BT5240 - Final Project Report: Testing for the Centrality-Lethality Hypothesis in the Context of Essential Genes

Arun A. R. [*] **Manishnantha M** [**]

[*] *CS19B002, Indian Institute of Madras, Chennai, India*
(e-mail: cs19b002@smail.iitm.ac.in)
[**] *CS19B031, Indian Institute of Madras, Chennai, India*
(e-mail: cs19b031@smail.iitm.ac.in)

Abstract: In this project, we attempt to test and validate the Centrality - Lethality Hypothesis in the context of essential genes using machine learning approaches. Protein-Protein Interaction networks of various organisms and their essential genes from the STRING database were used. Various centrality measures and ReFeX features of these networks were extracted. These features were used to train a machine learning model and it was then used to predict the essential genes of an organism previously unseen by the model. The organisms used in this project are the same as the organisms used in the NetGenes paper. In this data set, using leave one out validation, a mean precision of 0.558 and a mean AUROC of 0.589 was achieved. Essential genes are, by definition, lethal and hence if we are able to show that there is a correlation between centrality and essentiality of a gene in a network, we will be showing, by extension, a correlation between centrality and lethality as well.

Keywords: Centrality-Lethality, STRING Database, Machine Learning, Essential Genes, Networks, Centrality Measures, ReFeX.

1. INTRODUCTION

The Centrality-Lethality hypothesis states that "nodes with higher centrality in a network are more likely to produce lethal phenotypes on removal, compared to nodes with lower centrality" [Raman et al. (2014)]. Here, centrality is defined by traditional centrality measures such as degree centrality, betweenness centrality, eigen-vector centrality, etc.

In the context of biology, a lethal component of an organism can be defined as a component, whose removal or change causes the death of the organism. In this project, we consider essentiality of gene as the "indispensability of a gene under rich media conditions" [Mobegi et al. (2017)].

1.1 Objectives

The main objective of this project is to test the Centrality Lethality Hypothesis. We attempt to test this hypothesis via machine learning approaches and predict the lethal nodes of a given network using centrality measures and comparing it to the lethal nodes obtained from experimental data.

As an additional objective, we would like to gain insights on how precise machine learning solutions can be in predicting the lethal nodes of a network using only centrality measures and a collection of recursively extracted graph features.

1.2 Applications

Traditional methods used to predict essential genes of an organism are resource expensive and time consuming. Hence, various in-silico processes to predict the essentiality of genes have been attempted. While the precision of the current model is far from being directly used in real life model, further developments can eventually lead to better more precise models with more reliable predictions.

2. METHODS

This project was fully implemented in Python. The list of libraries used are given in Appendix A. The code utilized to implement this project are put up in this `GitHub` repository. This project can be broken down into the following 5 stages.

- Data Collection
- Features Selection and Extraction
- Feature Scaling
- Model Selection
- Training and Prediction

2.1 Data Collection

Large scale PPI networks of organisms and their list of essential genes were required for the project. The data collected for the NetGenes paper [Karthik et al. (2018)] from the STRING database [Szklarczyk et al. (2019)] and the Database of Essential Genes (DEG) [Luo et al. (2014)] has been used here. The data set contains the PPI

networks of 27 different organisms and their corresponding list of essential genes. There were a total of 90435 genes in the data set. The list of organisms in the data set is given in Appendix B.

2.2 Features

Selection A collection of 10 well-known centrality measures have been chosen for our classification objective.

The Centrality measures used are:

(i) Degree Centrality
(ii) Eigen-Vector Centrality
(iii) Closeness Centrality
(iv) Betweenness Centrality
(v) Sub-graph Centrality
(vi) Load Centrality
(vii) Harmonic Centrality
(viii) Reaching Centrality
(ix) Clustering Centrality
(x) Page Rank

A set of ReFeX [Henderson et al. (2011)] features have also been included. Different sets of ReFeX features were obtained for each graph, hence the ReFeX features that were common among all graphs were only considered.

A total of 5 ReFeX Features were thus selected. They are as listed below.

(i) degree(mean)
(ii) degree(mean)(mean)
(iii) external_edges
(iv) external_edges(mean)
(v) internal_edges

The above 10 centrality measures and 5 ReFeX features were combined and used for predicting gene essentiality.

Extraction The Centrality measures were then extracted from these networks by using inbuilt functions from the `networkx` python package.

The `graphrole` python package was used to extract ReFeX features.

The extracted features are available in the repository for each of the 27 networks.

2.3 Feature Scaling

It was observed that the features obtained were in different ranges or scales. Hence, it was important that the features were scaled before being used to train the machine learning model, for more precise results.

The `StandardScaler` from the `scikit-learn` package was used to scale the features appropriately.

2.4 Model Selection

The objective is to train a machine learning classifier on a large data set of genes from different PPI networks. The classifier algorithm, should be capable to classify a previously unseen gene as essential or non essential based on the set of features listed above.

Among the models used, XGBoost provided the most precise and consistent results when tested. XGBoost is considered one of the most accurate modeling for structured data. XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

The python package `xgboost` was used to implement the model.

2.5 Training and Prediction

The leave one out validation method was used for training and prediction in this project. That is, out of the 27 PPI networks in the data set, a graph is picked and placed on the test set and the remaining 26 graphs are used to train the model. Training and then testing is done for every graph in the data set. Hence, the model is run 27 times and its performance is recorded each time.

3. RESULTS AND DISCUSSIONS

3.1 Results

The performance of the model was evaluated based on the following 4 metrics.

- Accuracy
- Precision
- F1 score
- AUROC

These statistics were recorded for each iteration of the leave one out validation by comparing the data of essential genes already available with the predictions made by our model. Functions for these metrics were imported from the `scikit-learn` library.

A summary of the statistics of performance of our model are given in Table 1.

	Accuracy	Precision	F1 score	AUROC
mean	0.8689	0.5585	0.2887	0.5893
min	0.6194	0.0000	0.0000	0.4817
max	0.9597	0.9333	0.7423	0.7816
std	0.0868	0.2752	0.1992	0.0701
25%	0.8424	0.3666	0.1775	0.5445
50%	0.9000	0.7055	0.2468	0.5800
75%	0.9282	0.7410	0.4195	0.6346

Table 1: Statistical Results of the Model

A mean precision of 0.5585 and an AUROC score of 0.5893 was obtained.

3.2 Interesting Observations

Features Importance An interesting question to ask could be, which of these features play a key role in predicting gene essentiality of an network?

To answer this question,the Feature Importance graph was plotted.

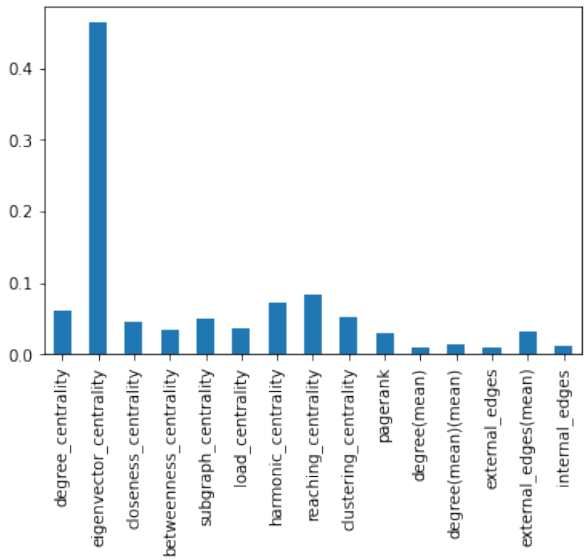

Figure 1: Feature Importance Plot

Eigen-Vector centrality stands out and is a key feature in predicting the essentiality of gene in an organism. Following that, in decreasing order of importance, we have reaching centrality, harmonic centrality, and degree centrality.

Among the ReFeX features, the external_edges(mean) measure is significant. Overall, ReFeX features do not seem to contribute as much as the centrality measures in prediction.

Additional Observations On an attempt to improve the precision of the gene prediction, Deep Neural Networks were experimented. Various Neural Networks architectures were tested, but inconsistencies in predictions were observed in all attempts.

Simple Neural Networks produced poorer results compared to XGboost. Whereas complex Neural Networks with just a few layers (1-2 hidden layers) over fit on the training set and failed to predict essential genes on the test set.

If there were a strong correlation between centrality and lethality, the Deep Neural Networks would have captured it. But the lack of consistency in results may indicate that the correlation between centrality and lethality may not be well pronounced.

3.3 Discussions

Supporting the Hypothesis Even as the model was only trained on centrality measures, it was able to predict the essentiality information with a reasonable mean precision of 0.5585 and a mean AUROC score of 0.5893. This may hint at the existence of correlation between centrality and essentiality. As observed, features with high values of feature importance in the given model, like eigen-vector centrality, may have a correlation to the essentiality of the gene.

Refuting the Hypothesis In Table 1, a minimum precision of 0 is observed. This shows that even the state of the art, XGBoost classifier algorithm fails to predict essential genes of the network in one of the iterations of leave one out validation. The model does not achieve the precision of traditional gene prediction methods.

Further, Deep Neural Networks as discussed in section 3.2, fail to capture patterns of correlation between centrality and lethality.

The above two statements indicate that there might not be a strong correlation between centrality and lethality in the context of essential genes.

From the arguments supporting and refuting the hypothesis, there is no sufficient proof supporting a strong correlation between centrality and lethality. However, a weak correlation might exist.

3.4 Future Work

Using centrality measures and using the well renowned XGBoost algorithm, a mean precision of 0.558 in predicting essential genes was obtained.

Although this beyond the scope of the project, for in-silico methods to be an alternative to traditional methods, much higher precision and AUROC score is required. To improve the precision even further of predicting gene essentiality in an organism, a few ideas are as follows.

Train and Testing among a group of organisms i.e. Homogeneity in the data set. This project includes data from a diverse set of organisms. This may have contributed to noise in the data. Better, more reliable predictions could possibly be made by restricting the data set to a group of similar organisms eg. different strains of the same bacteria.

This is because, due to inherent differences in network structure and properties, one centrality measure could be strongly correlated to essentiality in a class of organisms whereas an entirely different centrality measure could be the most correlated to essentiality for another group of organisms.

More Useful Features Centrality measures can only get us this far in predicting gene essentiality. To go beyond, one could include a set of biological features of a particular gene, which possibly captures the essentiality of a gene.

Alternatively, attempts at predicting gene essentiality only by using biological properties of the gene could be enhanced with the most important centrality measures, like eigen-vector centrality,reaching centrality, harmonic centrality and degree centrality.

4. CONCLUSIONS

In this project, the objective was to test and validate the Centrality-Lethality Hypothesis. This was performed by extracting various centrality measures from PPI networks of 27 different organisms and then using them to find the essential genes via a XGBoost Classifier. We find that correlation between centrality and lethality may not be well pronounced.

For a stronger claim, in future work, the hypothesis can be tested by analysing a more diverse set of organisms. Statistical tests to quantify the correlation between centrality and essentiality can also be implemented.

5. ACKNOWLEDGEMENTS

We would like to thank everyone involved in the BT5240 Computational Systems Biology course (Jan - May 2021 Semester) for the wonderful learning opportunity, especially Prof. Karthik Raman and the TAs.

We would also like to thank the open source contributions to the software packages used which enabled us to work on our project with ease.

Appendix A. LIST OF PYTHON PACKAGES USED

- pandas
- graphrole
- networkx
- sklearn
- xgboost

Appendix B. ORGANISMS IN THE DATA SET

Organism Name	Taxonomical ID
Staphylococcus aureus N315	158879
Pseudomonas aeruginosa UCBPP-PA14	208963
Salmonella enterica serovar Typhi Ty2	209261
Bacillus subtilis	224308
Mycoplasma genitalium	243273
Burkholderia thailandensis E264	271848
Francisella tularensis novicida U112	401614
Streptococcus pyogenes NZ131	471876
Caulobacter crescentus NA1000	565050
Streptococcus pneumoniae R6	171101
Campylobacter jejuni	192222
Pseudomonas aeruginosa	208964
Shewanella oneidensis	211586
Bacteroides thetaiotaomicron VPI-5482	226186
Vibrio cholerae	243277
Burkholderia pseudomallei K96243	272560
Mycoplasma pulmonis	272635
Streptococcus sanguinis SK36	388919
Sphingomonas wittichii RW1	392499
Porphyromonas gingivalis ATCC 33277	431947
Escherichia coli K 12 substr MG1655	511145
Acinetobacter sp ADP1	62977
Haemophilus influenzae	71421
Mycobacterium tuberculosis H37Rv	83332
Helicobacter pylori 26695	85962
Staphylococcus aureus NCTC 8325	93061
Salmonella typhimurium LT2	99287

REFERENCES

Henderson, K., Gallagher, B., Li, L., Akoglu, L., Eliassi-Rad, T., Tong, H., and Faloutsos, C. (2011). It's who you know: Graph mining using recursive structural features. In *Proceeding of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, New York, NY, USA.

Karthik, A., Ravindran, B., and Karthik, R. (2018). Network-based features enable prediction of essential genes across diverse organisms. *PLoS ONE 13(12): e0208722*. https://doi.org/10.1371/journal.pone.0208722.

Luo, H., Lin, Y., Gao, F., Zhang, C.T., and Zhang, R. (2014). An update of the database of essential genes that includes both protein-coding genes and non-coding genomic elements. *Nucleic Acids Research 42*. http://www.essentialgene.org/.

Mobegi, F., Zomer, A., de Jonge, M., and van Hijum, S. (2017). Advances and perspectives in computational prediction of microbial gene essentiality. *Brief Funct Genomics*. https://doi.org/10.1093/bfgp/elv063.

Raman, K., Damaraju, N., and Joshi, G. (2014). The organisational structure of protein networks: revisiting the centrality-lethality hypothesis. *Syst Synth Biol*. https://doi.org/10.1007/s11693-013-9123-5.

Szklarczyk, D., Gable, A., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., NT Doncheva, J.M., Bork, P., Jensen, L., and von Mering, C. (2019). String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. https://doi.org/10.1093/nar/gky1131.