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

Nicolás Antonio López Rozo

August 3, 2020

Abstract

Previous Network-based efforts to predict unmapped protein-protein interactions (PPI's) suggest that proteins with multiple paths of length 3 (L3) are more likely to be connected. This paper extends this so-called L3 principle by taking into account feature extraction and using `XGBoost` techniques for prediction. In particular, we train the model using hand-crafted features as well as features learned from embeddings using `Node2Vec`. Our main result shows that while L3 remains an important principle for predicting links, the approach is outperform by using embedded features. The mentioned approaches are compared using the human and the rice interactomes.

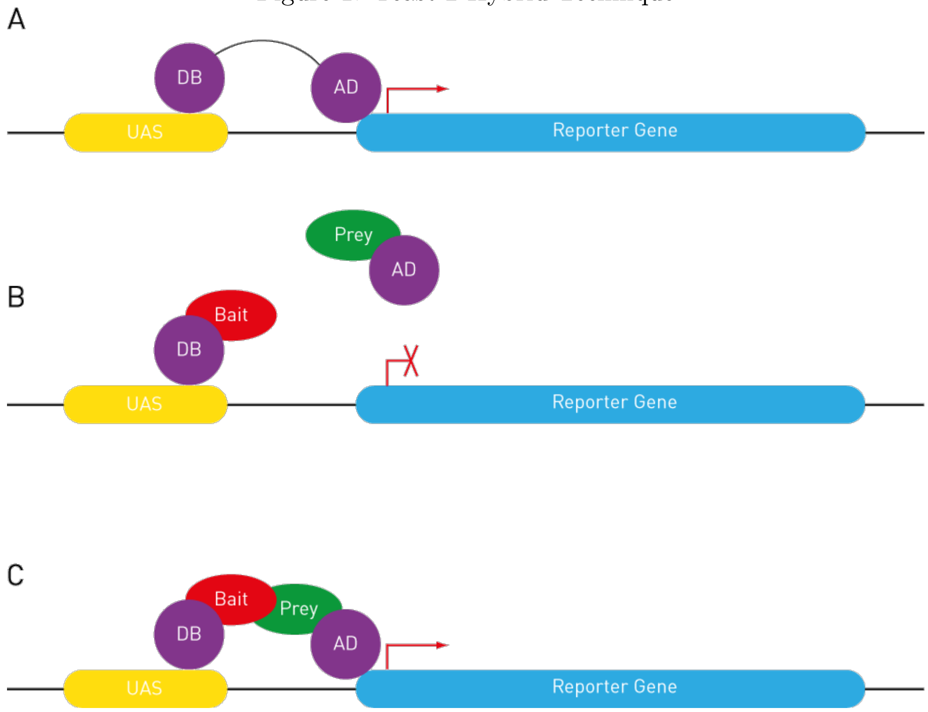
1 Introduction

Proteins, as a fundamental constituents of any living being and their "building blocks" in terms of biological functionality, interact among themselves and with other molecules according to their amino acid sequences and specific three-dimensional structure and folding. These interactions also depend on many physical and chemical properties of each macro-molecule and on the cellular organelle where they operate, such as pH and temperature. For this reason, experimental ways to reproduce all of the necessary conditions to validate if an interaction actually occurs within the cell are generally expensive.

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. 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.

Having several of these Y2H results allows scientists to establish a PPI network, where all known interactions for each protein are represented. Several

Figure 1: Yeast-2-Hybrid Technique

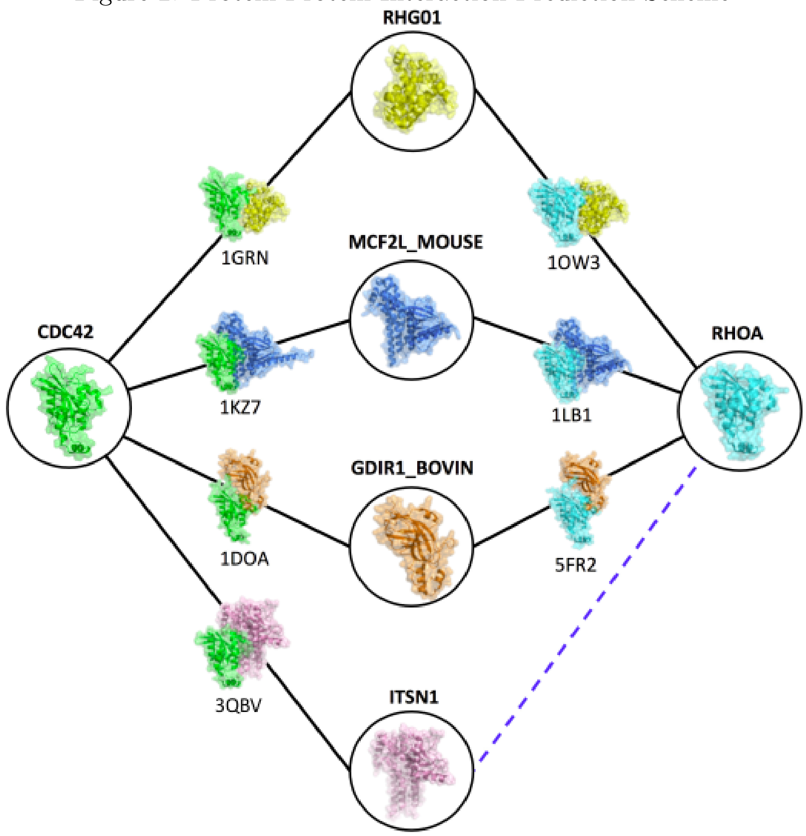


algorithms are proposed over these networks in order to predict unknown interactions. In this report, three prediction methods are presented and their results are shown: Common Neighbors (**CN**), which uses the length-2 path count; length-3 path count (**A3**); and degree-normalized length-3 paths score (**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 in the next sections, the mentioned approach fails because it does not consider the structural and chemical properties of the proteins.

For achieving the described results, several networks are used and compared using state-of-the-art methods, as well as the proposed ones (CN, A3, L3). For this report, the human interactome was used (*HI-II-14*). A curated version of the mentioned network (*HI-TESTED*) was also used in order to compare the influence of research bias on the network predictions. A massive experimental assay was carried on and its results are consolidated and used to build a validation network (*HI-III*).

Figure 2: Protein-Protein Interaction Prediction Scheme



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 [2], 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 the previous work, only PPIs with evidence from curated databases were used (rows where *databases* column has a value greater than zero), resulting in a network with NNNN nodes and MMMM edges.

2.2 Code Implementation

Previous code implementation was adapted from `C++` to `Python` (3.6), 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* (V3.6) 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 and Precision Counting

Sampling proceeds as follows: first, a fraction (10%) of the edges is removed at random from the network. Then, for each method (**CN**, **A3**, **L3**) the prediction of the removed fraction is performed. The new edges are then compared against the original (removed) edges according to the ranking that each method yields. Edges belonging to the original connected component are considered positive in the sense that they are experimentally validated interactions. For each of the tested networks and a rank r, the precision was calculated as

$$P_r = \frac{L_r}{r}$$

where L_r is the total number of positive edges in the range $[1, r]$.

The previous procedure was applied for a fraction of 0.90 remaining edges, and a mean of 10 simulations was obtained.

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

For applying the test results, a partition of the whole dataset was made: 80% of the data was randomly selected and used for training, while the remaining 20% was used for validation. The whole training-validation procedure was applied 5 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

As explained above, the precision trend for each of the networks are presented for the predicted top 2000 interactions.

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.

4 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] Kovács, I.A., Luck, K., Spirohn, K. et al. Network-based prediction of protein interactions. *Nature Communications*, 2019.

[2] 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.

Figure 3: Edge Prediction Paradigms

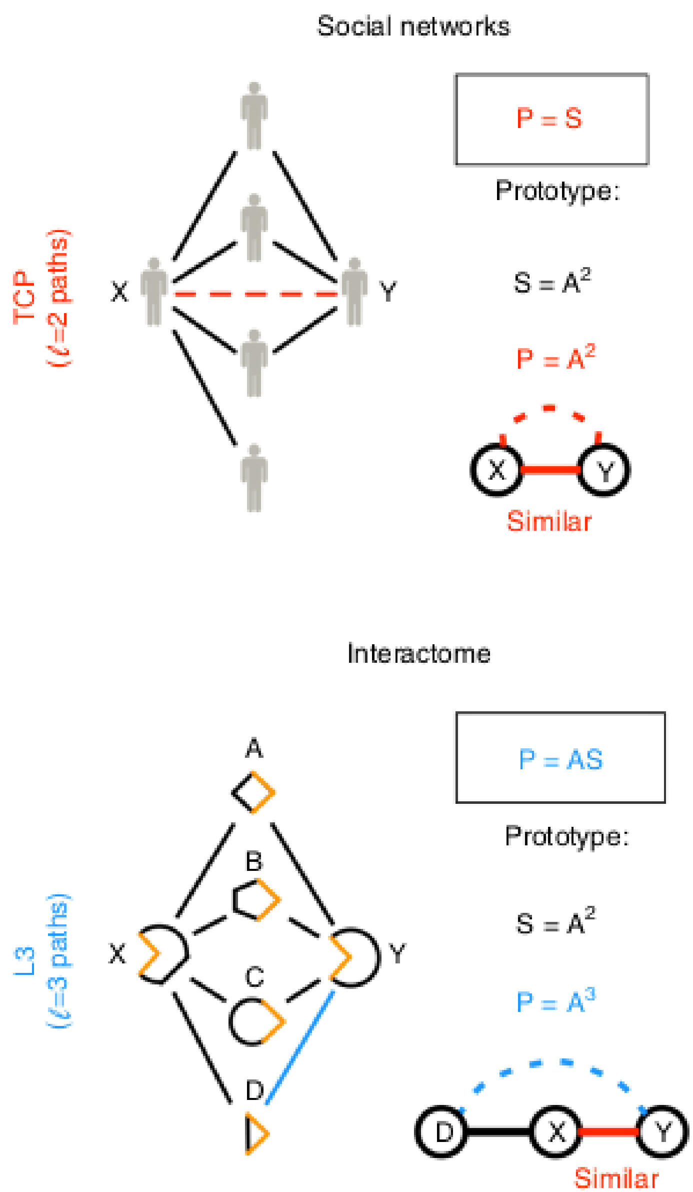


Figure 4: Methods Comparison for *HI-II-14*

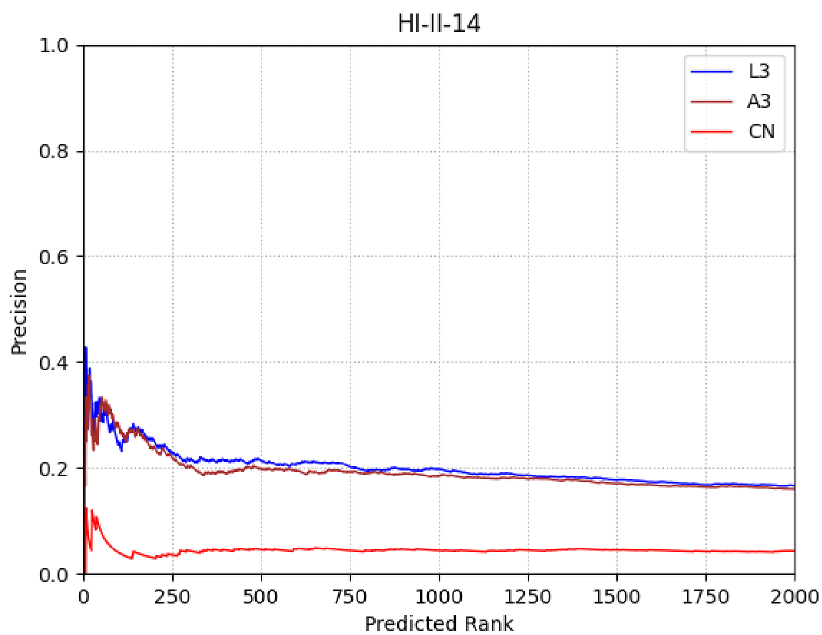


Figure 5: Methods Comparison for *HI-TESTED*