**NPI-RGCNAE: Fast predicting ncRNA-protein interactions using the Relational Graph Convolutional Network Auto-Encoder (Supplementary File)**

***k*-mer features as side information**

To compare the influence of sequence-based features on our method, we applied the *k*-tuple frequencies as sequence-based features. Sequences of RPI2241 and RPI369 were provided in the original datasets. For the NPInter10412 dataset, sequences of proteins and ncRNAs were collected from the Uniprot [1] and NONCODE [2] databases, respectively. Sequences of the RPI7317 were extracted from the GENCODE [3] and Uniprot [1] databases. We used 3-mer frequencies of reduced amino acid residues for protein sequences and 4-mer frequencies of nucleotides for RNA sequences. Twenty types of conventional amino acids are divided into seven groups as {*A*,*G*,*V*}, {*I*,*L*,*F*,*P*}, {*Y*,*M*,*T*,*S*}, {*H*,*N*,*Q*,*W*}, {*R*,*K*}, {*D*,*E*} and {*C*}, according to their physicochemical properties [4]. One letter from each group was used to substitute the other letters in the same group. The 3-mer frequencies of the seven letters can be represented as a 343-D vector. Similarly, the 4-mer frequencies of nucleotides on an RNA sequence can be represented as a 256-D vector.

Let **X***p* be the *k*-tuple frequency matrix of protein sequences, **X***r* the *k*-tuple frequency matrix of RNA sequences. The final node feature matrices **Z***p* and **Z***r* can be computed as follows:

,

,

where

,

,

*ψ*(.,.) the column-wise concatenation operator on matrices, *σ*(.) the Rectified Linear Unit (ReLU) activation function, **W***p,*1, **W***p*,2, **W***r*,1 and **W***r*,2 four fully connected layers, and **b***p* and **b***r* two bias vectors.

For convenience in implementation, ncRNA and protein *k*-tuple frequency features have been aligned to the same dimensions in the way of complementing zeros, before they are entered into the fully connected layers.

The final node features matrices **Z***p* and **Z***r* merged node embeddings and sequence-based features. They were input into the decoder for prediction. The performances of our method with sequence-based features and without sequence-based features will be compared.

[1] The UniProt Consortium, “UniProt: a worldwide hub of protein knowledge,” *Nucleic Acids Research*, vol. 47, no. D1, pp. D506–D515, Jan. 2019, doi: 10.1093/nar/gky1049.

[2] Y. Zhao *et al.*, “NONCODE 2016: an informative and valuable data source of long non-coding RNAs,” *Nucleic Acids Res*, vol. 44, no. D1, pp. D203–D208, Jan. 2016, doi: 10.1093/nar/gkv1252.

[3] J. Harrow *et al.*, “GENCODE: The reference human genome annotation for The ENCODE Project,” *Genome Research*, vol. 22, no. 9, pp. 1760–1774, Sep. 2012, doi: 10.1101/gr.135350.111.

[4] J. Shen *et al.*, “Predicting protein-protein interactions based only on sequences information,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, Mar. 2007, doi: 10.1073/pnas.0607879104.