Evolutionary Time and Protein-Protein Interaction Networks

Preliminary Analysis

STA 596: Practical Data Science
Jesse Hautala
Shawn Houser
Angyalka Valcsics

December 16, 2021

Contents

I. Introduction

Proteins control all biological systems in a cell and, through various interactions with each other, enable cells to complete tasks such as: enzyme activation, gene regulation, and intercellular communication. A system of proteins can be modeled by an undirected network of protein-protein interactions (a PPI network, or PPIN), with nodes representing proteins and edges representing interactions (see Figure 1). The complete set of such interactions for a species is called the protein interactome. When interaction relationships between proteins break, possibly due to environmental factors or random mutations, such breakage can cause disease and death, of the cell and of the organism. We hypothesize that the evolutionary time of a species, which is defined as the total branch length from the root to the leaf representing that species in the tree of life, is directly related to network statistics which describe the topological stability of the species' protein interactome.

II. Background

Advances in proteomics allow researchers to study the protein interactome, but limitations of experimental methods in practice prevent PPI networks from being comprehensive and free of noise. We regard extant PPIN data, including the data used in this project, as a noisy sample from the true protein interactome. For example, the yeast-two-hybrid method for mapping protein interactomes was first developed in 1989 by Fields and Song using *Saccharomyces cerevisiae* as a biological model. The accuracy of this experimental method is estimated to be less than 10 percent. Consequently, the population being studied is the true protein interactome for each species and the variables of interest are the network statistics derived from PPIN data.

Understanding how protein interactomes evolve and developing methods for analyzing PPI networks is a central goal of evolutionary systems biology (Maddamsetti (2021)). In a paper by Rohan Maddamsetti they provided evidence that protein interactomes in E-Coli appear to show a generational increase in network resilience. Marinka Zitnik (Zitnik *et al.* (2019)) defined network resilience as the measure of how quickly a network breaks down as edges between nodes are randomly removed. A resilience rating of 1 implies that the network is most resilient while a rating of 0 implies a complete loss of connectivity in the PPI network. The current research identified a positive linear relationship between the resilience of an interactome and evolutionary time of the species.

III. Methods

III.I. Acquisition of Data

Our dataset comes from the Stanford Network Analysis Platform (or SNAP). This data was collected using the Search Tool for the Retrieval of Interacting Genes/Proteins (or STRING), from the European Molecular Biology Laboratory, and is organized in multiple text files, joined by species ID. It comprises taxonomy information, an edge set for each interactome, and a numerical variable for evolutionary time of the species.

In order to apply common network analysis techniques, we restructured the PPIN data as a series of adjacency matrices. To reduce the computational burden during initial algorithm development, we selected an arbitrary subset of 75 species of Proteobacteria, a major phylum which includes a wide variety of pathogenic genera such as Salmonella. For network statistics that only pertain to connected graphs, we use the largest connected subgraph (or LCSG).

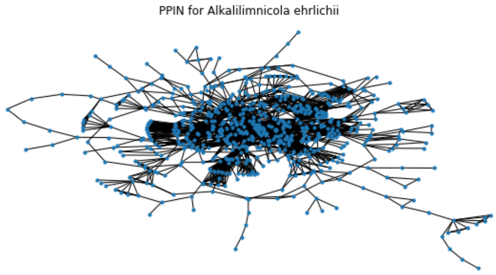

Figure 1: An example PPIN

To extract pertinent information about network stability from each network we used the concept of Exponential Random Graph Models. The basic assumption of these models is that the structure in an observed network can be explained by a vector of sufficient statistics which are a function of the observed network. To construct our models, we first need to find the vector of sufficient statistics for each network.

Using functions of NetworkX, a Python library for network analysis, we calculated statistics for each PPIN, including: average degree centrality, number of triangles, modularity, and maximal clique stats for the complete network and for the LCSG.

Average centrality for a network describes the average number of edges for all nodes. In other words, this statistic describes the average number of connections for all nodes. If a network has a high average degree centrality then we interpret this network as being dense with respect to the number of nodes in the network.

The **Number of Triangles** statistic counts the number of triangles in the network–we sum up the number of triangles each node is a part of then divide this number by three. A triangle is a set of three nodes where each node has a relationship to the other two–this is sometimes referred to as a 3-clique. Networks that have a large number of triangles tend to be highly interconnected. However, networks that have a low number of triangles turn out to be poorly connected and may suffer from instability.

Modularity is a measure of the structure of networks or graphs which measures the strength of division of a network into modules. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.

Cliques are fully connected subgraphs, meaning each node in a clique is directly connected to every other node in the clique. Therefore any clique of size $n > 1$ necessarily includes $\binom{n}{n-1}$ sub-cliques. For our network statistics (e.g. **Clique-Size Mean**) we only measure maximal cliques, i.e. those cliques which are not sub-cliques of a larger clique. Where clique metrics are derived from the LCSG, the metric name is prefixed with “LCSG”.

GiantProportion is a simple metric that seems to bring unique information into the model. It is the ratio of nodes in the PPIN that are members of the LCSG, this is *LCSG Node Count* divided by *Total Node Count*. We use this ratio to represent the health of the system during simulated “network failure”, inrementally removing random nodes until fewer than 10% of the nodes belong to the “giant component” (**GiantProportion** $< .1$). We include simulation results among network statistics as the **Critical Threshold**, the mean number of removals to failure, over 10 simulations.

Additionally, we wrote a function to find the number of k-stars in each PPIN from one to the maximum size star in the network. For example, a 1-star is the number of nodes in the network, a 2-star is the number of edges, a 3-star is the number of nodes with three edges, and so on. Finally, we combined all of these statistics into a data matrix for our models (see appendix for complete details).

III.II. Models

With these network statistics we measure topological stability of each PPIN, intending to identify a subset of key network statistics that suffice as significant predictors of the evolutionary time of a species. The question is, which among these many network statistics has the most influence on the response variable? With many more variables than data points, we have to be particularly careful about overfitting our model to the training data; many machine learning algorithms would produce models that suffer from poor generalization. Moreover, these network statistics are highly correlated with each other, which presents an added challenge, particularly for linear regression models. Consequently, we need to leverage models that afford feature selection. We chose two supervised learning models: LASSO regression and Random Forest regression.

The LASSO method regularizes model parameters by shrinking the regression coefficients, using the L1 norm, reducing some of them to zero. The feature selection phase occurs after the shrinkage, where every non-zero value is selected to be used in the model. Hence, important network statistics for this model where taken to be those which have non-zero coefficients. While building our initial model we used alpha = 0.6. This is mostly due to convergence issues and will likely change for the final model.

Random forests build a collection of de-correlated decision trees, by randomly choosing only m predictors from the full set of predictors when performing a split, the split is only allowed to use one of those m predictors. Finally, the average of the resulting trees is taken. In random forest regression, features are selected that improve the variance reduction. That is, correlation between trees is reduced without increasing the variance too much. For our initial model we used a max depth of four–a parameter that we plan to tune before choosing a final model.

IV. Preliminary Results

The network statistics with the most influence on the evolutionary time of each species for the LASSO model were the number of triangles, clique count, LCSG clique count, node count, LCSG node count, LCSG Degree Max, and number of 1-stars (see Figure 2). The sign of coefficients tells you if the network statistic is positively or negatively related to the response (evolutionary time). This project is not focused on finding a linear relationship however we will try to interpret these coefficients.

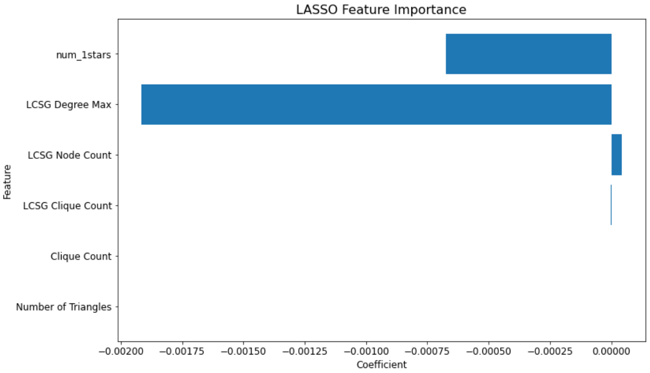

Figure 2: LASSO Feature Selection

In Figure 2 we can see that LCSG node count (the number of nodes in the LCSG) and LCSG clique count (the number of maximal cliques in the LCSG) are both positively related to a species' evolutionary time. Recall that network resilience is the measure of how quickly a network breaks down as nodes are randomly removed and it has been shown that networks with low network resilience also have a low evolutionary time. Hence as the LCSG clique count rises, the LCSG becomes more stable and so we can expect an increase in evolutionary time of the species. On the other hand if the LCSG has a node with a large number of edges (LCSG max degree), those edges only lead to other nodes in the LCSG, meaning that if this number is high then a LCSG in the network has a high density. This may or may not be a good thing for the survival of the species. If this LCSG also has a small amount of nodes, then I would assume the evolutionary time would be low but if the number of nodes is high then it may be beneficial. The number of triangles and maximal clique count have been included in the model but the coefficients are close to zero. The mean squared training error for the LASSO model was 0.05.

The random forest regression model was consistent with the results from above, please see the important variables for this model (See Figure 3). Arguably, the network statistics with the most influence on the evolutionary time of each species for this model were the number of nodes, modularity, clique count, average centrality, LCSG clique-size mean, and a number of other k-stars statistics.

We explained before how modularity and average degree centrality influence the topological stability of a network. The LCSG clique-size mean denotes the average size of a maximal clique within the LCSG. If this number is large, then we can expect the stability of this network to be strong and this translates to a higher network resilience. The k-stars statistics are likely to be significant since a network with many high degree nodes would be dense. The mean squared training error for the random forest regression model was 0.009.

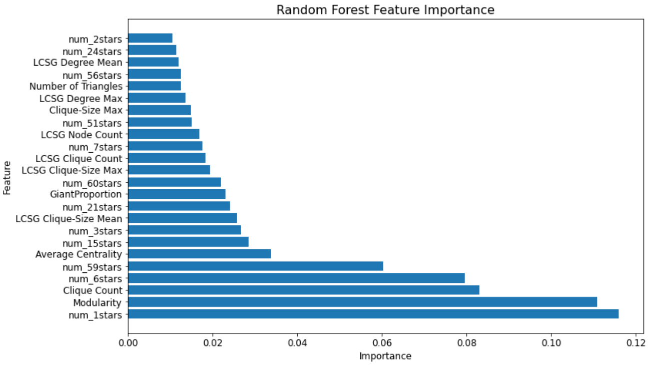

Figure 3: Random Forest Feature Selection

V. Conclusion

There are many possible confounders which should be considered along with our preliminary results: investigative biases towards modeling common or popular organisms, network size, and genome size. It seems that the network statistics we have found so far and entered into the models do a fair job of describing topological stability. Moreover, many of these statistics do have a significant relationship to evolutionary time of a species. In the next few weeks, we plan to fine-tune the hyperparameters of the models. For example we intend to add cross validation to the LASSO model and use the *RandomizedSearchCV* function to find optimal parameters for the Random Forest model.

VI. Appendix

VI.I. Exhaustive Enumeration of Cliques is Memory-bound

We tried implementing exhaustive enumeration of *all* cliques (as opposed to *maximal* cliques) but found this to be impractical, as the supporting NetworkX algorithm (*enumerate_all_cliques*) is memory-bound. We tested this limitation via execution with 64GB of RAM and 3 $Worker$ processes; execution failed with a $MemoryError$ after $\sim$ 8 hours of processing, when one of a workers attempted to allocate additional memory beyond available capacity (see Figure 2).

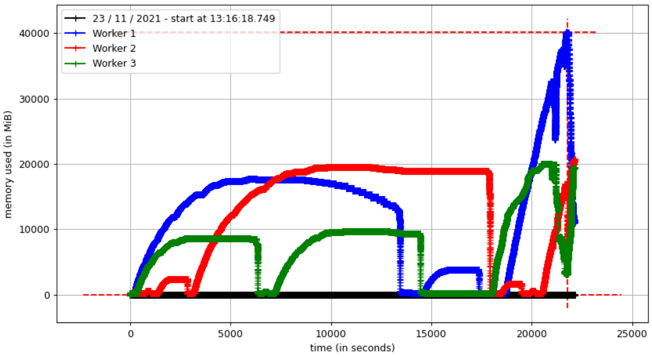

Figure 4: $Worker$ memory usage

VI.II. K-Stars Algorithm

Below is pseudocode for how we constructed the data frame of stars counts for each network.

Algorithm 1 Get Stars Algorithm

```
LCSG ← giant component for undirected network
A ← convert LCSG to adjacency matrix
d ← sum row elements of A
values, counts ← find unique elements and counts for each
stars ← pandas DataFrame of counts with index names being values
```

Table 1: Count K-Stars Data Frame

Species_ID	882	883	36870	52598	56780	⋯
num_1stars	109.0	39.0	22.0	17.0	26.0	⋯
num_2stars	105.0	57.0	54.0	30.0	70.0	⋯
num_3stars	65.0	63.0	31.0	20.0	53.0	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋱

It is important to note that the maximum star count is likely different for each network so when we concatenate these smaller data frames together to form the data frame which holds all network statistics it must be dynamic. Additionally, this means that once we train the model–if we choose to include all star variables–then we cannot test the model on a new set of networks. A quick fix for this may be to only include the first ten rows of this data frame.

VI.III. Predictors Using NetworkX

Table 2: Predictors Using NetworkX Data Frame

Species_ID	882	883	36870	52598	56780	⋯
Average Centrality	0.013	0.019	0.023	0.030	0.020	⋯
Number of Triangles	12742.000	8017.000	689.000	49.000	11135.000	⋯
Modularity	0.679	0.559	0.674	0.741	0.556	⋯
Clique Count	1060.000	3580.000	209.000	823.000	547.000	⋯
Clique-Size Max	26.000	19.000	10.000	6.000	27.000	⋯
Clique-Size Mode	2.000	5.000	2.000	2.000	2.000	⋯
Clique-Size Mean	4.726	5.004	2.923	2.335	4.075	⋯
LCSG Clique Count	916.000	371.000	180.000	249.000	451.000	⋯
LCSG Clique-Size Max	26.000	19.000	10.000	5.000	27.000	⋯
LCSG Clique-Size Mode	2.000	2.000	2.000	2.000	2.000	⋯
LCSG Clique-Size Mean	5.102	4.647	3.039	2.116	4.237	⋯
LCSG Node Count	736.000	502.000	217.000	139.000	536.000	⋯
LCSG Degree Max	60.000	58.000	27.000	11.000	66.000	⋯
LCSG Degree Mode	1.000	3.000	2.000	2.000	2.000	⋯
LCSG Degree Mean	9.465	9.661	5.060	4.144	10.590	⋯

VII. References

[1] Zitnik, M., Sosič, R., Feldman, M. W., Leskovec, J. (2019). Evolution of resilience in protein interactomes across the tree of life. Proceedings of the National Academy of Sciences, 116(10), 4426-4433. https://doi.org/10.1101/454033

[2] Maddamsetti, R. (2021). Selection maintains protein interactome resilience in the long-term evolution experiment with Escherichia coli. https://doi.org/10.1093/gbe/evab074

[3] Evolution of protein interactomes across the tree of life. (n.d.). Retrieved from http://snap.stanford.edu/tree-of-life/

[4] Sumit Mukherjee. (2011). Exponential Random graph models. Retrieved from https://artowen.su.domains/courses/319/smukherjee.pdf