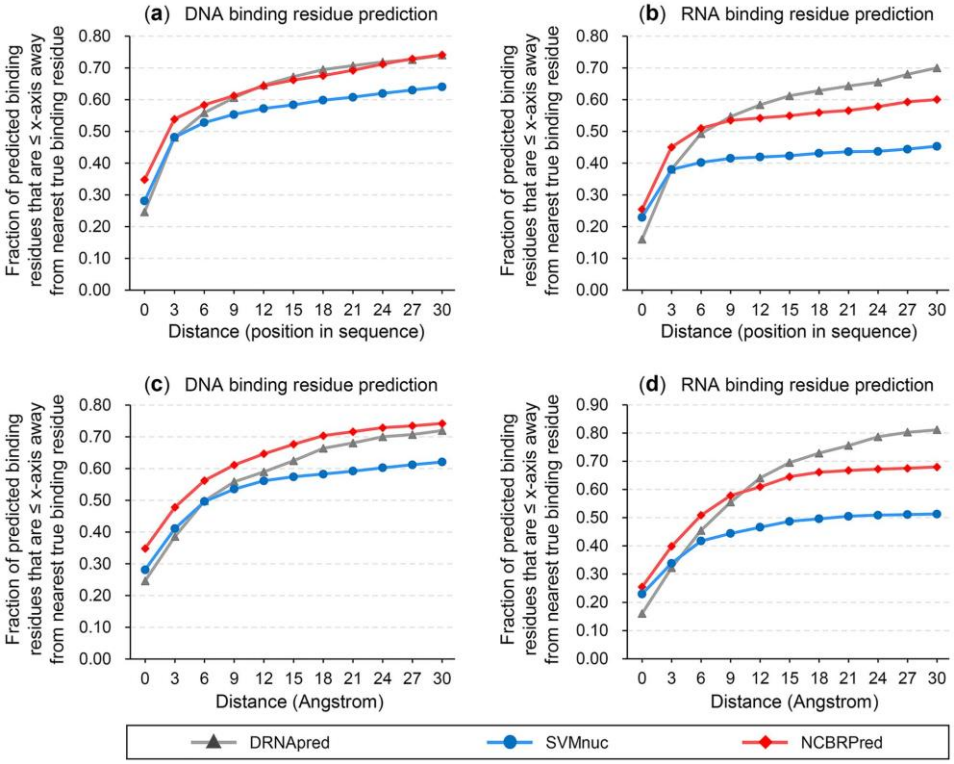
- The predictive performance of NCBRPred based on **different features and their combinations**.
- Impact of the **window sizes** on the predictive performance of NCBRPred.
- Comparison of different models
- Performance comparison of various computational methods (sequence similarity $\leq$ 30%)
- Performance of various methods on the independent dataset MW15 (sequence similarity $\leq$ 25%)
- Analysis of the predicted nucleic acid binding residues
- Predictive results visualization

Conclusion

NCBRPred

Analysis of the predicted nucleic acid binding residues

Fraction of predicted binding residues that are ≤ x-axis away from nearest true binding residue



Predictive results visualization