Evolutionary Time and Protein-Protein Interaction Networks

Preliminary Analysis

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

November 25, 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. Protein interactions can be modeled by an undirected network of protein-protein interactions (a PPI network, or PPIN), with nodes representing proteins and edges representing interactions. 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 by 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.

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. The current research identified a relationship between the resilience of an interactome and evolutionary time of the species.

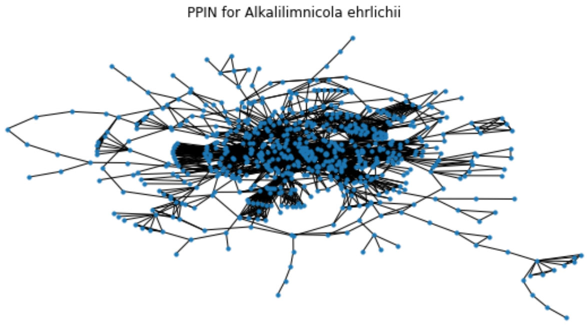
III. Methods

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, EMBL 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.

Shawn, creating the pickle is all you.

After transforming the data into a workable format, we selected an arbitrary subset of 75 species of Proteobacteria, a major phylum which includes a wide variety of pathogenic genera such as Salmonella. Our goal was to select a subset of networks which had a smaller size when compared to more complex species in the data since larger networks would increase computational time for our algorithms.



Jesse, briefly we should both talk about how we got the two separate dfs with all of our network statistics. We can go into detail about the code in the appendix–you'll see that I designated a section for it there.

III.II. Models

Our research question aims to find network statistics, which measure topological stability of the species' protein interactome, that are significant predictors of the evolutionary time of a species. That is, we suspect that there is a relationship between some of these network statistics and the response–evolutionary time. The nature of our data is that there are more features than data points. For this reason we need models that work well with such data–we need models which perform feature selection. Here I would like to explain what the ERGM is (using the pdf in the references below) and how we used it, as well as explain how we concatenated the two dfs above together

IV. Preliminary Results

LASSO info Random forest info

V. Conclusion

Need to talk about how we plan to fine-tune the above models over the next two weeks here. I.e. Adding CV to lasso and tuning hyperparameters of random forest using RandomizedSearchCV.

VI. Appendix

VI.I. Predictors Using NetworkX

Table 1: Predictors Using NetworkX Data Frame

Species_ID	882	883	36870	52598	56780	⋯
Average Centrality	0.013	0.019	0.023	0.030	0.020	⋯
Average Closed Triangles	51.938	47.910	9.525	1.058	62.323	⋯
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	⋯
Node Count	1023.000	809.000	262.000	495.000	757.000	⋯
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	⋯

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

```
giantC ← giant component for undirected network
A ← convert giantC 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 2: 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 unique to 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.

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