生物智能算法 神经网络组

Personal information

- Name: 贾磊
- Student ID: 11816008
- Email: jialei0701@foxmail.com

Timeline

Task	Date	Done
1.选择论文	**Mar. 14**	√
2.精读论文，理解模型	**Mar. 21**	√
3.复现论文	**Mar. 28**	√
4.完成对比实验	**Apr. 4**	√
5.形成报告	**Apr. 11**	√

1. 选择论文

Title:

CNNsite: Prediction of DNA-binding Residues in Proteins Using Convolutional Neural Network with Sequence Features.

> IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2016

Abstract:

> Protein-DNA complexes play crucial roles in gene regulation. The prediction of the residues involved in protein-DNA interactions is critical for understanding gene regulation. Although many methods have been proposed, most of them overlooked motif features. Motif features are sub sequences and are important for the recognition between a protein and DNA. In order to efficiently use motif

features for the prediction of DNA-binding residues, we first apply the Convolutional Neural Network (CNN) method to capture the motif features from the sequences around the target residues. CNN modeling consists of a set of learnable motif detectors that can capture the important motif features by scanning the sequences around the target residues. Then we use a neural network classifier, referred to as CNNsite, by combining the captured motif features, sequence features and evolutionary features to predict binding residues from sequences.

摘要

蛋白-DNA复合体在基因调控的过程中扮演着重要的作用。对参与到蛋白-DNA互作的残基（residues）的预测对于理解基因调控有重要意义。现在已经有一些预测方法，但是这些方法忽视了基序（motif）的特征。基序特征是亚序列，其对蛋白质和DNA的识别具有重要意义。为了有效利用基序特征进行DNA绑定残基的鉴定，本研究应用卷积神经网络来提取目标残基周围序列的基序体征。

---

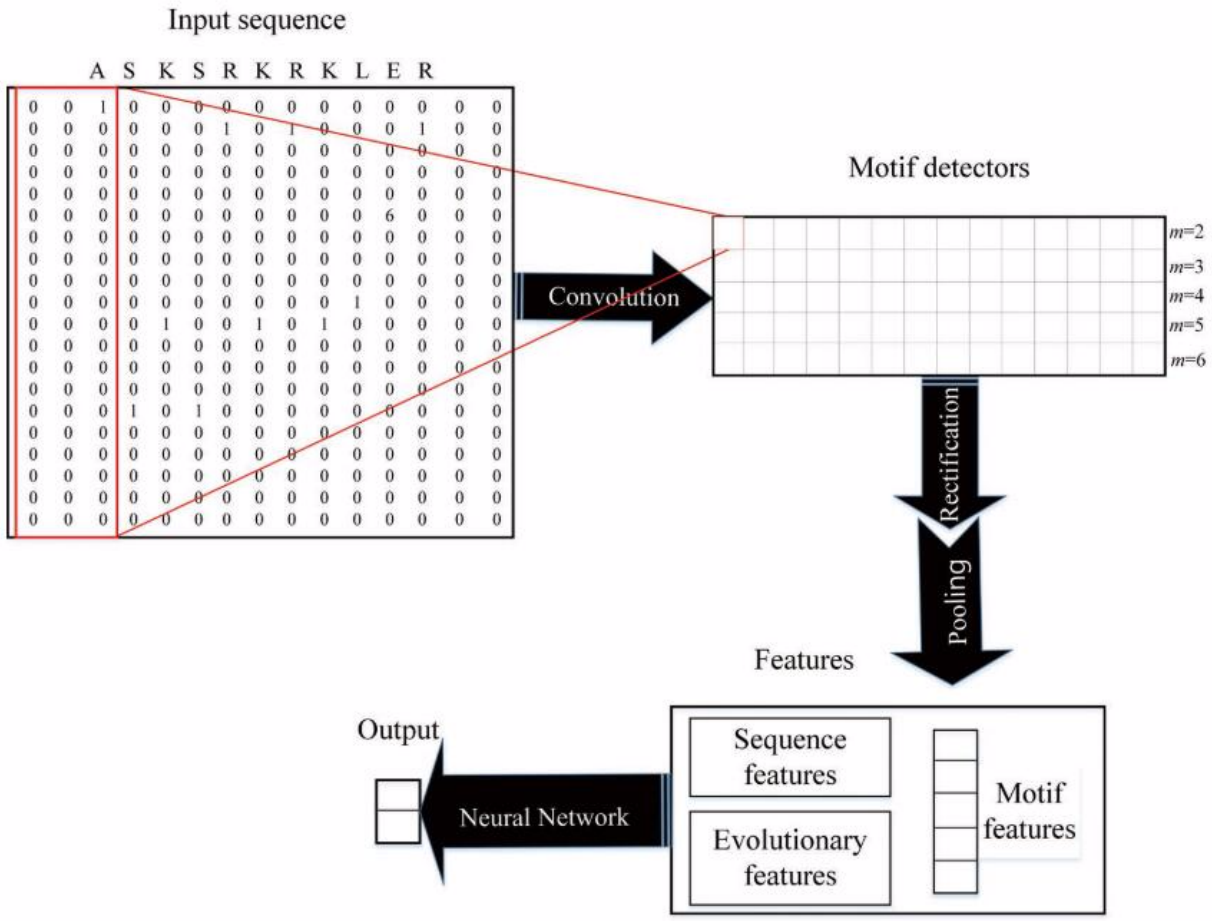
2. 精读论文，理解模型

数据集

TABLE I
THE DETAILS OF THE THREE DATASETS

datasets	PDNA-62	PDNA-224	TS-72
binding residues	1,215	3,778	1,040
non-binding residues	6,948	53,570	13,226

Framework



Convolution layer

输入residue-wise数据S左右填补（m-1）的unuseful residue，转换为矩阵M（类图像像素数据）；

$$M_{i,j} = \begin{cases} 0.5 & \text{if } i < m \text{ or } i > n - m \\ 1 & \text{if } S_{i-m+1} = j^{th} \text{ base} \\ 0 & \text{if otherwise} \end{cases}$$

输出为矩阵X，其中$X_{i,k}$表示第k个motif detector在第i个位置的得分；D具体表示未提供!

$$X_{i,k} = \sum_{j=1}^{m} \sum_{l=1}^{20} M_{i+j,l} D_{k,j,l}$$

Rectification layer

$$Y_{i,k} = max(0, X_{i,k} - b_k)$$

通过模型训练阶段学习得到mofit得分阈值，过滤非高效motif特征

Pooling layer

$$Z_k = max(Y_{1,k}, \cdots, Y_{n,k})$$

最大池化

Neural network layer

综合motif特征、sequence特征、evolutionary特征进行预测。
采用dropout technique避免overfitting。

不同特征比较

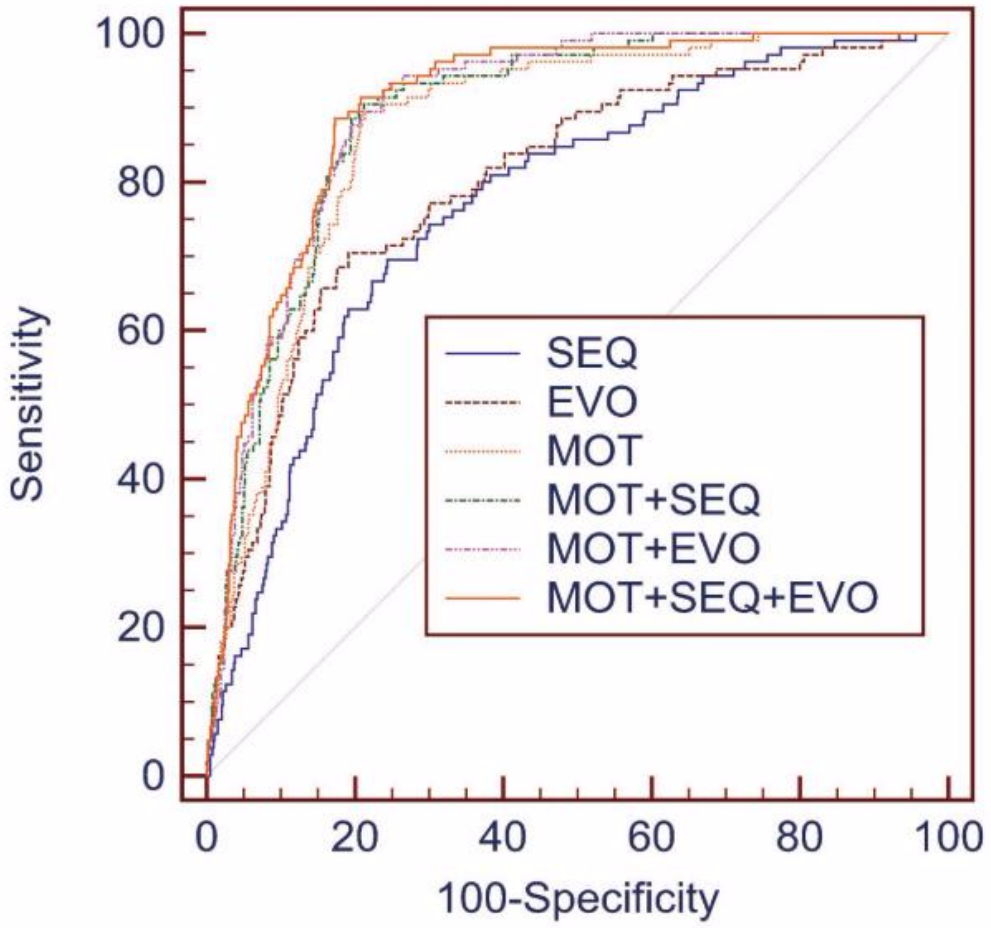


Fig. 2. ROC curves of CNNsite with different combination of features on PDNA-62.

3. 复现论文

代码见附件

文中没有提到模型优化的全过程，此外，模型中有些过程缺乏足够信息，我这边难以进行可视化。

4. 对比实验

方法间比较

TABLE IV
THE PREDICTING PERFORMANCE COMPARED WITH OTHER COMPUTATIONAL METHODS ON PDNA-62

Method	ACC(%)	MCC	SN(%)	SP(%)	ST (%)	AUC
Dps-pred	79.10	–	40.30	81.80	61.10	–
Dbs-pssm	66.40	–	68.20	66.00	67.10	–
BindN	70.30 –	69.40	70.50	69.95	0.752	
Dp-bind	78.10	0.490	79.20	77.20	78.20	–
DP-Bind	77.20	–	76.40	76.60	76.50	–
BindN-RF	78.20	–	78.10	78.20	78.15	0.861
BindN+	79.00	0.440	77.30	79.30	78.30	0.859
PreDNA	79.40	0.420	76.80	79.70	78.30	–
CNNsite	80.63	0.509	85.87	79.78	82.67	0.911

Sensitivity (SN), Specificity (SP), Strength (ST), Accuracy (ACC), and Mathews Correlation Coefficient (MCC).

5.模型解释

Motif特征有效性

Explanation for the effectiveness of motif features for the prediction of DNA-binding residue

Discriminant power (DP) of a motif t in CNNsite is calculated by the following formula:

$$DP(t) = \sum_{i}^{p} \sum_{j}^{d} f_{i,j}(t) \tag{14}$$

$$f_{i,j}(t) = \begin{cases} Z_j & \text{if } argmax\ (Y_{1,j},\ \cdots,\ Y_{n,j}) = pos \\ 0 & others \end{cases} \tag{15}$$

We find that the residues R, K, G are the important compositions of these motifs. This finding is consistent with the study of Szilgyi and Skolnick, in which they found that R, A, G, K and D are important for the formation of protein-DNA interactions. The importance of R for the formation of protein-DNA interactions is further confirmed by Sieber and Allemanns work, which states that R can indirectly interact

with DNA by interacting with both the phosphate backbone and the carboxylate of E(345). Since these residues are important for the formation of protein-DNA interactions, we speculate that they often occur in the context of the DNA-binding residues and their occurrences are important features for prediction.

THE TOP 15 MOTIF FEATURES OF VARIOUS LENGTH WITH THE LARGEST DISCRIMINANT POWER

Length	2	3	4	5	6
1	KR	RNR	KNWV	NRRRK	SNRRRK
2	GR	RMR	WVSN	KGNRS	KGRRGR
3	GN	RGR	CKGF	TRGRV	VSNRRR
4	GK	RLP	KGFF	GRRGR	VSRGRT
5	NR	RKR	GHRF	TRKRK	TTRKRK
6	EK	KTR	HSPA	RGHRF	KKRRKT
7	KT	HSP	VSNR	KRVRG	GIGNIT
8	RN	LKG	YRPG	VSNRR	YKGNRS
9	RT	TRK	KTRK	SNRRR	KSIGRI
10	KG	ALR	IKNW	RGRVK	MKRVRG
11	GT	IQI	FGKM	KGRRG	RKSIGR
12	IS	DSL	SIGR	KTRGR	GSGNTT
13	DK	RKT	FMKR	RVRGS	NKRMRS
14	TR	MRN	KRMR	KRMRS	SKTRKT
15	SR	RKE	RGHR	SRGRT	KTRGRV

THE PROPOSITION OF R,A,G,K AND D IN THE TOP 15 MOTIF FEATURES OF VARIOUS RESIDUES WITH THE LARGEST DISCRIMINANT POWER

length	2(%)	3(%)	4(%)	5(%)	6(%)
A	0.00	2.22	1.67	0.00	0.00
G	16.67	4.44	11.67	16	13.33
K	20.00	13.33	15.00	12	16.67
D	3.33	2.22	0.00	0.00	0.00
Others	36.67	44.44	55.00	29.33	41.11