Atlas of Science
another view on science
http://atlasofscience.org

A novel deep learning-based method for predicting RNA-protein interactions

RNA-binding proteins (RBPs) take over 5–10% of the eukaryotic proteome and regulate the gene localization and translation. On the other hand, the mutations in RBPs have been discovered to be associated with disease risk, such as FUS and TDP-43 in amyotrophic lateral sclerosis. Thus, decoding the links between RNAs and proteins can facilitate the insights into the mechanism behind them. Identification of ncRNA interactions through experimental methods is still challenging and high-cost, which can be complemented by the use of computational models. How to accurately and automatically identify whether a RNA binds to a protein is urgently needed.

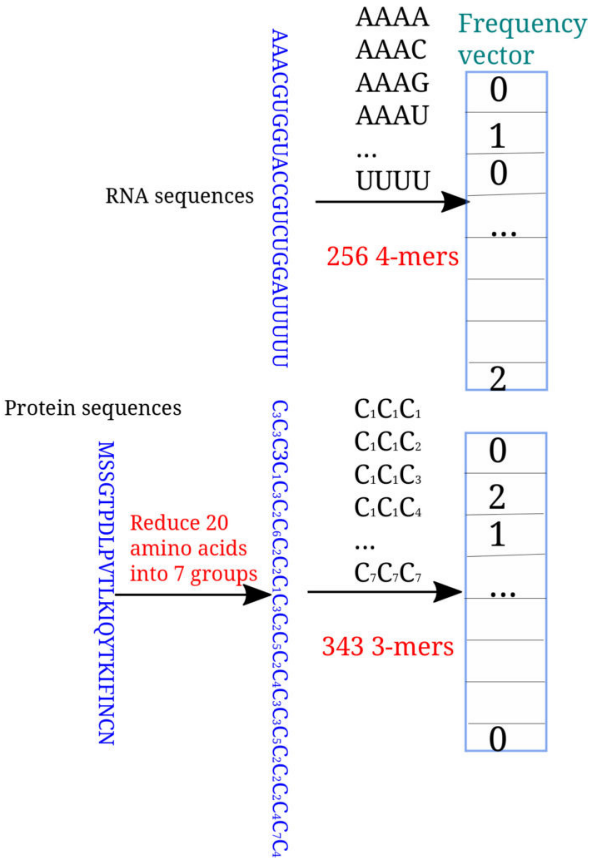

Fig. 1. Encoding RNA and protein sequences into a vector of k-mer frequency. The 20 amino acids are grouped as follows: (Ala, Gly, Val), (Ile, Leu, Phe, Pro), (Tyr, Met, Thr, Ser), (His, Asn, Gln, Tpr), (Arg, Lys), (Asp, Glu) and (Cys).

We develop a deep learning-based method, IPMiner, to automatically predict the RNA-protein interactions directly from sequences, which can be applied for any RNA and protein pairs. The new IPMiner proceeds with the following 4 steps:

In the first step of IPMiner (Fig. 1), it encodes simple k-mer sequence features both for RNA and protein sequences. For RNA sequences, we extract the frequency of 4-mers, which is the number of times a 4-mer appears in the sequence. For protein sequences, we first divide the 20 amino acids into 7 groups, then we get the frequency of 3-mers using the reduced amino acid alphabet.

In step 2, we use stacked autoencoder to further refine the presentations of raw k-mer features for proteins and RNAs, respectively (Fig. 2). Stacked autoencoder consists of multiple layer of neural networks, and each layer reconstructs original input after nonlinear transformations.

In step 3, the learned high-level features for proteins and RNAs from stacked autoencoder are concatenated, which are fed into a random forest classifier to predict whether this RNA-protein pair interacts or not. To remove the potential bias caused by a single classifier and enhance the accuracy, we also trained 2 other random forest classifiers: one is using the raw k-mer frequency features without any post-processing as the input, and the other is using the abstracted features from unsupervised stacked autoencoder without fine tuning using labeled RNA-protein pairs as the input. In total, we will have 3 random forest classifiers for different input features as a complement to each other.

In step 4, finally we integrate the outputs from these 3 different classifiers using stacked ensembling, where the outputs from the 3 different classifiers are inputted into a logistic regression to learn the weights for the 3 different classifiers. Compared to the traditional majority voting, it can automatically learn the different contributions of diverse classifiers to the final decision.

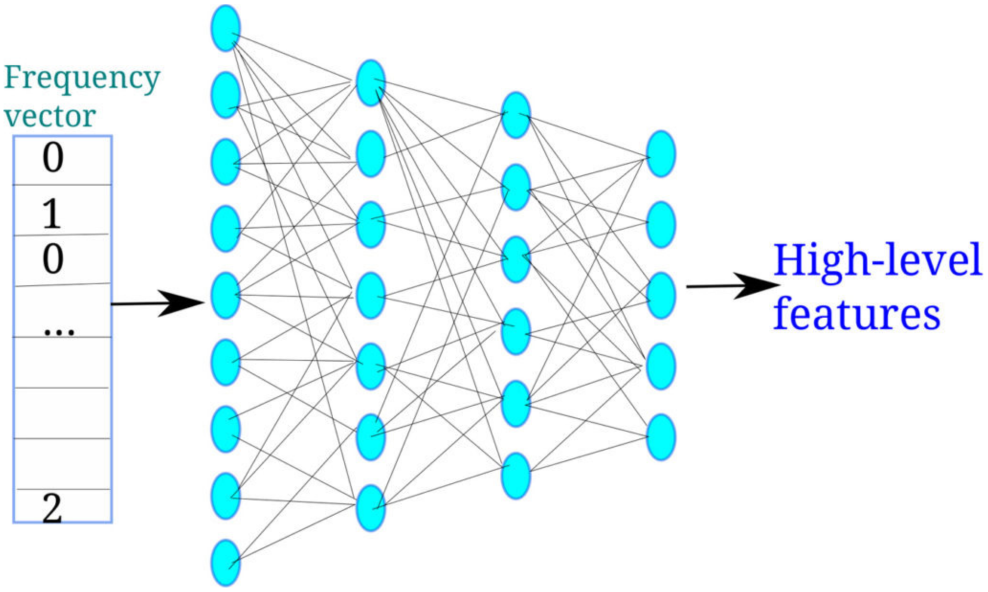

Fig. 2. Stacked autoencoder is used to further refine the presentations of raw k-mer features for proteins and RNAs, respectively. The refined features are further fed into random forest to classify RNA-protein interactions.

Due to the new IPMiner is only requiring the sequences as the input, it can be used to predict the probability of interaction for any pair of RNAs and proteins. Its efficacy has been demonstrated on multiple RNA-protein datasets. To make our IPMiner serve the academic community better, an easy-to-use standalone software has been released at http://www.csbio.sjtu.edu.cn/bioinf/IPMiner/ and https://github.com/xypan1232/IPMiner. When using this IPMiner, the users only need prepare two Fasta files for RNAs and proteins respectively, then IPMiner will automatically calculate the interaction potential between any pair of RNAs and proteins in both files.

Xiaoyong Pan [1], ***Hong-Bin Shen*** [2]
[1] *Department of medical informatics, Erasmus Medical Center, Rotterdam, The Netherlands*
[2] *Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China*

Publication

IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction.
Pan X, Fan YX, Yan J, Shen HB
BMC Genomics. 2016 Aug 9