YTLR: extracting yeast transcription factor-gene associations from the literature using automated literature readers

Tzu-Hsien Yang*, Chung-Yu Wang, Hsiu-Chun Tsai, Ya-Chiao Yang, and Cheng-Tse Liu

Supplementary Materials

1 Learning curves for TF-gene sentence understanding networks in YTLR Phase II

In YTLR Phase II, the extracted sentence descriptions for TF-gene pairs are first comprehended by deep learning auto readers. The auto readers were trained using the TF-gene sentence description datasets. Notice that by summarizing the classified sentence descriptions for a TF-gene pair using the deep sentence description understanding network, better performance can be achieved in recognizing the TF-gene binding/regulatory associations.

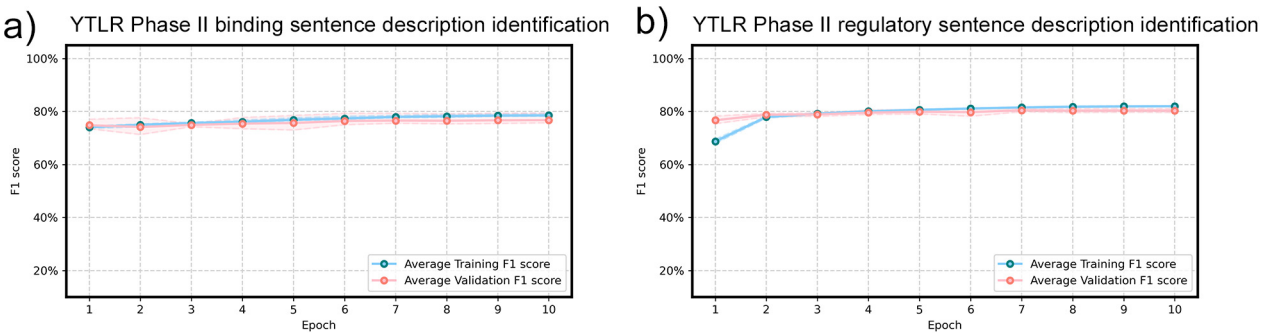

Figure S1: (a) The learning curves of the TF-gene binding information understanding network. (b) The learning curves of the TF-gene regulatory information understanding network.

2 Retraining of the CNN and LSTM models for identifying literature with TFB/TFR evidence

We retrained the CNN and LSTM models proposed by Burns *et al.* using the BioBERT word embeddings. The hyperparameters of TFB/TFR CNN models were chosen by five-fold cross-validation: (1) learning rate schedule: learning rate = 2.45e-4/2.4e-4 with exponential decay (decay rate = 0.85); (2) Optimization method: Adam; (3) number of epochs: 50; (4) neuron initialization: Xavier initialization; (5) Batch training size: 64. (6) dropout = 0.6. The hyperparameters of TFB/TFR LSTM models were also selected by five-fold cross-validation: (1) learning rate schedule: linear warm-up followed by cosine decay (max learning rate = 2e-5/3e-5); (2) Optimization method: Adam; (3) number of epochs: 50; (4) neuron initialization: Xavier initialization; (5) Batch training size: 64. (6) dropout = 0.5/0.6. The CNN and LSTM models were trained using NVIDIA RTX 2080Ti GPUs.

The learning curves for retraining the CNN and the LSTM models are shown in Figure S2 and Figure S3, respectively.

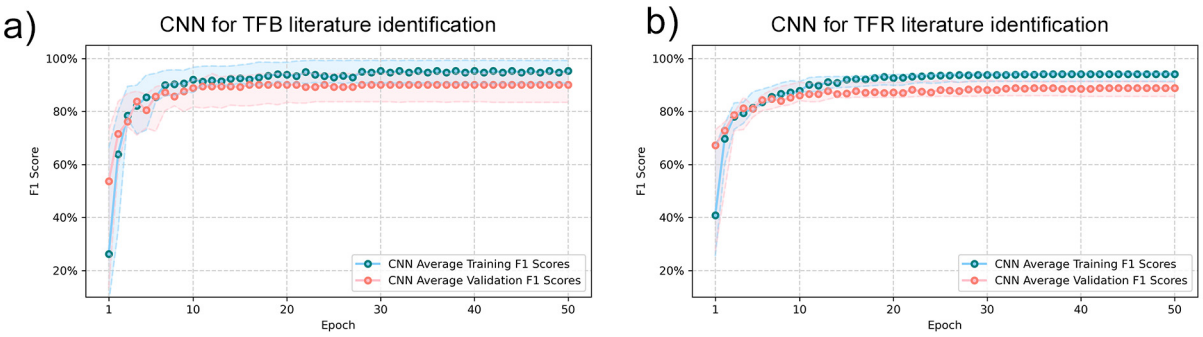

Figure S2: The learning curves for obtaining the CNN model proposed by Burns *et al.* [1]. (a) On the TFB-tagged article training/validation set. (b) On the TFR-tagged article training/validation set.

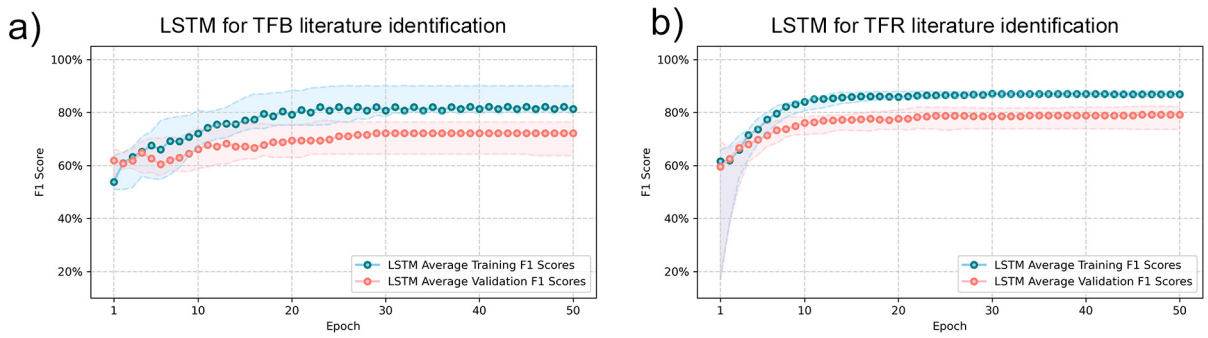

Figure S3: The learning curves for obtaining the LSTM model proposed by Burns *et al.* [1]. (a) On the TFB-tagged article training/validation set. (b) On the TFR-tagged article training/validation set.

References

[1] G. A. Burns, X. Li, and N. Peng, "Building deep learning models for evidence classification from the open access biomedical literature," *Database*, vol. 2019, 2019.