Disease Protein Pathway Discovery using Higher-Order Network Structures

Kevin Q Li, Glenn Yu, Jeffrey Zhang

1 Introduction

A disease pathway is a set of proteins that influence the disease. The discovery of disease pathways is an important problem because these pathways can reveal the underlying disease mechanism, uncover significant protein interactions, and elucidate potential treatment options. Finding the protein-protein interactions that accelerate or mitigate disease development may provide novel insight into safer drug targets. A greater understanding of the disease protein network further helps to trace diseases to their root biological causes and identify unforeseen side effects of current treatments.

The disease pathway discovery problem is usually set up as follows: given a protein-protein interaction (PPI) network and a set of proteins known to be associated to the disease, predict other proteins in the network that are also part of the disease pathway. Generally speaking, previous research on this problem has primarily focused on local connectivity features of the known associated disease proteins. In other words, new proteins were predicted by building from the known associations using some lower-order network heuristic such as path length or common neighbors. It has been shown recently, however, that many disease pathways have a considerably more complicated network structure than a single tightly-connected component of protein nodes, as suggested by the use of local heuristics. Disease pathways actually exhibit more loosely connected subgraphs of network nodes that contribute to disease development through sophisticated interactions. Thus, it is critical to develop new methods for the analysis and determination of common motifs within such PPI network structures to more effectively elucidate disease pathways with similar patterns. In this project, we will focus on studying these higher-order network structures and applying the methods to a PPI network to robustly uncover disease protein pathways.

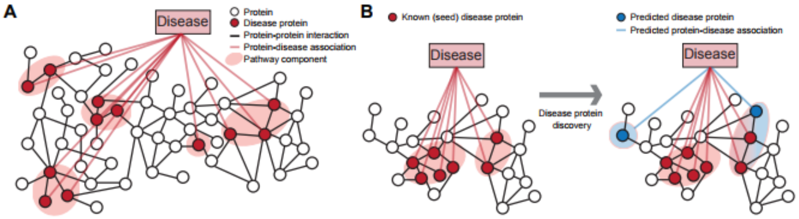
A
Disease
○ Protein
● Disease protein
— Protein-protein interaction
— Protein-disease association
● Pathway component
B
● Known (seed) disease protein
Disease
Disease protein discovery
● Predicted disease protein
— Predicted protein-disease association
Disease

Figure 1: Visualization of disease pathways and the prediction task.

2 Prior Work

2.1 Preliminary methods

As aforementioned, previous research has focused on exploiting local features of the known associated disease proteins. Agrawal *et al.* [1] listed a few examples of such "network-based" methods: neighborhood scoring, random walk, and DIAMOnD. They further showed that some of these methods do not perform particularly well on many disease pathways. For example, neighborhood scoring weights each node based on the number of connections it has to a known disease protein, which means that it is extending the disease pathway by merely enlarging it as it were. This method would fail to predict proteins that are not so directly linked to the known disease proteins or if it were to predict them, would also predict many irrelevant proteins along the way.

2.2 Higher-order network structures

In 2010, Milenković´ *et al.* [2] interestingly used the local topological signatures of protein nodes in order to find motifs that might recur throughout the network. A network motif is a statistically significant recurring subgraph of the network. Rather than focus on general network properties like high node degrees as neighborhood scoring does implicitly, the specific local pattern around each node is taken into account when identifying motifs. This study considered n-graphlet motifs. A n-graphlet is one of all possible connected non-isomorphic induced subgraphs of n nodes (this paper used $n = 2$ to 5, 30 distinct graphlets in total). The topological signature of a protein node is then a vector of graphlets V, where $V[i] = 1$ if the node is part of a graphlet i.. In this way, they were able to capture the topology or neighborhood around any protein node. Given these graphlet vectors, different clustering methods were then used to group the protein nodes based on signature similarities. Although this approach worked well to distinguish between cancer and non-cancer proteins, the local topological signatures did not confidently distinguish between types of cancers, which is arguably the more important problem.

Agrawal *et al.* gives a similar approach for identifying higher-order network structures and employing them in disease pathway discovery. They also considered the graphlet vector of each node as Milenković´ *et al.* did, but made this vector more fine-grained by considering all possible positions, or orbits, within the graphlets. That is, they considered exactly where the node was positioned in a graphlet for each graphlet. There are 73 orbit positions for the 30 distinct n-graphlets for $n = 2$ to 5, after reducing symmetrical positions to avoid double counting. They further showed that disease-linked proteins tend to have characteristic graphlet orbit vectors that make them easier to identify. In fact, considering orbit positions classifies most proteins into their specific corresponding disease category with significance, a result the topological signature approach did not achieve. Therefore, orbit signatures may prove to be a significant feature of disease-linked proteins.

Benson *et al.* [3] also designed an algorithmic framework to help understand the higher-order organization of complex networks. The algorithm in this paper provides an efficient way to investigate how motifs, or recurring small network subgraphs, cluster together in the larger network. In order to quantify the clustering problem, Benson et al. defined a metric called motif conductance, which is used to maximize instances of a motif inside a cluster and minimize those across cluster boundaries. This heuristic is an extension of the conductance metric for spectral clustering, which we learned in class. More precisely, given a motif, the algorithm aims to minimize the following for a cluster S:

$$\phi(S) = \frac{\text{cut}(S, \overline{S})}{\min\{\text{vol}(S), \text{vol}(\overline{S})\}}$$

Finding the exact cluster of nodes that minimizes this heuristic is a hard problem. The paper provides a framework to find a near-optimal solution with high computational efficiency. The algorithm extends techniques employed in spectral graph clustering, using eigenvectors and eigenvalues of matrices associated with motifs. This algorithm scales with large networks: for computing triangular motifs, it can handle billions of edges.

2.3 Critique

While empirically the results do seem to be better when employing higher-order connectivity, there is not a clear theoretical foundation for why this is the case. In general, there seems to be insufficient support for why certain methods prove useful - for example, optimizing for the recall-at-k objective in the case of Agrawal *et al.*'s research. Thus, we plan to investigate the significance of higher-order structures more deeply. One approach is to build upon looking at all possible orbitals for graphlets with small amounts of nodes ($n = 2$ to 5) by incorporating larger graphlets. To do this, we can remove smaller graphlets that do not prove to be useful in determining a disease signature. Furthermore, the papers discussed do not provide a systematic way to learn which motifs are important in an unsupervised manner. For the disease pathway discovery task, we sometimes do not know what motifs to look for the first place and hence cannot apply the algorithms discussed above directly. One possible improvement is to find a base set of motifs through literature research and identify other potential motifs in a bootstrapping manner. We could also explore all possible motifs from small graphlets and identify ones that are relevant to the disease and proteins at hand. In this paper, we will first establish baseline standards using local network features, then apply Benson *et al.*'s motif conductance minimization algorithm to the PPI network at hand. We will also test whether network-based methods may perform better than motif-based ones for certain disease categories given the diversity of disease pathway structures. Finally, we will implement some of the above proposed improvements to more sophisticatedly discover significant disease motifs.

3 Data

3.1 Datasets

SNAP has collected a group of three convenient datasets for this problem described below [1].

Human protein-protein interaction (PPI) network: A biological network with 21,557 proteins (nodes) and 342,353 interactions (edges). The graph is undirected and unweighted, but all protein-protein interactions have been verified experimentally.

Protein-disease associations A list of tuples (p, d) meaning that protein p is associated with disease d, for all known $p - d$ associations. These links come from DisGeNET and there are over 21,000 tuples, spread across 519 diseases each of which have at least ten associated proteins. Some more complicated diseases, however, have up to hundreds of protein associations.

Disease categories Second-level ontological categorizations (i.e. cancers, cardiovascular system diseases, etc.) of 290 out of the 519 diseases in the previous dataset. We can also obtain first-level disease categories by using the Disease Ontology and investigating each disease's Unified Medical Language System (UMLS) code.

3.2 Disease pathways

Formally, a disease pathway for disease d is a subgraph G_d of the PPI network consisting of all nodes n associated with disease d. Thus, G_d only contains edges connecting nodes both in G_d. All 519 pathways in the dataset have an average largest connected component size of 15.32, average fraction of nodes in largest connected component of 0.30, average density of 0.09, and average conductance of 0.54. The density of an undirected graph is $\frac{2|E|}{|V|(|V|-1)}$, a measure of how interconnected the graph is. The conductance γ of a cut in the graph, in our case $(G_d, \overline{G_d})$, is defined as follows:

$$\gamma(G_d, \overline{G_d}) = \frac{\sum_{i \in G_d, j \in \overline{G_d}} a_{i,j}}{min(E(G_d), E(\overline{G_d})},$$

where $a_{i,j} = 1$ if i, j have an edge between them and $E(G_s)$ is just the number of edges in the original graph G that have a node in G_s, the subgraph. As the equation suggests, conductance measures how isolated a disease pathway G_d is from the rest of the network (lower conductance value means more well-knit). An average disease pathway conductance of 0.54 means that a protein in a pathway is on average as likely to connect to another pathway protein as it is to connect to a protein outside of the pathway. This observation shows that the PPI network is quite interconnected between disease pathways, making the classification problem even more difficult.

3.3 Representative pathway statistics

We chose 5 representative disease pathways for testing: 1) Cardiomyopathies, 2) Cognition Disorders, 3) Pancreatic Neoplasm, 4) Diabetes Mellitus, Experimental, and 5) NADH:Q(1) Oxidoreductase deficiency. We tried to capture a diversity of network size, density, and conductance values in choosing these pathways. These pathways have an average largest connected component size of 41.40, have an average fraction of nodes in largest connected component of 0.60, average density of 0.13, and average conductance of 0.52. It is notable that the average values for density and conductance are very similar to those for all of the pathways. The average largest connected component size here is considerably greater than the average over all pathways because we wanted to have a sufficient set of seed proteins as training data. The specific statistics for each disease pathway are shown in the table below.

Disease	Proteins	Edges	Max node degree	Largest component	Density	Conductance
Cardiomyopathies	90	139	11	59	0.035	0.576
Cognition Disorders	23	12	3	3	0.047	0.434
Pancreatic Neoplasm	80	150	24	49	0.047	0.554
Diabetes Mellitus, Experimental	106	210	17	79	0.038	0.588
NADH:Q(1) Oxidoreductase deficiency	20	95	17	17	0.850	0.439

Table 1. General network statistics for the subgraphs of representative disease pathways.

4 Analysis methods

We use the following two evaluation metrics to measure the prediction quality of the algorithms implemented. Each metric takes the 3 inputs as follows and outputs an evaluation score proportionate to the accuracy of the algorithm's predictions.

1. S: a map from protein to score (higher score indicates higher likelihood of being a disease protein)
2. T: a set of known disease proteins to be predicted (not including seed proteins)
3. k: the maximum number of protein predictions to be considered

Normalized Discounted Cumulative Gain (NDCG): Discounted cumulative gain takes into account the cumulative accuracy of predictions. For a given i, the discounted gain at i is just the fraction of correct predictions out of the top i proteins as determined by S all divided by a $\log i$ factor in order to give greater weight to smaller i. We sum these values over i from 1 to k to get the cumulative evaluation DCG_k. To get $NDCG_k$, we just divide DCG_k by the ideal DCG_k, or $IDCG_k$, where the top i predictions are all correct for i from 1 to k. As shown, this metric measures prediction accuracy while putting greater emphasis on the top predictions. The formal definition of $NDCG_k$ is given below, where S_i is the set of the top i proteins based on their score in S.

$$NDCG_k = \frac{DCG_k}{IDCG_k} = \frac{\sum_{i=1}^{k} S_i \in T}{|S_i| \cdot \log_2(|S_i| + 2)}$$

Recall-at-k: This metric measures the overall accuracy of the top k protein predictions. Formally, recall-at-k is defined to be $\frac{|S_k \in T|}{min(k,|T|)}$ considering the cases where $k < |T|$.

5 Baseline: Local network-based algorithms

As our baseline approaches, we implement a variety of local network-based algorithms. We take these to be our baseline because they are more traditional approaches that just take into account a known disease protein's surroundings to make predictions on other proteins in the pathway.

5.1 Diffusion (random walk)

Diffusion-based methods set seed proteins as random walkers and determine protein predictions based on the frequency with which they are visited. Specifically, in our case, we set all seed proteins to be the starting walkers and for each seed protein, we choose a random neighbor to be the next walker. We then iterate over this new set of walkers and for each one choose a random neighbor to be the next walker. We continue this process for a certain number of iterations keeping track of the number of times each protein is visited. The score for each protein then is just its frequency of visits. This method hypothesizes that starting from seed disease proteins, other disease proteins will be visited more often than non-disease proteins in random walks starting from the seed proteins because disease proteins are assumed to be tightly clustered.

5.2 Mean direct neighborhood scoring

Traditional neighborhood scoring takes into account the number of neighbors of non-disease proteins that are disease proteins and extends the set of disease proteins by sampling from the weighted distribution of non-disease proteins created using this metric. The process is then repeated for a certain number of iterations and the score of each protein is based on the order in which they were added to the disease protein set. We modify this algorithm slightly in an attempt to extend our predictions beyond those with short path lengths to the seed proteins, downplaying the effects of direct protein adjacency. Our approach is mean direct neighborhood scoring. At each iteration, the score of each non-disease protein is updated by taking the average of the scores of its neighboring proteins. The initial disease seed proteins maintain a constant score of 1.0 and all other proteins start with a score of 0 at iteration 0. The final score of protein p after k iterations is $\sum_{i=1}^{k} \text{avg}(N_{i-1})$ where N_i is the set of scores of the neighbors of p after iteration $i-1$.

5.3 DIAMOnD

DIAMOnD, or DIseAse MOdule Detection Algorithm, is an algorithm part of a family of algorithms known as community detection algorithms similar to the direct neighborhood scoring approach. The algorithm takes advantage of the fact that the disease linked proteins in the protein-protein interaction network seem to be in the same neighborhoods.

The algorithm starts with a set of seed proteins that are known to be linked with a certain disease and tries to grow that set incrementally to be the set of proteins assumed to be disease-linked. At each iteration, all other proteins not in the set are given scores based on the probability that in a random graph (one where neighbors are distributed uniformly at random among all other nodes), at least k_i neighbors belong to the current set of disease-linked proteins.

More explicitly, if we assume the graph is random and there are s proteins currently in the disease-linked set, and a protein v has k neighbors, the probability that exactly k_i neighbors belong to the set is

$$p_v(k_i) = \frac{\binom{s}{k_i}\binom{N-s}{k-k_i}}{\binom{N}{k}},$$

where N is the number of proteins in the network.

We chose to approximate this value for the sake of efficiency as follows

$$p_v(k_i) \approx \binom{k}{k_i} \left(\frac{s}{N}\right)^{k_i} \left(1 - \frac{s}{N}\right)^{k-k_i}$$

During an iteration, if we know v has k_i neighbors in the set of disease-linked proteins, then in reality it has at least k_i disease-linked proteins as neighbors, so the probability, or p-value, is

$$\sum_{j=k_i}^{k} p_v(j).$$

At each iteration, we assign all proteins their p-values, and at the end we choose the most anomalous protein (lowest p-value) to add to our set of disease-linked proteins. The order in which the proteins were added to the set then serves as the final ranking of the proteins.

5.4 Results and discussion

We ran each of our algorithms on the set of diseases listed in 3.3. Diffusion was run for 1000 iterations, neighborhood scoring for 50 iterations, and DIAMOnD for 200 iterations. For each disease, we randomly chose 75 percent of the known proteins to be seed proteins (this is a lower percentage than that used in the project milestone to more robustly assess the predictive power of these methods). We evaluated our algorithms using NDCG-at-200 and recall-at-200.

Disease	Diffusion		Neighborhood		DIAMOnD	
	NDCG	Recall	NDCG	Recall	NDCG	Recall
Cardiomyopathies	0.009	0.087	0.002	0.043	**0.030**	**0.174**
Cognition Disorders	0.000	0.000	0.000	0.000	0.000	0.000
Pancreatic Neoplasm	0.009	**0.050**	0.000	0.000	**0.028**	**0.050**
Diabetes Mellitus, Experimental	0.000	0.000	**0.018**	**0.111**	0.006	0.037
NADH:Q(1) Oxidoreductase deficiency	0.000	0.000	**0.114**	**0.600**	0.000	0.000

Table 1. Prediction scores of the network-based algorithms applied to five disease pathways.

As shown in the table, the optimal algorithm for each disease given a particular evaluation metric was also the optimal algorithm given the other evaluation metric. Mean direct neighborhood scoring performed especially well on NADH:Q(1) Oxidoreductase deficiency because this disease pathway had a very significant connected component. Out of the 20 known proteins in this pathway, 17 are part of the largest connected component, which means that the disease proteins are all within the same neighborhood of each other. On the other hand, the diffusion algorithm performed optimally out of the three methods for cognition disorders and pancreatic neoplasm because these disease pathways were more spread out. The sparsity of the disease proteins was in this way better captured using a random walk method. Lastly, DIAMOnD likely performed optimally for cardiomyopathies and diabetes mellitus because the disease proteins were less spread out and each disease protein has a low p-value. As the table suggests, these algorithms do not predict disease pathway proteins very well in general and we need to explore other approaches to improve on pathway discovery.

6 Motif-based prediction

6.1 Model and Algorithm

We first try to use Benson *et al.*'s model for our motif-based prediction scheme. Given a motif M, the model aims to find a cut $(S, \bar(S))$ in the network that minimizes the ratio in 2.2. Intuitively, minimizing the ratio leads to a set S with nodes that occur in M many times and avoids splitting the nodes in M between the cut.

For each selected disease, we run a motif clustering algorithm on a series of motifs that we think might be relevant to the disease. Specifically, we follow the algorithm below:

1. Compute a motif "co-occurrence" matrix $\mathbf{A}$, where each entry $A_{i,j}$ is the number of times nodes i and j co-occurs in an instance of the motif M.

2. Compute the Laplacian matrix corresponding to the co-occurrence matrix $\mathbf{A}$. Specifically, compute the matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D}$ is a diagonal matrix whose elements satisfy $D_{i,i} =\#$ times node i appears in an instance of the motif M.

3. Compute a spectral ordering σ of all the nodes using the second smallest eigenvector v of the Laplacian matrix L. Here we use a technique from the Shi-Malik algorithm for spectral clustering commonly used in image segmentation.

4. Find the prefix set of σ with the minimum motif conductance. In other words, we look for

   $$\arg\min_i \phi_M(\{\sigma_1, \cdots, \sigma_i\})$$

   where ϕ_M is the motif conductance metric

   $$\phi_M(S) = \frac{\text{cut}_M(S, \overline{S})}{\min(\#\text{ motifs in } S, \#\text{ motifs in } \overline{S})}.$$

5. Evaluate the set using similar methods as before (NDCG-at-200 and recall-at-200).

6. Iterate this over all motifs to select the motif with the best NDCG and recall.

6.2 Approximation Algorithm

Finding the cluster with the global minimum motif conductance turns out to be too slow for our protein network. We then explore an approximation algorithm based on the approximate personalized PageRank method, except adapted for motif conductance. We use a modified version of an algorithm from Yin *et al.*'s work [8]. We look for a cluster with a local minimum motif conductance instead. Similar to before, we compute an ordering of the nodes and look for the prefix set with minimum motif conductance. In our experiments, we also start with a set of nodes, the seed nodes, to find a local cluster - hence we will build the ordering based on the existing seed nodes. We use the Approximate Weighted Personalized PageRank algorithm for the ordering of nodes, adapted for motifs:

1. Initialize the PageRank vector $p(v) \leftarrow 0$ for all nodes v.

2. Initialize a residual vector $r(u) \leftarrow \frac{1}{n}$ for all seed nodes u, where n is the total number of seed nodes, and $r(v) \leftarrow 0$ for all other nodes.

3. Compute the modified degrees $d(v) =\#$ motifs containing v for each node v.

4. While at least one node v has residual value $r(v)$ greater than $\epsilon d(v)$, perform updates to $p(v)$ using the original APPR algorithm (see Andersen *et al.*) for more details.

6.3 Results and discussion

We ran the above motif-based algorithm on the dataset for the following motifs: clique-3, clique-4, and clique-5. Clique-n is a complete graph on n nodes. We decided to test these motifs first because there is evidence in literature [1] that supports the significance of these motifs. For each disease, again we randomly chose 75 percent of the known proteins to be seed proteins and evaluated our algorithm using NDCG-at-200 and recall-at-200.

Disease	Clique-3		Clique-4		Clique-5	
	NDCG	Recall	NDCG	Recall	NDCG	Recall
Cardiomyopathies	0.041	**0.217**	0.020	0.174	**0.043**	0.174
Cognition Disorders	0.003	**0.167**	0.003	**0.167**	**0.007**	**0.167**
Pancreatic Neoplasm	0.000	0.000	0.000	0.000	**0.001**	**0.050**
Diabetes Mellitus, Experimental	**0.027**	**0.111**	0.019	0.037	0.001	0.037
NADH:Q(1) Oxidoreductase deficiency	0.000	0.000	0.000	0.000	0.000	0.000

Table 2. Prediction scores of the motif-based algorithm applied to five disease pathways.

The results for certain motifs are promising and outperform the corresponding results of the baseline methods. For example, clique-3 obtained a higher or equal NDCG and recall score compared to any of the baseline methods for cardiomyopathies, cognition disorders, and diabetes mellitus. However, one baseline result that stands out is mean direct neighborhood scoring on NADH:Q(1) Oxidoreductase deficiency, which has a recall score of 0.600. By contrast, none of the motifs used were able to predict any other known proteins. This dichotomy demonstrates that higher-order network algorithms may not always be superior to local network searches because certain disease pathways are clustered around a particular region in the network. This local centrality attribute thus actually favors baseline methods like neighborhood scoring over motif-based methods. However, to confirm this result, the motif-based algorithm would have to be run on the diseases over many other motifs. At the moment, it seems quite plausible that the predictive power of a particular method is based on the network structure of the disease pathway, which is only to say that a motif-based method may not always be the best solution to discover novel pathway proteins.

7 Conclusion

Community detection algorithms applied to the protein-protein interaction network proved to be useful in disease pathway classification. Many of the baselines techniques applied achieved results that took advantage of the fact that proteins that are part of the same disease are often times related to each other in some aspect of the structure of the graph. As the results showed, these results varied widely depending on the type of disease pathway involved. Higher order techniques, like the motif-based algorithm used, account for more general structural properties of a network and thus were able to perform, on average, better than the baseline techniques.

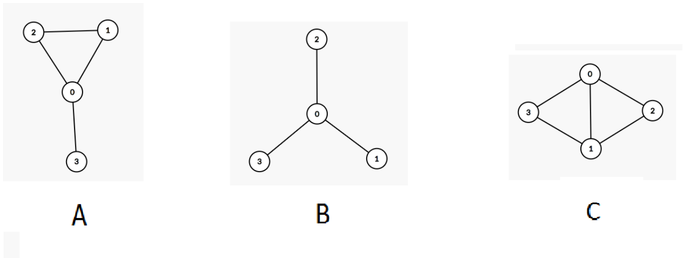

Figure 2: Future motifs to be explored on the human PPI network.

Furthermore, we learned from the above experiments that in general it is hard to predict characteristic features of disease pathways. However, we did notice that some methods performed significantly better on certain disease pathways than others, suggesting that perhaps an ensemble model could be effective. As such, we can learn representations for disease pathways to find which types of pathways each method performs well on. Then we could weight predictions from methods that are favored for a certain pathway more heavily. Another future direction this project could take is to explore further the potential of simple motifs, such as the ones in Figure 2. If these motifs prove to be useful in disease pathway discovery, the identification of new disease proteins becomes a much more efficient task, as these motifs only have four nodes. There is literature support for the importance of these motifs in the characterization of disease proteins, specifically for musculoskeletal system, nervous system, and cardiovascular system diseases [1].

A final future approach is to allow for multiple motif-based search. That is, disease pathway may be better characterized by a combination of different motifs as opposed to a single motif. Given several different motifs, we can consider a feature vector of the frequency of motifs incident to the protein in question and learn a classifier to discover proteins that are disease-linked. The advantage of this approach is that instead of making a broad assumption overall diseases that disease-linked proteins are more likely to be in the same neighborhood, we can learn more complex relationships across all diseases. Agrawal et al. explores this by taking all possible motifs of a small amount of nodes (5 or less) and using the frequencies of each motif as features for part of the signature to identify disease-linked proteins. Overall, we find all of these future approaches to be promising because of our findings that suggest higher-order network structures are important in disease pathway discovery. It remains to be determined how best to use these structures to effectively characterize disease, though network motifs should be strongly considered in any future solution.

8 Acknowledgments

We would like to thank Hao Yin for his assistance in optimizing the motif-based approximation algorithm as well as modifying the algorithm to enable multiple seed nodes.

References

[1] M. Agrawal *et al.*, “Large-Scale Analysis of Disease Pathways in the Human Interactome.” (2017).

[2] T. Milenković´ *et al.*, “Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data.” *Journal of the Royal Society Interface* **7**, 423 (2010).

[3] A. Benson *et al.*, “Higher-order organization of complex networks.” *Science* **353**, 6295 (2016).

[4] S.D. Ghiassian *et al.*, “A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome” *PLoS Comput Biol* **11**, 4 (2015).

[5] W. Kim *et al.*, “NemoProfile as an efficient approach to network motif analysis with instance collection” *BMC Bioinformatics* **18**, 423 (2017).

[6] J. Grochow *et al.*, “Network Motif Discovery Using Subgraph Enumeration and Symmetry-Breaking” *Research in Computational Molecular Biology* Lecture Notes in Computer Science, **4453** (2007).

[7] J. Shi *et al.*, “Normalized Cuts and Image Segmentation.”, *IEEE Transactions on PAMI*, **22**, 8 (2000).

[8] H. Yin *et al.*, “Local Higher-Order Graph Clustering.”, *KDD*, 555-564 (2017).

[9] R. Andersen *et al.*, “Local partitioning for directed graphs using PageRank.” *Internet Mathematics*, **4863** (2008).

Individual Contributions

We tried to divide the project as equally as possible between the three of us. Below are each of our individual contributions to the project.

Kevin Li: Implemented motif-based algorithm using MAPPR and wrote up respective part in the report

Glenn Yu: Implemented mean direct neighborhood scoring, collected data and summary statistics, coded NDCG evaluation metric, wrote up respective parts as well as project background and conclusion

Jeffrey Zhang: Implemented diffusion and DIAMOnD algorithms, coded recall evaluation metric and test data, wrote up the respective parts as well as part of the conclusion