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

Nicolï¿œs Antonio Lï¿œpez Rozo

October 9, 2020

Abstract

Past work on applying network analyses to predict unmapped protein-protein interactions (PPI) suggests 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 results based on the L3 principle by taking into account the representation learning of node features of the PPI network. In particular, the proposed `XGBoost` model uses L3 and other handcrafted features, as well learned features from `Node2Vec` embeddings. Our main result shows that while L3 is an important feature for predicting links, better performance is achieved when the measure is combined with edge embeddings. Our approach is evaluated for the human and rice interactomes.

1 Introduction

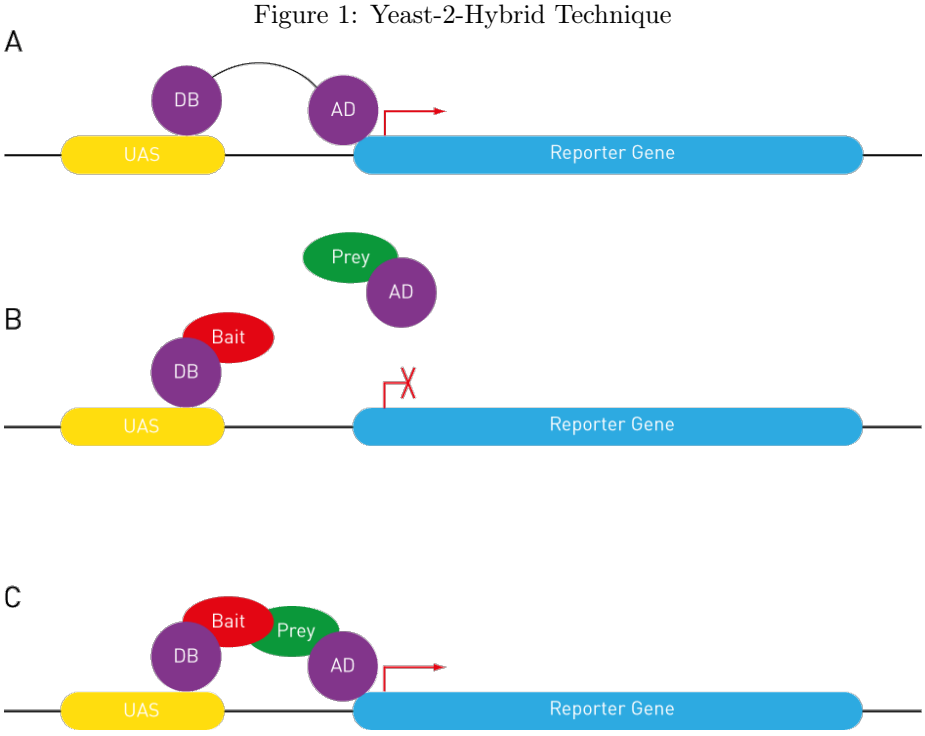
Proteins are key actors of biological processes inside cells. They function as are part of dynamic networks of protein-protein interactions (PPI) rather than carrying out a variety of tasks as isolated agents [**?**]. At the cellular level, PPI networks play a key role in a number of interdependent mechanisms, including signal transduction, homeostasis control, stress responses, plant defense and organ formation. At the molecular level, PPI networks also play an essential role in physiological and developmental processes, including protein phosphorylation, transcriptional co-factor recruitment and transporter activation [**?**].

Several authors have introduced different methods for extrapolating information from PPI networks. Kovacs et al. (2019) proposes an approach which relies on counting paths of length 3 (L3) to predict interactions among proteins for a variety of model organisms. The proposed approach outperforms previous methods for PPI networks of yeast (*S. cerevisiae*), Arabidopsis (*A. thaliana*), worm (*C. elegans*), fly (*D. melanogaster*), fission yeast (*S. pombe*) and mouse (*M. musculus*), as well as for the human interactome [**?**].

A common way to uncover (or validate) PPIs is the *yeast-two-hybrid* or *Y2H* technique (also known as *two-hybrid screening*). Y2H is based on the expression of a so called reporter gene that is activated by the binding of a DNA-binding

domain and an activation domain of a transcription factor. This transcription factor then binds to an upstream activation sequence. For the Y2H technique, a protein is fused to the DNA-binding domain (known as *bait*) and another protein to the activation domain (known as *prey*). Only when the proteins interact, the reporter gene is expressed. Otherwise, the reporter gene expression is activated by the activation domain.

Figure 1: Yeast-2-Hybrid Technique



PPI networks are constructed based on the outcome of numerous Y2H experiments in which known interactions for each protein are represented. Several algorithms are proposed over these networks in order to predict hidden interactions. This paper evaluates PPI predictions based on three measures: the count of paths of length 2 or Common Neighbors (**CN**); the raw count of paths of length 3 (**A3**); and the degree-normalized count of paths of length 3 (**L3**).

Our focus is to evaluate different methods for predicting PPIs using the existing knowledge of the network of interactions, which is represented as an undirected graph.

Human and rice PPI networks are used to compare state-of-the-art methods , as well as the proposed approach (CN, A3, L3). In the case of the human network, the human interactome dataset denoted by *HI-II-14* and its curated version (*HI-TESTED*) were used for evaluation. Results on experimental assays are consolidated to build a validation network denoted by *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 [**?**]: the datasets *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 [**?**], corresponding to the *Oryza sativa* subspecies. The file labelled *4530.protein.links.detailed.v11.0.txt.* contains more than 8 million PPIs from several sources. For the purpose of this study, only PPIs with evidence from curated databases are considered (i.e. rows where the column *databases* has a value greater than zero). The resulting network consists of 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, respectively.

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.

It is worth mentioning that some exploration experiments were carried out with interactions removal percentages of 2, 5, 10 and 20%, as well as train-test partitions of 15-85, 80-20 and 75-25, and the explained parameter set was selected

because it either represented a marginal gain (results not shown) or because it was a common parameter selection in related literature.

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.

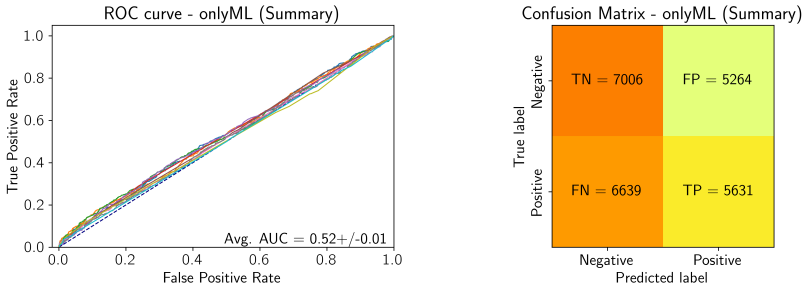
The chosen metric for validation was the Area under the Curve (**AUC**) of the Receiver Operating Characteristic (**ROC**). This curve corresponds to plot the sensitivity (probability of predicting a real positive as positive) against 1-specificity (probability of predicting a real negative as positive). 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 **??**. 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. This result can also be assessed when looking at the confusion matrix, where a precision of 0.5168 and a recall of 0.4589 can be derived.

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



Next experiment carried out was to add each prediction method score as a prediction feature, that is, use the score as the eleventh input feature for the ML algorithm. Results for the Common Neighbors method (**CN**), which uses paths of length 2, are shown in Figure **??**. It can be observed again that all 10 experiments have small variability among them and perform significantly better than the baseline, with an area under the ROC curve of 0.90, with small standard deviation. When looking at the confusion matrix for this model, turns out that most of the guesses are true positives and true negatives, resulting in a precision of 0.93 and a recall of 0.74. Same analyses are done with the count of paths of length 3 (**A3**) and with the degree-normalized length-3 score (**L3**) and results are presented in figures **??** and **??**, respectively, resulting also in area under the curve of 0.9 for both cases.

Figure 3: Results for `Node2Vec` model with CN feature

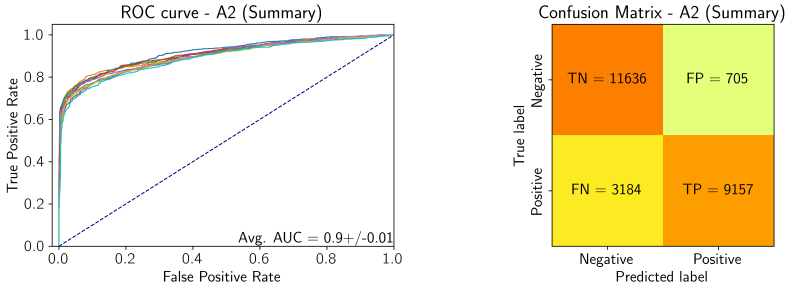


Figure 4: Results fir `Node2Vec` model with A3 feature

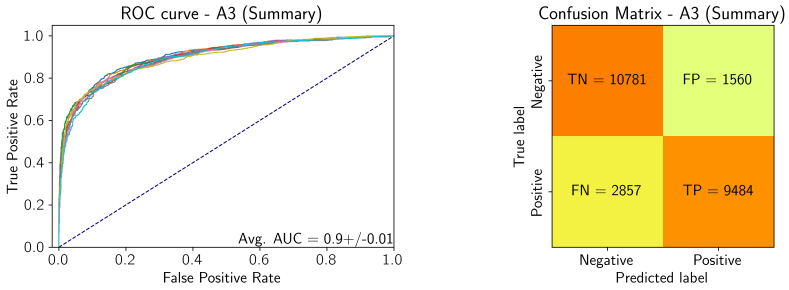
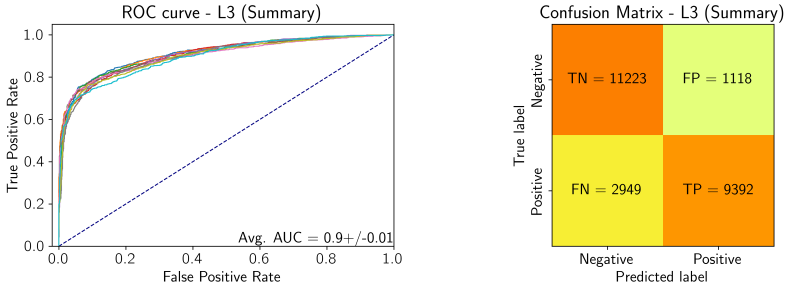


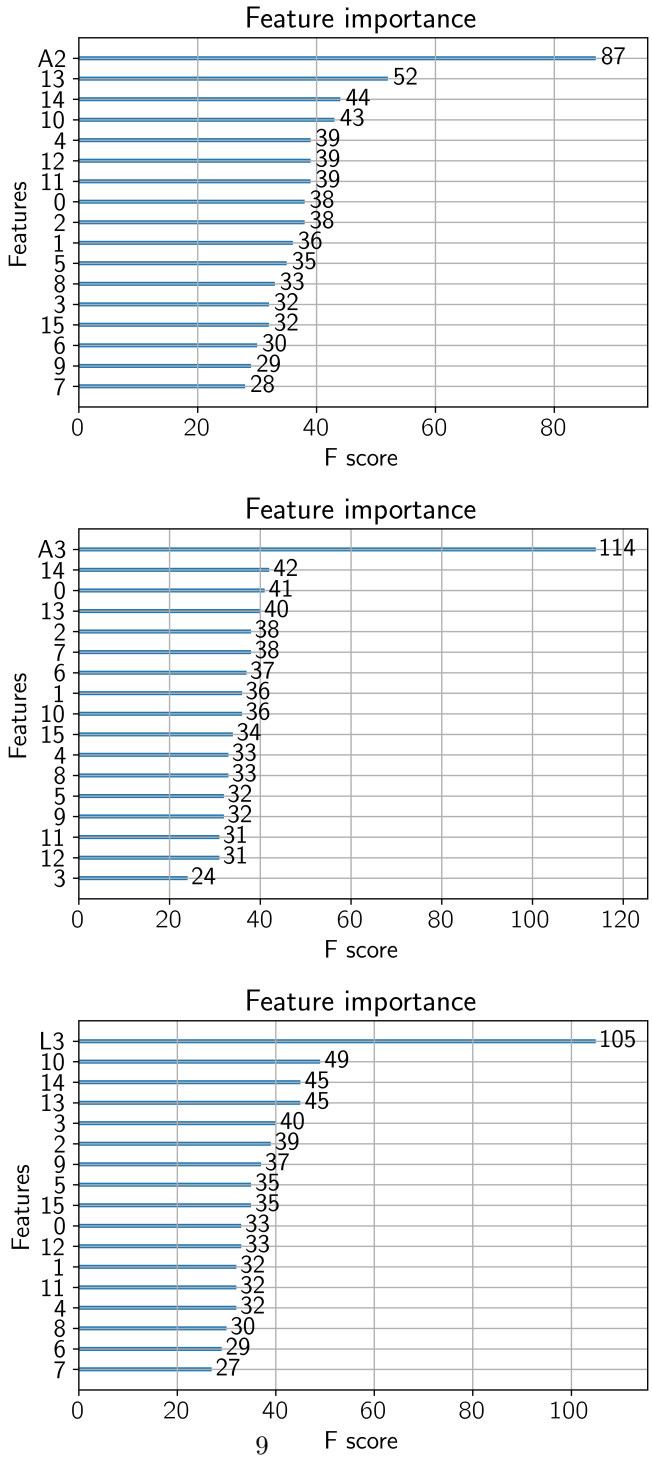
Figure 5: Results for `Node2Vec` model with L3 feature



The precision when using the count of paths of length 3 (A3) is 0.86 while the recall is 0.77. It is interesting to observe that precision decreased 7.5% when compared to the previous case, but recall increased 3.6%. A similar situation is obtained for the model with the normalized score (L3), whose precision decreased to 0.89 (-3.8%) and recall increased to 0.76 ($+2.6\%$). It can be seen also that the ROC curve rapidly achieves higher values of true positive rate in the A2-featured model when compared to the other models.

Finally, an relevant evaluation on the models that include a handcrafted feature is necessary: How important is the appended feature for the model result? The answer can be observed by the feature importance plots for CN, A3 and L3 in Figure **??**. Each plot resembles the number of times that each feature creates a bifurcation in the underlying decision trees that `XGBoost` uses. The more bifurcations, the higher the importance a feature has on the model itself. All models have the appended feature as the most important by a large margin.

Figure 6: Importance for `Node2Vec` models: with L3 feature



3.2 Human Interactome

(PENDING FROM HERE)

Figure 7: Methods Comparison for *HI-II-14*

Figure 8: Methods Comparison for *HI-TESTED*

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.

4 Conclusions

Taking into account the different results validated in this paper, 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.