Motifs in Protein-Protein Interaction Networks for Disease Pathway Prediction

Sahil Arora sarora@gatech.edu

Thomas Hu thomashu@gatech.edu

1 Abstract

In this work, we showcase the effectiveness of motif-enhanced embeddings in disease pathway prediction on protein-protein interaction (PPI) networks. Given an undirected PPI network with 30 different associated diseases, we performed node classification with an array of supervised learning methods. Specifically, we applied node2vec with logistic regression, graph convolutional networks, and GraphSAGE to the network to classify the associated diseases for each node. By including the orbital values as well, which is the frequency of specific motifs occurring in the graph, we show a significant improvement in classifying each node with its associated diseases. We also show that the subsampling is just as effective as exact search for this task, and that much of the improvement is from embeddings derived from smaller motifs.

2 Introduction

Most cellular components exert their functions through interactions within the interactome, a network that describes physical interactions within a cell [1]. Discovered pathways are systems of interacting proteins and molecules that, when mutated or otherwise altered in the cell, manifest themselves as distinct disease phenotypes [2] and are encoded in the interactome. Computational discovery of disease pathway is a crucial problem that can give insights for medical diagnosis and drug-disease treatment but their discovery is a challenging computational task as it requires identifying all disease-associated proteins and their pathways.

Disease pathways can be characterized and discovered through disease-associated proteins in a PPI network, which maps the interactions between the different proteins in the human body. In the network, a node is a protein and they share an undirected edge if the proteins interact in some biochemical process. Computational analysis can be used on known disease-related proteins to identify other proteins in a disease pathway. Because of this, classifying disease associations of proteins in PPI networks has become an area of interest in computational biology. Training machine learning classifiers on node embeddings of these networks via node2vec [3], struct2vec [4], and DeepNet has been one approach. Using graph convolutional neural networks (GCNs) directly on the graph is a novel approach that has been used sparingly.

Motif detection and discovery algorithms have been used in PPI networks to find key groups of proteins and sets of interactions involved in particular diseases. Network motifs are subgraphs that are statistically overrepresented relative to a random network [5]. It has been shown that PPI networks exhibit interesting features in terms of repeated occurrences of certain modules and that these modules have biological meaning, since they may represent evolutionary conserved topological units of cellular networks [6]. Thus, much attention has been paid to the identification of small subgraphs, particularly those occurring significantly often within the biological networks. In fact, work has shown that these motifs can be potentially used for disease pathway discovery [2].

Despite the effectiveness of both approaches, there is little work combining motif data with learned embeddings for disease pathway prediction. Detected motifs have been concatenated to neural embeddings for disease

classification with logistic regression [2]. But there has not been a comprehensive exploration of incorporating detected motifs into various classifiers. We investigate this topic to attempt to answer the following questions:

- **How effective can graph neural networks be in disease pathway prediction?** Learning has only been done on node embeddings in this task, so extending this task to methods such as Graph Convolutional Networks and GraphSAGE would be novel.
- **Can these network motifs be correlated with diseases to improve classification algorithms?** This will be tested by training classifiers on embeddings with and without the motif data.

3 Related Work

Menche et al. [7] presents a network-based framework to identify the position of disease modules (a connected subgraph formed by disease-related proteins) in the interactome and predict the relationship between pairs of disease by looking at overlap of disease modules. The authors show that disease modules can only be uncovered for diseases whose number of associated genes exceeds a critical threshold determined by the network incompleteness. They claim that diseases sharing pathobiological similarity tend to be clustered in the same neighborhood of the interactome. If two disease modules overlap, local perturbations causing one disease can disrupt pathways of the other disease modules as well, resulting in shared clinical and pathobiological characteristics.

Agrawal et al. [2] focus on discovering disease pathways in PPI network. They proposed to look at higher-order connectivity by investigating motifs and count the number of times a protein from a disease pathway is located at a specific position in these motifs. They used the neural embeddings of these features fed into a logistic regression classifier to predict the disease related proteins from the PPI network. They also showed promising results by feeding the motifs as well to the logistic regression classifier, improving its performance by 11%. This is one of the only works to do this in disease pathway prediction, and it is the main paper we will be extending.

Bajaj et. al is another relevant work that extends Agrawal by using GCN instead of logistic regression with pretrained node2vec embeddings for disease classification [8]. GCN provides a semi-supervised learning framework based on node embeddings which is an efficient variant of convolutional neural networks to operate directly on graphs [9], which was ideal where only a few proteins have diseases associated with them. They also employed GraphSAGE, which generates its own embeddings by by sampling and aggregating features from a node's local neighborhood. As a general inductive framework, it is useful for leveraging node feature information to efficiently generate node embeddings for previously unseen data. [10] They do not incorporate motif information in this work, however, so we extend what they have accomplished with infusing detected motifs into embedding information.

Motif discovery has been explored extensively in application to biological networks, as the biology of a living cell involves many intricate networks of interdependent events and interactions among molecules such as transcriptional or gene regulation networks, PPI networks, metabolic pathways, neural networks, etc [5]. As motifs gained interest, multiple algorithms arose with differing approaches. There are network-centric algorithms that enumerate all subgraphs with that occur in the target network, and there are motif-centric algorithms that first enumerate all subraphs of a certain size. There are also exact search strategies and sampling strategies that approximate to find larger motifs. Ciriello and Guerrera did a review of motif discovery and detection algorithms specifically on PPI networks, and found that network-centric approaches were more effective [6]. However, they did not comment on whether the tradeoff for sampling is worth it. We plan on exploring this by using exact search and subsampling for FANMOD, a network-centric algorithm.

4 Experimental Setup

4.1 Dataset

The PPI network from Menche et al. was culled from 15 databases and contains physical interactions experimentally documented in humans [7]. The network is unweighted and undirected with 21,557 proteins and 342,353 experimentally validated physical interactions. A protein-disease association occurs when the alteration of a protein to a disease. This is the PPI networks we used. As for the protein-disease associations, we used DisGeNET, a platform that centralized the knowledge on Mendelian and complex diseases [11]. We examined over 21,000 protein disease associations, which are split among the 519 diseases (which are simplified to 30 categories) that each has at least 10 associated disease proteins. The figures below show that a few diseases are more prevalent than others but most of the proteins have no diseases associated with them, and a small few have multiple associations.

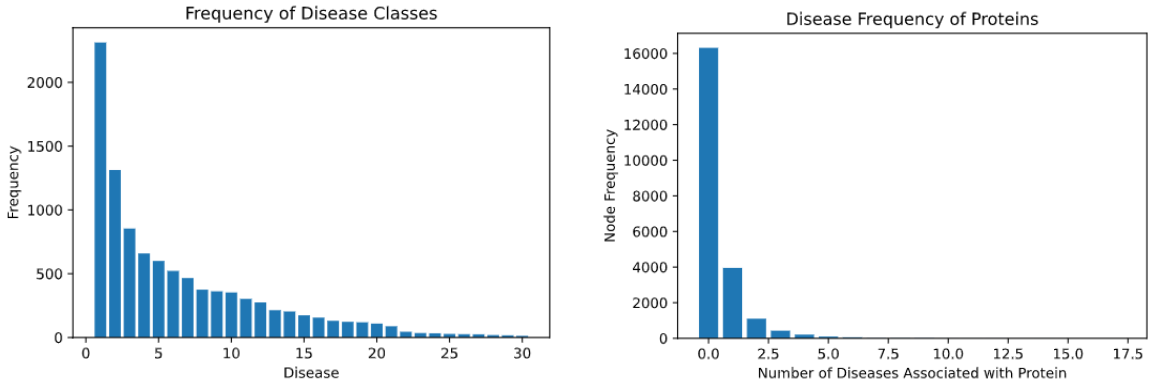

Figure 1: Distribution of Disease Frequency and Classes

4.2 Supervised Learning Methods

All of the methods were trained on 20% of the data, with 40% of the data being used for validation, and the remaining 40% for testing. Stratified splits were used to ensure equivalent class distributions between the split datasets. The metrics used to evaluate these classifiers were accuracy and F1-micro. Accuracy was not sufficient on its own as 75% of the nodes had no disease association at all, so a classifier can do this well by simply predicting the dominant class. We use F1-scoring, which is the average of precision and recall, to evaluate the predictive power of our classifiers more fairly. F1-micro is the average of the F1-scores across all of the classes, which gives equal weight to each possible label. This was chosen as certain disease categories are more frequent than others, such as cancer, and our evaluation did not want to be similarly biased towards a more dominant class. Accuracy, F1-micro, average precision, and average recall scores for the classifiers are all shared in the Results section.

GCN and GraphSAGE are both neural networks, and similar training mechanisms were applied to both classifiers, as after hyperparameter tuning it was found that different optimizations for each did not have a significant effect on downstream classification accuracy. The outputs of the classifiers are both fed to a dense layer of size 30 with a softmax activation to predict the disease classes. The cross entropy loss function was optimized with Adam, an adaptive learning rate optimizer that combines stochastic gradient descent with momentum and gradient scaling. The initial learning rate was 0.005. This network was trained for 200 epochs with early stopping for when the validation and training losses converge. To prevent overfitting, dropout of 0.5 is applied to the layers.

4.2.1 Node2Vec Embedding with Logistic Regression

Each node in the full network was embedded into a 128 length vector, which was generated through 10 random walks of path length 100 through the network for each node. These were then modeled with windows of various sizes separately, as well as with varying probabilities for returning to and moving away from source node for a random walk. Tuning these hyperparameters had little effect downstream, so we settled on using a window size of 5, and equal probabilities of 0.5 for a random walk to return to a source node and moving away from a source node. A logistic regression classifier was then applied to this generated embedding, which had the task of classifying each node in the embedding with the 30 diseases. Since a protein could have 0 to multiple diseases associated with it, this was a multilabel classification problem, which logistic regression is not specifically built for. The task was binarized, so for each of the 30 labels, a logistic regression classifier trained on the embedding made a binary prediction for each association. The final accuracy scores are a culmination of this ensemble of predictions.

4.2.2 Graph Convolutional Network

A graph convolutional layer is similar to a layer found in a neural network, but also contains the normalized adjacency matrix of the graph, as well as the node features. This encodes the structure and features of the graph into the trainable layers. After testing multiple hyperparameters, the network presented here consists of two layers of size 32, with a ReLU activation in between the layers. It is trained with the configuration mentioned above.

4.2.3 GraphSAGE

GraphSAGE does not use the graph directly like GCN, as it generates its own embeddings by sampling and aggregating features from a node's local neighborhood. Like GCN, GraphSAGE architecture here is also a two layer structure of size 32. For the first layer, neighborhoods of size 100 are sampled for each node, and for the second, neighborhoods of size 50 are used. The generated embeddings from here are trained in a neural network architecture in the configuration described earlier in this section.

4.3 Motif Discovery and Detection

In order to better combine graph embedding and motifs detection. We looked at graphlet orbital positions [12]. Graphlets are small connected non-isomorphic subgraphs of a large network. Since our PPI network is undirected, graphlet is well defined for characterizing the network. We can accomplish this by generalizing degree distribution into distributions measuring the number of 'nodes' in graphlets. For subgraph of size 2 to 5 nodes, the node can be classified into 73 position of graphlets shown in Figure 2.

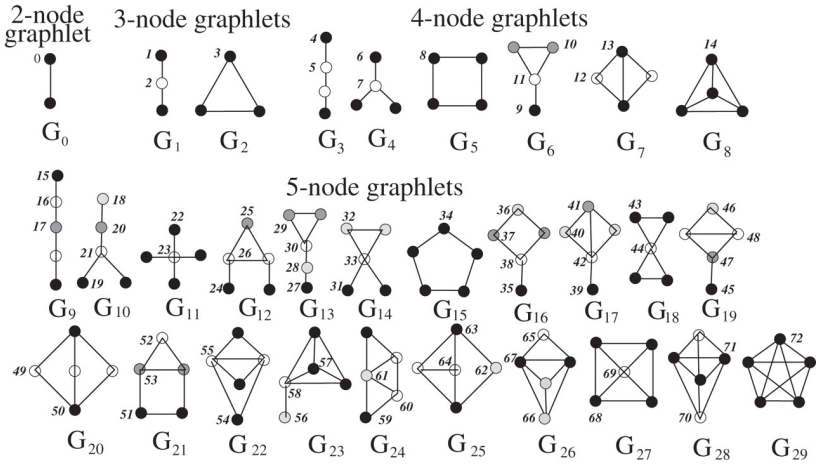

Figure 2: 73 possible motifs ranging from size 2 to 5

We implemented FANMOD, which employs Rand-ESU, an exact search algorithm that uses pattern growth trees to find maximal and unique subgraphs using partitioned pattern growth trees, which while effective at finding motifs, can also return redundant isomorphic candidates. It has random sampling such that each leaf is visited with equal probability without traversing the entire tree [6]. We followed the graphlet orbital types in the PPI network. We did not focus on 2 node graphlets as they are trivial as any edge can be described as such. We successfully extracted positions for size 3 motifs using the exact search and subsampling to compare its effectiveness. For size 4 motifs on our network, we solely used subsampling due to the computational infeasibility to run exact search on our large network. For this same reason we were unable to find 5 node graphlets through exact search or a sufficient subsampling. However, Agrawal et. al. provided the orbital positions up to 5-node graphlets for the same dataset, so we used those orbital positions. The subsampling technique we used was graphlet orbital estimation by random sample of vertices with cutting probabilities of 0.5 for each level of the search tree. The motif features is therefore the number of orbital positions that a protein is in the PPI network.

5 Results

From the accuracy score we can see that Node2Vec is not a good classifier with the additional motif information, which does not align with the previous results. Even using the provided motifs from previous research, we were not able to generate embeddings or train logistic regression to acheive the purported 11% increase in accuracy [2]. In fact from our experiments we found that embeddings generated from sampling nearby nodes, whether it is through a random walk in Node2Vec or neighborhood sampling in GraphSAGE, were not effective at discriminating disease classes.

While accuracy scores give us some sense of the effectiveness of the classifiers compared to each other but they do not tell the full story. When looking at the F1-micro scores we can see that motif information did improve GCN significantly in comparison to the baseline. Using all motifs up to size 5 caused the most improvement, but the difference between all the motif-enhanced GCN results are within 1.5% of each other. This shows that subsampling is just as effective as the exact search, as it still imparts useful information about the distribution of the graphlets for different pathologies. It also shows that there are diminishing returns on examining higher order connectivities, as including size 5-motifs does not significantly improve the F1-micro or accuracy scores compared to size 3-motifs.

	Node2vec		GCN		GraphSAGE	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
baseline	**0.7569**	**0.04830**	0.7562	0.1500	**0.7641**	0.0676
3-motifs exact	0.0068	0.0163	0.7551	0.5277	0.7583	0.0715
3-motifs subsampled	0.0731	0.0292	0.7530	0.5224	0.7563	0.0799
4-motifs subsampled	0.0336	0.0288	0.7599	0.5192	0.7602	0.0593
5-motifs from Agrawal	0.0011	0.0194	**0.7605**	**0.5348**	0.7464	**0.0842**

Table 1: Accuracy and F1-micro scores of Methods with and without Motifs. Bold indicates best score across different motif embeddings

The ineffectiveness of logistic regression makes sense when we visualize the node2vec embeddings and the associated diseases for each node. Below is a t-SNE visualization of the embeddings reduced to two dimensions, with the colors representing the fact that a protein is associated with a disease. While 75% of the nodes have no relation to any disease in the corpus, there is no separable distinction in the embedding, which will make it difficult for any algorithm to learn which proteins and nodes are associated with particular diseases. This also explains why GraphSAGE had similar issues. The random walks for Node2Vec are likely to contain nodes in similar neighborhoods that were sampled. Even a more powerful model like a 2 layer neural network will still struggle to find meaningful separation in a similar embedding. This would have to be experimentally verified by attempting to train larger and more complex GraphSAGE models in future work.

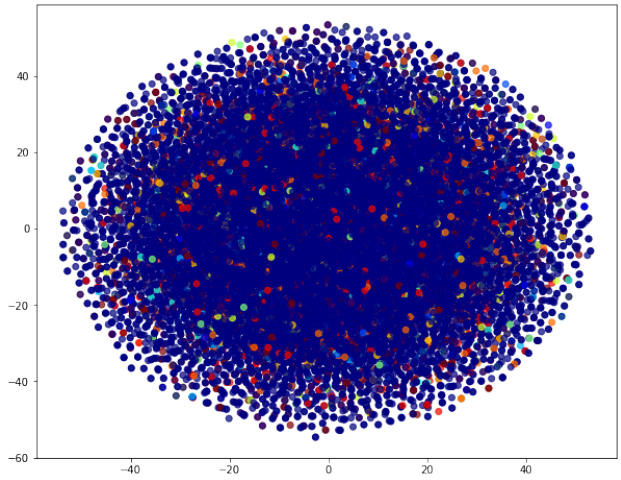

Figure 3: t-SNE visualization of node2vec embeddings

Average precision and recall over the disease classes, presented in the table below, reinforce earlier findings as well. Logistic regression on the Node2Vec embeddings has high recall and low precision, because the classifier trivially predicts that all proteins have no disease associations. Both GCN and GraphSAGE did not do this, as they have more balanced precision and recall scores. GCN however was much more effective at identifying diseases, as their average precision and recall scores were almost an order of magnitude more accurate than GraphSAGE.

	Node2vec		GCN		GraphSage	
	Precision	Recall	Precision	Recall	Precision	Recall
baseline	0.0061	0.5131	0.1400	0.1900	0.0649	0.0715
3-motifs exact	0.0085	0.6615	0.5105	0.5462	0.0742	0.0700
‘ 3-motifs subsampled	0.0153	**0.8884**	0.5088	0.5368	0.085	0.0765
4 motifs subsampled	0.0151	0.8262	0.4995	0.5406	0.0626	0.0571
5-motifs from Agrawal	**0.0142**	0.4442	**0.5152**	**0.556**	**0.0856**	**0.084**

Table 2: Average Precision and Recall

6 Conclusion

We have shown the effectiveness of motif-enhanced embeddings in disease pathway prediction on protein-protein interaction (PPI) networks. Given an undirected PPI network with 30 different associated diseases, we performed node classification with an array of supervised learning methods. Specifically, we applied node2vec with logistic regression, graph convolutional networks, and GraphSAGE to the network to classify the associated diseases for each node. While we were not able to generate successful results with embedding based approaches like Node2Vec and GraphSAGE, we did show the promise of motif enhancement with GCNs. We found a significant improvement in classifying each node with its associated diseases. With GCN results, we also show that the subsampling is just as effective as exact search for capturing the motif distribution information necessary for this task, and that much of the improvement is from embeddings derived from smaller motifs. There are many future directions for this work. We can try to improve the embedding based approaches to acheive more comparable results with existing literature, explore other motif discovery algorithms, and extend this methodology to other tasks that employ GCNs.

References

[1] S. Ghiassian. Network medicine: A network-based approach to human diseases. 2015.

[2] Monica Agrawal, M. Zitnik, and J. Leskovec. Large-scale analysis of disease pathways in the human interactome. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 23:111–122, 2018.

[3] Aditya Grover and J. Leskovec. node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[4] Leonardo F. R. Ribeiro, Pedro H. P. Saverese, and Daniel R. Figueiredo. struc2vec: Learning node representations from structural identity. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017.

[5] E. Wong, Brittany Baur, S. Quader, and C. Huang. Biological network motif detection: principles and practice. Briefings in bioinformatics, 13 2:202–15, 2012.

[6] G. Ciriello and Concettina Guerra. A review on models and algorithms for motif discovery in protein-protein interaction networks. Briefings in functional genomics & proteomics, 7 2:147–56, 2008.

[7] Jörg Menche, A. Sharma, M. Kitsak, S. Ghiassian, M. Vidal, J. Loscalzo, and Albert-László Barabási. Uncovering disease-disease relationships through the incomplete interactome. Science, 347, 2015.

[8] Payal Bajaj, Suraj Heereguppe, and Chiraag Sumanth. Graph convolutional networks to explore drug and disease relationships in biological networks.

[9] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

[10] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In Advances in neural information processing systems, pages 1024–1034, 2017.

[11] Sagar Maheshwari Marinka Zitnik, Rok Sosič and Jure Leskovec. BioSNAP Datasets: Stanford biomedical network dataset collection. http://snap.stanford.edu/biodata, August 2018.

[12] Nataša Pržulj. Biological network comparison using graphlet degree distribution. Bioinformatics, 23(2):e177–e183, 01 2007.