Prediction of Protein-Protein Interactions on the Human and Rice Interactomes

Nicolás Antonio López Rozo

August 27, 2020

Abstract

Network analysis for predict unmapped protein-protein interactions (PPI) suggest that the higher the number of paths of length 3 (L3) between two proteins, the more likely they are to interact. This paper extends previous work based on the L3 principle by taking into account the representation learning of node features of the PPI network. In particular, we train an XGBoost model using L3 and handcrafted features, as well as embeddings from `Node2Vec`. Our main result shows that while L3 is an important feature for predicting links, best performance is achieved when combined with edge embeddings. The proposed approached is evaluated for the human and rice interactomes.

1 Introduction

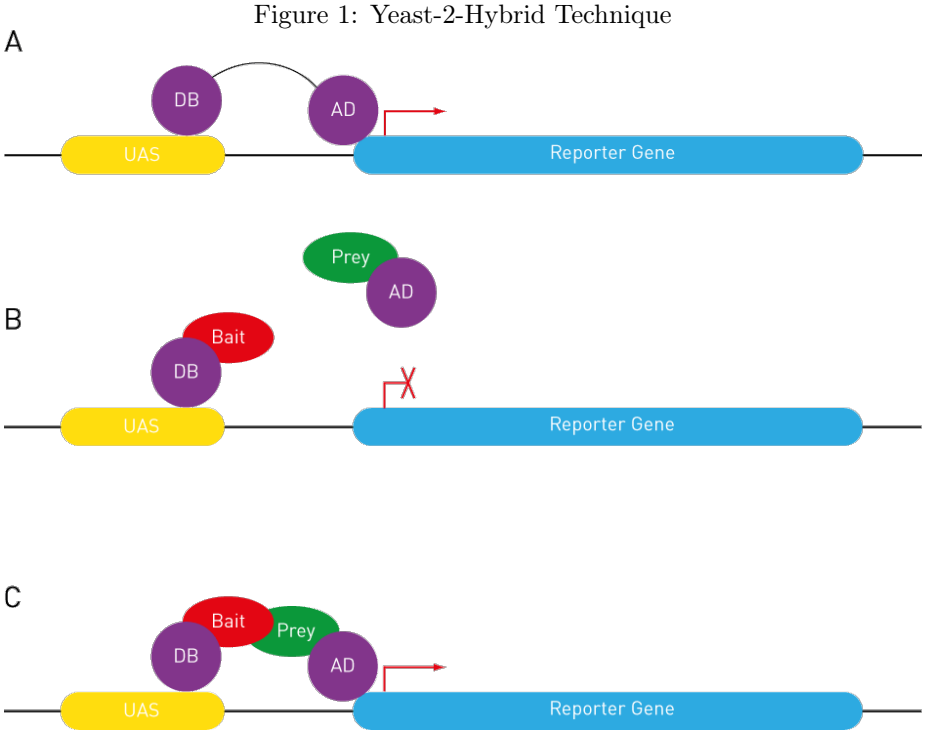
Proteins are the key actors of biological functionality inside the cell. As they carry out a variety of tasks, they do not work as single substances but rather as a part of dynamic networks of protein-protein interactions (PPIs) [2]. PPIs have a key role in a variety of biological processes such as signal transduction, homeostasis control, stress responses, plant defense and organ formation. At the molecular level, PPIs play an essential role in many physiological and developmental processes, including protein phosphorylation, transcriptional co-factor recruitment, transporter activation [4].

Prediction of potentially relevant, yet unexplored PPIs is a current research topic on bioinformatics and therefore, several authors have proposed different methods for extrapolating information from the existing PPI networks. As presented by Kovacs et al (2019), information related to counting paths of length 3 (L3) seems to outperform state-of-the-art methods when it comes to predict interactions among proteins for a variety of model organisms such as yeast (*S. cerevisiae*), Arabidopsis (*A. thaliana*), worm (*C. elegans*), fly (*D. melanogaster*), fission yeast (*S. pombe*) and mouse (*M. musculus*), as well as on the human interactome [1].

The most common way to validate PPI's is the *Yeast-Two-Hybrid* technique (also known as *two-hybrid screening* or *Y2H*), which is based on the expression of

a specific reporter gene that activates by the binding of a DNA-binding Domain (DB) and an Activation Domain (AD) of a Transcription Factor that binds to an Upstream Activation Sequence (UAS). For the Y2H technique, a protein is fused to the DB domain (known as *bait*) and another one to the AD (known as *prey*). If the proteins do not interact, then the reporter gene is not expressed. Otherwise, the reporter gene expression is activated by the activation domain.

Figure 1: Yeast-2-Hybrid Technique



Having several of these Y2H results allows scientists to establish a PPI network, where all known interactions for each protein are represented. Several algorithms are proposed over these networks in order to predict unknown interactions. In this report, three prediction methods are presented and their results are shown: Common Neighbors (**CN**), which uses the length-2 path count; raw count of paths of length 3 (**A3**); and degree-normalized length-3 paths score (**L3**).

Focus of the present study is to evaluate different methods for predicting protein-protein interactions (PPIs) using the existing knowledge of the network, which is an undirected graph. The traditional way is usually based on social networks analysis, more specifically on the Triadic Closure Principle (TCP), that states that the more common shared friends that two people have, the more likely that they know each other. As shown by previous studies, the mentioned approach fails because it does not consider the structural and chemical

properties of the proteins [1].

For achieving the described results, human and rice PPI networks are used and compared using state-of-the-art methods, as well as the proposed ones (CN, A3, L3). In the case of the human network, human interactome (*HI-II-14*), as well as a curated version of it (*HI-TESTED*) were used. A massive experimental assay was carried on and its results were consolidated and used to build a validation network (*HI-III*).

2 Materials and Methods

2.1 Data Availability

Human interactome data and base source code were downloaded from the repository of the length-3 degree normalized paths methodology [1]: the dataset *HI-II-14* and *HI-TESTED* are used for prediction and the dataset *HI-III* is used for validation.

Rice interactome information was downloaded from the STRING database [3], corresponding to the *Oryza sativa* subspecies. The downloaded file was *4530.protein.links.detailed.v11.0.txt.* and contains more than 8 million PPIs from several resources. For the purpose of this study and based on previous work, only PPIs with evidence from curated databases were used (i.e. rows where the column *databases* has a value greater than zero), resulting in a network with 5025 nodes and 164420 edges.

2.2 Code Implementation

Previous code implementation was adapted from `C++` to `Python` (V3.6), in order to unify the algorithms into one single script. For the purpose of algorithmic validation, the three methods were implemented from scratch with basic functionalities and data structures of the `Python` language.

2.3 Data Preprocessing

Information for the human interactome was used as-is, which corresponds to networks of 4298, 3727 and 5604 proteins and 13868, 9433 and 23322 interactions.

For the rice interactome, an additional preprocessing was performed. The filtered network for rice consists of 5025 proteins (nodes) and 164420 interactions (edges) distributed among 178 connected components. The connected component with the greatest number of edges was selected in this case. The extracted connected component consists of $n = 4390$ nodes and $m = 163319$ edges, which corresponds to 99.33 of filtered edges. Further investigation is applied to this network, which is very similar in number of nodes to the curated information on the human interactome, although rice network is much more connected.

2.4 Edge Prediction

For the interaction prediction for each network, the algorithms described below were used. It is important to keep in mind how the protein-protein interaction (PPI) network $G = (V, E)$ is conceptualized: each node ($v_i \in V$) represents a protein and each undirected edge ($e_b = \{v_i, v_j\}$, $e_b \in E$) represents and interaction among proteins v_i and v_j.

Common Neighbors (CN) This method is based on the Triadic Closure Principle: "the more common friends two individuals have, the more likely that they know each other". For the implementation of this method, A^2 matrix is calculated, being A the adjacency matrix of the network.

Length-3 Paths (A3) This is the simplest implementation of the proposed insight of "if my friends and your friends interact, then we might interact too". The calculating is carried on with A^3, i.e, the third power of the adjacency matrix.

Degree-normalized L-3 Score (L3) The previous approach might overestimate the importance of some edges due to intermediate hubs which add many shortcuts in the graph. To address that issue, a degree normalization for the path $X \to U \to V \to Y$ is applied by considering the degree k of the intermediate nodes U and V, as follows.

$$p_{XY} = \sum_{U,V} \frac{A_{XU} \cdot A_{UV} \cdot A_{VY}}{\sqrt{k_U \cdot k_V}}$$

where A_{ij} represents the value of the adjacency matrix for nodes i and j: 1 if the edge $\{i, j\}$ exists, 0 otherwise.

2.5 Sampling Procedure

For each network of protein interactions, the following procedure was performed 10 times in order to address the stochastic nature of the process and have a consensus:

- A percentage of interactions is removed at random from the network (20%).
- The same amount of removed interactions are then predicted using the main methods for prediction mentioned by Kovacs et al (2019): Common Neighbors (**A2**), raw path count of paths of length 3 (**A3**) and the Length-3 degree-normalized score (**L3**).
- A test dataset is created as follows: all removed edges are included (as observed positives for the ML algorithm) and from the predicted edges of **A2**, **A3** and **L3** that don't lie in the previous classification (observed negatives), a random subset is chosen such that the dataset is balanced, that is, the amount of observed positive labels is equal to the observed negative labels.

- Once the dataset is ready, it is randomly partitioned: 80% is used for `XGBoost` model training and 20% is used for validation. It is important to have in mind that balanced distribution of the positive and negative labels in the datasets was satisfied.

2.6 Feature Extraction with `Node2Vec`

The `Node2Vec` module was used for extracting features of the rice interactome graph. The parameters and considerations for the model were:

- All paths in the random walks are equally likely (`p=1, q=1`)
- Use a modest number of dimensions and threads for calculation (`dimensions=16, workers=4`)
- Since length-3 paths are the defining property in this study, there is no necessity for longer walks. However, it is important to try out many possible redundant routes and to consider a window of at least 4 (`walk_length=5, num_walks=300, window=5`)
- Other standard parameters were left with default values (`min_count=1, batch_words=4`)
- Edge embeddings were calculated using a geometric ratio of the node embeddings (`HadamardEmbedder`)

2.7 Handcrafted Feature

Due to the poor results of the *raw* Length-3 counting (**A3**), a different approach for this information was carried out in the present study: As it still gives a lot of information that might be useful for a predictive routine, this counting was normalized (dividing by the greatest counting in the **A3** top predictions) and then used as a feature for the Machine Learning algorithm. For completeness, also **CN** and **L3** information was used as a possible feature. Finally, the case were no handcrafted feature was also considered, that is, only the features extracted from the structure of the network.

2.8 Feature to Predict: Existence

The feature to predict corresponds to the possible existence (*True/False*) of a link based on the existing information of the network, using the network itself in a random sub_exploration (`Node2Vec`) as well as in a structured search (A3). This property is evaluated by taking out a fraction of the edges and then trying to predict for a given set of possible edges if they have a high probability to belong to the original network.

2.9 Machine Learning Algorithm

The Extreme Gradient Boosting implementation of gradient boosted trees is applied in this study to evaluate the existence of an edge. Gradient boosted trees are usually used for supervised learning problems, where the training data X_i has multiple features and pretends to explain (or predict) a target variable Y_i. The corresponding implementation applied for this study is `XGBoost`, available publicly.

The selected parameters for the model were: `max_depth=3`, `colsample_bytree=0.6` and `eval_metric=''auc''`.

2.10 Result Validation

As mentioned before, 80% of the final dataset was randomly selected and used for training, while the remaining 20% was used for validation. The whole training-validation procedure was applied 10 times.

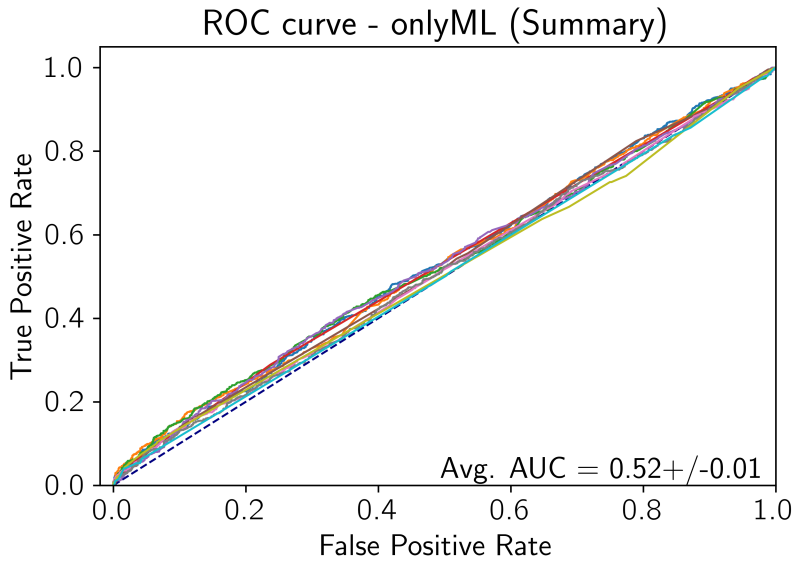
The chosen metric for validation was the Area under the Curve (**AUC**) of the Receiver Operating Characteristic (**ROC**). This curve corresponds to plot the sensitivity (probability of predicting a real positive as positive) against 1-specificity (probability of predicting a real negative as positive). It is worth to remind that AUC values move in the range $[0, 1]$, where 1 is a perfect prediction and 0.5 corresponds to a random guess. Normally, values over 0.8 of AUC are considered good.

3 Results and Discussion

3.1 Rice Interactome

For the rice interactome, the different model-features combinations were trained and validated. After executing the mentioned routines, the results are shown in Figure 2. First, one should have a baseline of comparison, which in this case corresponds to `Node2Vec` without any additional feature included. The plot below shows those results, and one can see that its mean performance using the AUC metric is 0.52, and that the results among the 10 repetitions are consistent. This results mean that the model using only the default features perform barely as good as a random choice of the labels.

Figure 2: Summary ROC curves for `Node2Vec` model alone



4 (PENDING FROM HERE ON)

As it can be inferred from the plots, L3-based predictions outperform their A^2 counterparts. Results also show that L3-score and A^3predictions follow a very similar trend.

When analyzing the robustness of each of the networks, the following values for the Weighted Spectral Distribution were found. For a robustness reference, the Erdos-Renyi model was used to generate a random network with the same number of nodes and edges and on those random networks, the WSD was measured.

Table 1: Validation of Weighted Spectral Distribution

Network	Network WSD	Erdos-Renyi WSD
HI-II-14	393.4939	198.7706
HI-TESTED	423.3902	276.1065
HI-III (VALID.)	373.9369	153.5329

The table shows that the three networks used in this report are robust, because the are significantly more robust than a network with the same density generated randomly.

5 Conclusions

Taking into account the different results validated in this report, one can conclude that length-3 path methodologies might work better on protein-protein interactions than its traditional length-2 (TCP based) counterparts. On the other hand, it can be seen that degree-normalization has little effect on the predictions, i.e., non-normalized A^3 matrix predictions are still a good methodology for edge prediction on PPI networks.

Previous result comes as no surprise when the biological basis of protein interactions is considered: It is necessary that protein A and protein B have complementary structures in order to interact, and when classical paths of length 2 are used, the predicted protein interactions usually have the same structures, and not complementary ones.

For the analyzed networks, it is also important to highlight their robustness when compared to a random network with the same density: if weighted spectral distribution (WSD) is used, then the used networks are on average twice as robust as their random network versions.

References

[1] István A. Kovács, Katja Luck, Kerstin Spirohn, Yang Wang, Carl Pollis, Sadie Schlabach, Wenting Bian, Dae-Kyum Kim, Nishka Kishore, Tong Hao, Michael A. Calderwood, Marc Vidal, and Albert-László Barabási. Network-based prediction of protein interactions. *Nature Communications*, 10(1), mar 2019.

[2] Jer-Sheng Lin and Erh-Min Lai. *Protein–Protein Interactions: Co-Immunoprecipitation*, pages 211–219. Springer New York, New York, NY, 2017.

[3] Damian Szklarczyk, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T. Doncheva, John H. Morris, Peer Bork, Lars J. Jensen, and Christian von Mering. String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(30476243):D607–D613, January 2019.

[4] Yixiang Zhang, Peng Gao, and Joshua Yuan. Plant protein-protein interaction network and interactome. *Current Genomics*, 11(1):40–46, mar 2010.