Supplementary file for "Pre-training Graph Neural Networks for Link Prediction in Biomedical Networks"

GO graph

To construct GO graph, following Peng et al. (2016), we first calculate semantic similarity matrix $M_g \in \mathbb{R}^{N_p \times N_p}$ for proteins based on their GO terms. N_p means the number of proteins. After that, we implement random walk with restart (RWR) algorithm on the semantic similarity matrix M_g to extract more important interactive pairs by selecting the top-t neighbors as interaction pairs for a given protein. Formally, RWR can be formulated as follows:

$$p_i^{t+1} = (1-\gamma)Mp_i^t + \gamma x_i,$$

where p_i^t represents the probabilities of reaching other nodes at the time t starting from the i-th node. M is transition probability matrix obtained by normalizing the similarity matrix M_g and γ is the restart probability (empirically, we set γ as 0.9). $x_i \in \mathbb{R}^{N_p}$ denotes original probability vector of node i and $x_{ij} = 1$ if $j = i$, otherwise 0. For a given node i, we prioritize the probability scores of all neighbors and select the top-t neighbors as its interactive pairs to construct GO graph $\mathcal{G}_g$.

Baselines

In this work, we validate the performance of our model via two tasks, i.e., SL prediction and DTI prediction. We introduce eight state-of-the-art baseline methods for SL prediction as follows:

- SL^2MF (Liu et al., 2019) is a logic matrix factorization-based method for SL prediction. It extracts gene features from GO (i.e., BP) and PPI network.
- GRSMF (Huang et al., 2019) is a graph regularized self-representative matrix factorization algorithm for SL prediction. It leverages gene functional similarity calculated based on GO (i.e., BP) as gene features.
- GCATSL (Long et al., 2021) is a novel graph attention network-based model developed for SL prediction. It constructs three different categories of feature graphs. The first two feature graphs are built from GO (i.e., BP and CC). The third feature is PPI network.
- SLMGAE (Hao et al., 2021) is a multi-view graph auto-encoder based method to predict SL pairs. It takes SL graph as main view and constructs three support views by extracting gene feature matrices from GO (i.e., BP and CC) and PPI network. For each view, SL adjacency matrix is utilized as node features and shared by different views.
- DDGCN (Cai et al., 2020) is a dual-dropout graph convolutional network model for SL prediction. It takes SL adjacency matrix as node features.

Meanwhile, we introduce five state-of-the-art deep learning methods for DTI prediction as follows:

- NeoDTI (Wan et al., 2019) is an end-to-end deep learning model to predict drug-target interactions (DTIs). It integrates multiple sources of biological data to construct a heterogeneous network for DTI prediction, such as drug structure similarity, drug-side-effect associations, drug/protein-disease associations, etc.
- DeepDTA (Öztürk et al., 2018) is a deep learning model that uses drug structures and proteins sequences to predict drug-target binding affinity.
- MolTrans (Huang et al., 2021) is a Transformer-based framework for DTI prediction. It encodes drug structures and protein sequences to learn features for drugs and proteins, respectively.
- DeepConv-DTI (Lee et al., 2019) is a convolutional neural network-based model for DTI prediction. It uses protein sequences and drug fingerprints to construct input features for proteins and drugs, respectively.
- GraphDTA (Nguyen et al., 2021) is a graph neural network based method for drug-target binding affinity prediction. It leverages drug structures and protein sequences to learn representations for drugs and proteins, respectively.

Reference

Peng, J. et al. (2016). Intego2: a web tool for measuring and visualizing gene semantic similarities using gene ontology. BMC genomics, 17(5), 553–560.

Liu, Y. et al. (2019). SL^2MF: Predicting synthetic lethality in human cancers via logistic matrix factorization. IEEE/ACM transactions on computational biology and bioinformatics, 17(3), 748–757.

Huang, J. et al. (2019). Predicting synthetic lethal interactions in human cancers using graph regularized self-representative matrix factorization. BMC bioinformatics, 20(19), 1–8.

Long,Y. et al. (2021). Graph contextualized attention network for predicting synthetic lethality in human cancers. Bioinformatics, 37(16), 2432–2440.

Hao, Z. et al. (2021). Prediction of synthetic lethal interactions in human cancers using multi-view graph auto-encoder. IEEE Journal of Biomedical and Health Informatics, 25(10), 4041–4051.

Cai, R. et al. (2020). Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers. Bioinformatics, 36(16), 4458–4465.

Wan, F. et al. (2019). Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions. Bioinformatics, 35(1), 104–111.

Öztürk, H. et al. (2018). Deepdta: deep drug–target binding affinity prediction. Bioinformatics, 34(17), i821–i829.

Huang, K. et al. (2021). Moltrans: Molecular interaction transformer for drug–target interaction prediction. Bioinformatics, 37(6), 830–836.

Lee, I. et al. (2019). Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS computational biology, 15(6), e1007129.

Nguyen, T. et al. (2021). Graphdta: Predicting drug–target binding affinity with graph neural networks. Bioinformatics, 37(8), 1140–1147.