

Nov 15,
2019

An improved deep learning method for predicting DNA-binding proteins based on contextual features in amino acid sequences [↗](#)

PLOS One

Ruixiong Ma¹¹USTB[1](#) Works for me [dx.doi.org/10.17504/protocols.io.2rdgd26](https://doi.org/10.17504/protocols.io.2rdgd26)
Ruixiong Ma
USTB

ABSTRACT

[With the explosively increased amount of newly discovered proteins, predicting the function of these proteins from amino acid sequences](#) is becoming one of the main challenges in functional annotation of genomes. [Nowadays a number of computational approaches have been developed to predict DNA-binding proteins effectively and accurately from amino acid sequences, such as SVM, DNABP and CNN-RNN. However, these methods do not consider the context in amino acid sequences, which](#) makes it difficult for them to capture sequence features adequately. [In this paper, we propose CNN-BiLSTM, a new method for predicting DNA-binding proteins, elaborately reconciling convolution neural network and bi-directional long short-term memory recurrent](#) neural network. CNN-BiLSTM can explore the potential contextual relationships of amino acid sequences to obtain more features than traditional models. [The experimental results show that the predication accuracy of the proposed CNN-BiLSTM method on the test set is 96.5%, which is 7.8% higher than that of SVM, 9.6% higher than that of DNABP and 3.7% higher than that of CNN-RNN respectively. Being tested on 20, 000 independent samples provided by UniProt that weren't involved in model training, the accuracy of CNN-BiLSTM is 94.5%, which is 12% higher than that of SVM, 4.9% higher than that of DNABP and 4% higher than that of CNN-RNN respectively. The model training process is visualized and compared with that of CNN-RNN, and it is found that the training process of CNN-BiLSTM support better generalization from the training data set, which shows that CNN-BiLSTM has a wider range of adaptations to protein sequences. On the independent samples set, CNN-BiLSTM presents better credibility, for its predicted scores are closer to the labels of the samples than those of CNN-RNN. Therefore, the proposed CNN-BiLSTM is a more powerful method for identifying DNA-binding proteins.](#)

EXTERNAL LINK

<https://doi.org/10.1371/journal.pone.0225317>

model structure.pdf

GUIDELINES

This is a method of recognizing DNA binding proteins by deep learning.

MATERIALS TEXT

WA+YsHNV4TofDdin9k/OdbiNG3EBVYxFwYq0nBXvwhBgN/nEhso4Ps39rYkUa0htUN+ Ae8FvecLVS9H474E2TNZWvp7D/8k7sT80QqTIpk9QFDiNz+Lj5olg1DRwiF9YbSKaVeRV0gz7ejEd1cZK55ULzmPB/luPFQDw2j60G4ecXKq+SGABNRW952c29b1708Cj/wM85S0yTzldVY2G2cYXyfg+WJTPWn1w9Ginw9Xn857bP04r2PpvXeW8WXjYCI0XeI651uSBnQbBGPCEtpHtG+haWIZ2aOu9c7Rjr2//dJnznNwZir35Wnt48PZd+HpqxDTKOE5TrLPmuHcinVNYZQVO5sV9HfBEeGq/hV8At/9XqympVN0iixMspIS84u3JaYwhsZs+FrcVChubnXISW9W8eU15A57vt/s8Vgts+TnJgH+aW05NdBQW7odBKns4MMw2Ac6NqgCIDYvQ1G3ZBIX2ajq+jLspz+mEuuflh+1A9cjzDT3+MIZcQoSNIjAHEqgAApuWhdxUbKdIzKuwQh2Tlp/nJDnFLa5JGJ96M5TiWlycmCQuDCLoTqHJnlE1PiL78ezc7dpFfwFz9HiuLLO8N0Lb/Nb/9tAyuUerSvuqU90Pp8Wg80l8oucrrn16/zTniloi+QCEe7bp64bG0QSaKtra6OoGfqYSAwEoemf8Z4bnu4M+qmdjivgwgUrwIZkiL0Sh38VN1YKgHIIo/zcilPfDVNZumqvS3YwRDZ+qV5W7I2QZ61oslhkKcDc/1/cCCO9MkSgBhOy5Xfwa62ty8uGSfD23HX3AgkcfBg2mfujyolnhus81Y3Wt26tIrA/RcArBpeXOCmpBTBF1+9olkMjScgTk7hcl9Nev1DmYgpJVvGThSsHomPZtHeUH5+0ADEPqqswnm55Ytf4UYdzxraQxqmWQ4dx0p/KigwfpCW0BPDWNRnbgkV9IF+zpxLAUF+fH8bmDvb1EUfEXQbjkNAJqWG6ib4/d8tkeE6AIIlfXDUReeWdU+tm7guTBvT6khsK66xQ8KUN5kFkCo11Kk8BrxZk00PsSvoS4JY4MOnm1hdmaiXAxI9UGb83w7K0Zl9fwdfGiszXHTGyHAFICYumess8+A1/cu8shXs2CkuXIIlFYHBXdTIIBAONP/Cwx64rlnhqMtFrAntSpG9Qd8JUwB7nC9VpqqdUdDP+gr9B/+L4CotOlx51lqlcpXvaar7whTH8XWlr2zc9dPu9ioj25mzLELaWltdtXw07xWzu6sVVyz

9jR7BGM6AeUy7MnTwYjDJScqpMKHN8XFsrLLCYdaUK3RJJVgfsqP9conLMFxrV8BC2RuDrKkP61x90tNz1gx73mOlwys7nckPb+7yiBS
Ql8cntHo+gR8WZ+8UmiC1r0kibnHTpzWGo1JNkfzVLzq/1PIF7FjUcmosm0SUEnm/POudG5x5B5gGit3JKn5I8SG780I3nFSTm7puiDvEO
mWSQg0v88VdqAU5b8q51ReR/E1PgAnW5+pkdydBxXCLWCWWbDflada1u8YunFkvoqJoumW+ApAEFOYT/Zrf+U/S07aZuuWu7jSY0Wa
W6RnBF13yD2bSsRGeo7mr1MEHkGh8IEOC6DK8f4dJa5ZH/hWi1+3pseXI5xB+oyJR9uGXdGTCPURPrYyas/S7dlPbvCTRcs0SYFnlQb+iK
MOhVlWu/lwsIm9XuYXnd8DamY4IEFZBAAdT/DMvv+NoP5I2Hg91xZCOAUe5HLNzV7bJRuDZDha9h9unlCwk0zNlyQufA2InsMjbLZ01uBa
//EtBGw30E4vZEPeWm08cDNqA8uBquU6QoBMEfDU8+b98VpKW+UetmXOTWIRUKP7s0wEx/ZtFTOriSnYvgsV2a5UGzDCAu6sk5zUM
80MHPPIK8PXRwOwOHw8Tp5gSM4PIXK6FjVtixjNP9NyFk+NI8MyZ8zoLkPlX/iq2P592ECC0E01XIM6zQT4TLKHU8UihkxvltszxROJVO
LRYatQN8fGSXD0CSb4lpUamP1A3qxAqtg/TPTy33KtLpVUa7SdKEgcEJ8qp4Mw0AMAqkZwtyTC0s+rDZ9FIL1rfCG0XUxJVL/ONvDgt3r
pH0ayuxMbUvDAm5WXcyWuBzN1gh7whsHO1g7bR+74bTmNufvX+/UTKHkKQOhUrkvhbz+ldz9vZdl5NAKBfc9YQ6/3PiMk8xsZjYW0wx
+oZmZ8YpT9KxMUKVdM9tQvvy0W+EuaKApUJeRI3dlEuCDB0CfYughE3ERj0XvNP2AkKLBjRkdngrbXGgbUdu4VScS69gikk8tJFfnfBW8
0BD7LCz4gAsfxRLLHbwKe/P9xRG73fzRLUHS8HE5osKVjteg0Wgpt/idHOqFm5SbCHwn8B9zGb03T5Hed3nT2Kc6po4vffOGXmUqFt
U7KQx2XLno9P8I4JWRtBcHYayDI3r8HWcoFyrFN9RvwXhZwWXiYV9IRIU8ZQZifKWGHfOC88G0m/99e2yl8asLKHrJzS83q/AbC8mG
3O3nxRWJlEjglP0Z/pdmOfa+rwKyUOD7p0w8JB/Akw2v6VFE7Ez5vDUQCGBzFYB4E7CXy6jH48e6K6KcsWLCZYlVWiW9P9KtG4yQqJ
m2wGLEvWaqWw9HGGIUOUQ4T+MiPK05vax3rBQL6/yUhgIC/GoagwmbUMxxc6/TR8RtxP1QABlksbKnsFjtziw1yKBdDR9IKoJUDu3/R
fbDNx8XrzSNYRQewJukwRDLftcM70oIxcMDQMatLd2ZrbCYVjgl4vMYfdyCTItHbZO3nEHfAf4u4tbAGLvgtrBlq5bePvSmin3hO441H3vb
xThpSeSE/ICynC4oSrgtnevQ2sRBHqMJL3GNaR99woWG2SeR88+UW743FzEvla9DGtyBhkyw7KGD70/AD3B3Xt1122zFCTOLOW1jY0
EpVXbPcc2HC7W07o4UwAOCziCXasaGY/0YInhsD5fP1+WLCcg+pxh+yQHaf9z+TfVp+DAbJK4idBmrCNOUZ4cBqV++xJgoaFVoL+YEm
Mt3OHdjKivHI4jbTsY6qml4w7TqsEOGblz7mYivJj/JWTI5rB0QjgU2C+Q44z2V8wb1nqPyKuFAshfhNiMk5zE4a13/R6pHU9MWXjxKx1ms
VjS1RO0QywKKZl1m0jW2bDSFfeMw/iwPq2fwTJ5ZZBtDYbkdxAKf1EJ56JkLadNHDq9c0n6GVSNI1frKpKx/2ox03xv8+yMmaNlfuKy1pS
aP7YUv/5d5fQH+MherVuDioLIT09oITUDdHmi5JC1lwVgI5ZV50km5FyP9dqgy/PKk0mrJlkiNbc8firhAqaBxP+3k90uBW1GcQFQ6njzBtp
iuy0yvcAC3lmyk6GJzu3A44hj8VbwF/YBClqrio+TPIVf71f0jztRa5z5L5QcyfaWTqA54g3KN3v/k6A19FOMolsiltuKYL7b4TDk8pzoDMAkV
C4/lyD01QI5ZriuLrjmiMplGOrQfQf/gfZCjDgbimje4KbgKCO7ZL9cK4pl15di6u0Qexqbwol3zQv1eTonoOnHiktljR7BhzTKVzzfujPz2Tg1an
n38ybMVFu1Gwobq9QINzHgZ+Ldx5QJpny6BT8omFGqJNWD3TSQimP9bB9aEk/6xiG0avYFn22xssARwCyh43qz5GZSLQamtSfgoRjsO
B3N5kC6+odle3osrja3ebSTQA1uQ2CxJvX0tOj8NCjvdgAEBJnwmY2P5a1a0mmsok5UgaLH6eemAHwQn4WU/cEbSXhNgcqZzuVWpkDt
K1QGeoDtydTDKa6EiDniul9wAnJEQacv9XVyUpD4RR4Pk7IsJ5pzLCNZkadl0Tz7zXayN22AehoNljQONS8mNkHuhwZf+F7ODv7BEaccjD
6y7J2RsoB8cmqjLffWnag6+SeMMwqRRbZuML8rbIN7Z+RCeOZjMw5b4rs+I55FolUVGMd/yEJPQparvzdR+Xs2+01IRyR+INwRbHmjO
1ugjhBq9DdbQI209KLydiend54LrTGU/BRQwi/AzFUm3nDyysomk4tz/UvyXTrrTqIQA0IJAN/uUk3MwGaMBOTX5iBCrAMW2DeGsWqGEUL
E30+DH3jJb0SiOrL5bCfJzItDxKzL+8wx9tDL2MiRRDQA4IKUJUQIOoznmXaRbvp7tr3vfzyUM

SAFETY WARNINGS

Pay attention to the proper use of the computer.

BEFORE STARTING

What you need to prepare:

Python 3
TensorFlow
Keras

- 1 Prepare the dataset:
[In the process of extracting protein data from UniProt](#), we removed those sequences with length less than 50 or greater than 1,280 amino acids, resulting in 17,651 [DNA-binding protein](#) sequences are selected as positive samples. At the same time, we got 50,500 non-DNA-binding protein sequences as negative [samples](#) in UniProt that are 50 to 1,280 in length. We took 500 sequences from both positive and negative samples as independent test samples, respectively. [For the remaining 17, 151 positive and 50,000 reverse samples, we randomly selected 85% of them as training sets and the remaining 15% as test sets to participate](#) in model training.
- 2 Build model:
The deep learning model is composed of four parts: coding layer, embedding layer, convolution layer and Bi-LSTM layer. The coding layer represents each [amino acid](#) as a particular number. The embedding layer translates amino acid sequences into continuous [vectors](#). The convolution layer consists of two convolutions and two maximal pooling operations. The mission of the Bi-LSTM layer [is](#) to grasp the context features of amino acid sequences. We use the Keras platform to build this model.

3 Model training:

The data is trained in the built model, and this process is carried out on the GPU. At the end of this process, we get a DNA binding protein predictor.



This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited