CS224W - Analysis of Networks
Project Proposal
Network Analysis of Human Protein Interaction Network

Vincent Billaut
vbillaut@stanford.edu

Pierre-Louis Cedoz
plcedoz@stanford.edu

Matthieu de Rochemonteix
mderoche@stanford.edu

Introduction

In the context of *CS224W: Analysis of Networks*, we have the opportunity to work on a research project, which has the objective of enabling us to apply network analysis techniques to a concrete problem, and hopefully discover new notions not necessarily covered in class. Biology, and particularly the study of human protein interactions, is amongst those fields that are fully taking advantage of the "data era" and revolutionizing themselves and that particularly interest us.

Studying these complex protein-protein interaction networks is key to better understand diseases and design compelling new drugs. This could explain the strong interest in such an interdisciplinary field and we felt that we could bring our contribution. We will start by reviewing a few articles that stroke our attention, discuss their strength and weaknesses and how they relate to each other, before building on them to present what we intend to study in this project.

1 Review of literature

As a starting point, we will review some related work that led us to consider working on Protein Protein Interaction (PPI) Networks and gave us some ideas as where our added value could lie.

1.1 First article: Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets

Content This first article [1] particularly drew our attention, because it very well fits within our will to combine relevant, non-trivial network analysis with insightful and impactful biological discoveries. The whole approach is very interesting and stems from purely network-theoretic notions.

The authors' goal is to use the notion of *controllability analysis* in network analysis to produce insight into PPI networks. The notion of control can be understood in directed graphs by picturing the propagation of information through the edges. The system is "controllable" if we can obtain any possible output (in our case, this has to do with protein concentrations) by only manipulating a subset of the nodes (that are called the driver nodes). From this notion, they derive the notion of *indispensability*. A node is *indispensable* if its removal causes the number of driver nodes to increase. A *neutral* node's removal doesn't influence the number of driver nodes, and a *dispensable* node reduces it. In other words, it is indispensable if it helps to control the network. What the authors found is that the set of nodes classified as *indispensable* in terms of network controllability correlates well with the set of proteins involved in disease causing mutations, human viruses targets and drug targets. The authors also showed that this notion of indispensability is related to key biological features, like evolutionary conservation and cell signaling for instance.

However, a concern that quickly arises when dealing with biological networks is the literature bias: proteins heavily involved in certain diseases tend to be much more studied and documented, resulting in an over-representation of the corresponding nodes in the networks. The authors correct for this bias, as well as for the degree bias (indispensable nodes tend, for the same reasons, to have higher degrees than neutral and dispensable ones).

A deeper study of this notion gives rise to the distinction between two kinds of indispensable nodes: type I and type II. The authors looked at the robustness of the indispensability notion by modifying the network in different ways and checking which nodes basically stayed indispensable whatever the modifications (type I) and which ones didn't (type II). The robustness of the indispensability notion, shown by the authors in the article, therefore suggests that those type II nodes could be undocumented target nodes of certain diseases or viruses and therefore highly valuable.

Critique While this article develops a very interesting characterization of important nodes of a PPI Network based on the network's topology, and very much explores the ways in which this feature is relevant in several contexts, the reader could feel frustration in the fact that this study does not attempt at including this feature in a more holistic model. Showing that indispensability makes a lot of sense in the context of understanding disease causing mutations could motivate its addition in a wider model that could for instance take into account features capturing orthogonal aspects of the network like communities — the notion of indispensability mostly makes sense at the node level.

1.2 Second article: Predicting Essential Genes and Proteins Based on ML and Topological Features

Content In this review [2], the authors summarize state-of-the-art computational methods for identifying essential genes and proteins in biological networks based on machine learning and network topological features. The identification of such genes is very important not only for understanding the minimal requirements for survival of an organism, but also for finding human disease genes and new drug targets.

The learning features that are mostly associated with gene and protein essentiality are based on gene expression, sequence and functional annotation and network topology. Network topological features numerically express the position of the elements (nodes and edges) in a network in relation to all other elements. Some of them are pure network topology features: Degree Centrality (DC), Betweenness Centrality (BC), Closeness Centrality (CC), Clustering Coefficient (CCo), Subgraph Centrality, Eigenvector Centrality, Information Centrality or Eccentricity Centrality. Others are integrated topological features with biological features. For instance in [3], Tang et al. defined the weighted degree centrality that integrates network topology with gene expression profiles and in [4] Peng et al. introduced ION that integrates the orthology with PPI networks using an iteration strategy.

Furthermore, various machine learning algorithms were built on top of these network topological features to predict essential genes and proteins. The most commonly used algorithms were support vector machine (SVM), nave Bayes (NB), neural network (NN), weighted k-nearest neighbors (WKNN), decision tree, gene expression programming (GEP), logistic regression and genetic algorithm.

To sum up, many computational methods have been proposed to predict essential genes, and at the same time, many features have been found to be related with gene essentiality.

Critique In this review, the authors are presenting more than 30 research papers on essential proteins identification from protein-protein interaction networks. In recent years, researchers have developed different combinations of learning attributes and machine learning algorithms to identify essential gene proteins. However, there is no quantitative comparison of the predictive power of such methods and so one can wonder which model should be considered the best to predict essential genes and proteins? It is difficult to answer this question since there is no benchmark: for a reliable comparison among models, it would be necessary that all models were based on the same biological network.

1.3 Third article: A DIseAse MOdule Detection (DIAMOnD) Algorithm (Analysis of Connectivity Patterns)

Content This article [5] has a somewhat different focus compared to the previous ones. However, some similarities led us to also consider it. Based on the assumption that proteins linked to a given disease are in the same graph neighborhood, the purpose of the article is to design an algorithm to retrieve all of the functional groups associated with a disease given a reduced number of seed nodes. This amounts to a community detection problem.

The article focuses on designing a criterion to detect the disease communities. It appears that purely topological community detection algorithms do not allow to identify disease-related clusters. Three methods are implemented and tested: a method based on link similarities, the Louvain method (based on the maximization of global modularity) and a Markov-based Clustering algorithm. The functional clusters identified through these methods are not significantly linked with diseases.

However, given a seed of disease-related nodes, it is possible to retrieve the other nodes of the disease community. Actually, a relevant criterion is the significance of the connections with seeds rather than the density. This criterion allows to exhibit communities that do not coincide with topological clusters and are relevant to the disease. An analysis of the robustness of this criterion to the removal of edges suggests that it is relevant, and that it is more significant as the completeness of the network increases. The performance is also tested on synthetic module reproducing the structure of disease modules.

Finally, the robustness of the algorithm to alterations of the seed set is evaluated, and shows that except for a minority of nodes, the behavior is not changed.

Critique Although the approach is quite different from the previous articles, some links are worth exploring. For instance, the selection and the power of the seed nodes may be linked to their importance in the network.

However, the article does not investigate this notion of seed nodes and the way they are selected, they simply assume that we already know part of the disease module. Given the two previous articles, the issue of finding the seed corresponding to a disease module seems important, and related to network robustness analysis and node importance classification.

This is part of a main flaw of this article, which is oriented towards finding a specific network feature corresponding to a biological concept, with a research process that does not seem neither exhaustive nor very reproducible. Actually, the method does not seem to have been tested on many PPI Networks or on other similar biological networks.

1.4 Brainstorming

In these articles, the common goal is to identify important proteins (drug targets or disease genes) from the analysis of protein-protein interaction networks. The first article introduces the notion of network controllability whereas the review presents computational methods, namely supervised algorithms based on network topological features such as node degree, clustering coefficient, or betweenness. The last article is not focused on characterization of a node importance, but identifies core seed nodes useful in community detection. Even if the perspective may seem different, there is an underlying notion of node importance in the disease modules.

We realized that most articles highlight relevant features in the detection of important proteins in a network but none of them actually compare the predictive power of these methods. It is clear that in recent years researchers have made great efforts to improve the prediction models of essential genes. For this purpose, different combination of learning attributes (biological and topological features) and machine learning algorithms have been tested as shown in the literature review.

However, one can wonder which model should be considered the best to predict essential genes and proteins? It is difficult to answer this question since the direct comparison of the prediction performances of the models is impractical. For a reliable comparison among models, it would be necessary that all models were based on the same values of network topological features and on the same type of biological networks. Unfortunately, studies are based on many different databases of a certain type of interaction (e.g., DIP, BioGRID and IntAct for PPIs) and, usually, these databases are regularly updated. Different databases or newer versions of a given database will have different sets of interactions that, in turn, will give rise to new networks with distinct structures and, consequently, different values of network topological features.

This need to unify and compare the results of the topological features goes along with a necessity to link the core nodes detection process with the module detection. Actually, none of the studies presented here provides a full pipeline of Disease module detection, they either retrieve the modules from seed nodes, or detect the most important nodes in the network.

One of the main features of all the articles presented here is the focus on a specific biological application of graph theory. From an external point of view, it seems that there may be a strong confirmation bias in the selection of the methods selected by the articles, in that they elaborated based on the structure of the networks they wanted to analyze until they found a good metric or a good algorithm, instead of designing several models based on graph theory and testing it on the real network. Those articles may therefore lack an agnostic, rigorously network analysis-based approach to the analysis of the Protein-Protein Interactions and the link with diseases.

In addition to this, the results shown in the articles, and especially in [5], have not been extensively cross-tested on diverse PPI networks. Using a unified database like STRING could be a way to cross-validate the relevancy of the methods on a single PPI network without having to redesign and adapt the framework for each method.

2 Project proposal

2.1 Our goal

To overcome this shortcoming, we propose to perform a quantitative comparative study of the predictive power of all network topological features presented in the previous articles. To this end, we will extract and combine all the topological features for all the nodes in a given interaction network and build a robust support vector machines classifier on top of them. We will then perform feature selection to quantitatively evaluate the importance of every features. The purpose of such an analysis is to predict the essential nodes of the network, both from a topological point of view (that is, the nodes which determine the robustness of the network), and from a biological point of view (that is, the nodes likely to be targeted by a disease), and to exhibit the links existing between those two concepts.

The efficiency of the node as a seed in Disease module detection, and even its belonging to such a module can also be linked to its overall importance, and checking the relationship between those two approaches can be seen as a way to validate our model, and may even be integrated as a new feature set if this is computationally realistic, or as a second step in the analysis.

As a result, we propose to systematize the detection of essential nodes based on a rigorous quantitative study. This would be done in a first time on the human PPI network, and if possible, should be extended to other organisms. We will then explore the relevance of using the essential node detection metrics to identify sets of seeds that can be used for Disease Module Detection, and try to establish links between those two approaches of the disease-protein interactions in a common framework. This would be a way to address the principal flaw of [5], which is the arbitrary selection of seed nodes.

2.2 The data

PPI Network data There are many databases publicly available that gather PPI Network data. One of the most common, that was used by several of the articles we reviewed, is the STRING[1] database. It contains a comprehensive state of knowledge on all the protein protein interactions known today, and we will have to restrict ourselves to a certain set of nodes and edges relevant to our problem (e.g. human genes, perhaps only related to a particular set of diseases) in order to have manageable volumes of data (the main dumps range from 17Go to 450Go, before decompression).

For a single PPI network from the STRING database, the data consists in a directed weighted graph, the weights representing the degree of certainty of the edges. Typically, the human PPI is a 450 Mb `.txt` file containing the list of edges. We will use the `networkx` `Python` library[2].

This will allow us to begin the Network-analysis part as soon as possible.

A more critical point is to obtain a database that allows us to have links between the proteins and diseases, so as to be able to validate the approach and actually link the biological importance of nodes to their importance in the network.

Drug Targets data One of the best ways to do this is to use a drug target database. Actually, we make the assumption that the proteins targeted by FDA approved drugs are important proteins of the network, and that they have a strong overlap with the proteins targeted by the diseases. Several drug target databases are available, amongst which DrugBank[3]. The only flaw of this approach is that as often for biomedical databases, the nomenclatures differ and the correspondence between databases is not easy to solve. We have already found a way to link STRING nodes to DrugBank targets (see Figure 1).

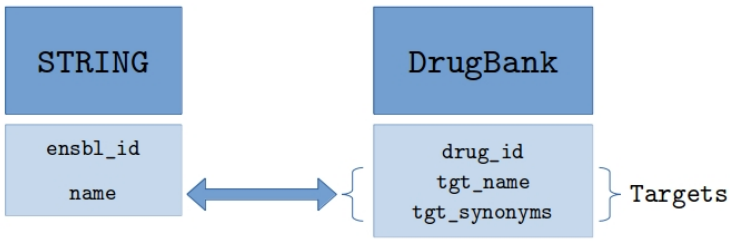

Figure 1: Illustration of the link between STRING and DrugBank

However, this matching is not exact. This will be enough to begin with, but we may need a third nomenclature database (such as *uniprot*) to have a more precise matching. In case a correspondence could be impossible to build, we will focus on the purely network approach and try to extend our methods for core nodes detection to other networks so as to extensively benchmark it.

To sum up, the essential part of the data (the PPI network) is already ready, and we have a prototype of the data needed for the second phase.

2.3 Methodology

To do this, we will list different topological features reflecting different conceptions of the importance of a node in a network. We will reproduce and validate the Disease Community detection algorithm and thoroughly test its robustness to changes in the seed set on our STRING graph.

If the Drug target database allows us to have enough correspondences, we will then build a **predictive model** (likely simple, given the reduced number of features). This will allow us to quantitatively and topologically **describe the biological importance of a node**. Once this analysis is complete, we will test its relevancy by plugging it into the disease module detection model, using the nodes detected by our algorithm as seeds.

[1] Official website: `https://string-db.org/`
[2] This choice is motivated by the fact that all standard methods in `networkx` are compatible with weighted graphs and take the weights into account
[3] Official website: `https://www.drugbank.ca/`

2.4 Potential changes to the project direction

Actually, our project can be separated in two phases:

- Topological characterization of the drug targets
- Use of this result as a base for disease community detection

We can imagine a case where we fail to establish a significant link between topological features and drug targets or biological importance. In this case, the side-task of validating the disease community detection algorithm could still be doable. However, an interesting thing in this case would be to define an alternative definition of what an important node is (an example of this is the notion of driver node defined in [1]), and focus on an algorithm able to detect the critical nodes for this metric in an efficient way, and on very large networks. Our project would then become more computational than biology-oriented.

2.5 Summary

To sum up, the main goal of this project is to properly define and identify the notion of importance of a node in a network, which goes along with the notion of robustness of this network. This will lead to a characterization of important nodes and an algorithm to detect them (this can be seen as an optimal percolation in terms of network structure alteration). Once we can detect important nodes, we will determine whether this algorithm is relevant from a biological point of view, and if such nodes are actually linked to a disease. A disease module detection algorithm will then be integrated to the pipeline to complete the protein-disease link detection. The integration of this algorithm will be an occasion to benchmark it in a non-biased way.

Even if we fail to actually link our findings to a biologically relevant interpretation, this study will allow us to explore links between the importance of a node in a network robustness and its importance in community detection.

References

[1] Arunachalam Vinayagam, Travis E Gibson, Ho-Joon Lee, Bahar Yilmazel, Charles Roesel, Yanhui Hu, Young Kwon, Amitabh Sharma, Yang-Yu Liu, Norbert Perrimon, et al. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the National Academy of Sciences*, 113(18):4976–4981, 2016.

[2] Xue Zhang, Marcio Luis Acencio, and Ney Lemke. Predicting essential genes and proteins based on machine learning and network topological features: a comprehensive review. *Frontiers in physiology*, 7, 2016.

[3] Xiwei Tang, Jianxin Wang, and Yi Pan. Identifying essential proteins via integration of protein interaction and gene expression data. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–4. IEEE, 2012.

[4] Wei Peng, Jianxin Wang, Weiping Wang, Qing Liu, Fang-Xiang Wu, and Yi Pan. Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. *BMC systems biology*, 6(1):87, 2012.

[5] Susan Dina Ghiassian, Jörg Menche, and Albert-László Barabási. A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS computational biology*, 11(4):e1004120, 2015.