**Approach**

The primary goal of this project is to determine the core proteins in the protein-protein interaction (PPI) network of exocytosis and endocytosis. In order to accomplish this goal, we will be using multiple strategies in order first infer the biological significance of the “core” proteins in this interaction network and then to determine what exactly these proteins are and what these protein’s functions are in relation to the processes of exocytosis and endocytosis.

*Data Generation*

First and foremost, we will generate a PPI network consisting of proteins that function in either exocytosis, endocytosis, or in both processes*.* First, the biologically relevant pathways (exocytosis and endocytosis) will be determined using the KEGG database(Kanehisa and Goto, 2000). Next, the proteins (and gene products) that make up these pathways will be identified and sent to the BIOGRID database (Chatr-Aryamontri et al., 2015). The interactions between these proteins will then be constructed into a single PPI network.

*Rich-Club Analysis*

In order to determine the biological significance of the “core” proteins in our full PPI network, it is necessary to utilize network analysis. One measure of determining the significance of the “core” nodes in an undirected network is called the rich-club coefficient. This measure is described by the equation,

(1)

where describes the value of the rich-club coefficient at the specified value of , is the number of nodes with degree greater than , and is the total number of edges of the nodes in the set . Mathematically, this coefficient becomes identical to the more commonly used clustering coefficient if it is only applied on the subset of nodes with degree greater than (Colizza et al.). However, there is an important failing of this measure since it is impossible to determine its significance without any comparison to some other network (Jiang and Zhou, 2008). An alternative rich-club coefficient has been developed, and utilizes the old version but also calculates it for “maximally random” networks, thus normalizing its value against all other networks with the same degree distribution (McAuley et al., 2007). This measure is described by the equation,

(2)

where is described in Eq. 1, and is calculated using Eq. 1 but over 100 random networks generated from the original network utilizing an edge-switching algorithm. This edge-switching algorithm preserves the same degree structure of the network while producing a random network, and is the standard for quickly generating random graphs with similar qualities to the original graph. Finally, is found by taking the mean of the values of for all 100 random networks. This allows calculation of the normalized rich-club coefficient (McAuley et al., 2007).

The next step in our analysis is to analyze the value of the rich-club coefficient function and to assess its value biologically. In most of the literature, a value of that is greater than 1 is thought to provide significant evidence for the existence of a “rich-club core” at that degree level. However, some doubts remain about the validity of this measure (Jiang and Zhou, 2008). In order to appease these doubts, we will use a bootstrap method of resampling against multiple sets of randomly generated graphs in order to assess its significance(Jiang and Zhou, 2008; Wuchty et al., 2009). Similarly, the value of the rich-club coefficient is generally considered to be monotonic as the values of increase, and we will look at this graph, as it will help us understand any biological significance to this measure. Finally, by analyzing this graph of the values of the rich-club coefficient, we will be able to determine if there is a rich-club core in the PPI network.

*Identification of Structurally Dominant Nodes*

Our next goal is to identify the most structurally dominant nodes in the PPI network. After thoroughly reviewing the literature, we have found that there is no single method for identifying structurally dominant nodes in a network(Pei Wang et al., 2014). Although some single measures describing the network (ie degree, PageRank, clustering coefficient) have been proposed, they all seem to fall short in various ways. Because of this shortfall, we will utilize two different measures in order to serve as both computational validation as well as a second check on our analyses. These two measures are the *First Principal Component* score and *core/periphery analysis* scoring.

FPC Scoring

The First Principal Component scoring method is a relatively new method of identifying structurally dominant nodes in networks (Pei Wang et al., 2014). It broadly acts as an integrative measure that brings together eight different network description measures in order to generate a score of the most structurally dominant nodes in the network. The eight measures that it integrates are as follows: degree, clustering coefficient, closeness, k-shell, eigenvector centrality, semi-local centrality, and network motif centrality. These are all integrated into the FPC score, which is described by the following equation,

where is the total score for the node, is one of the eight measures involved in the score and is that specific measure’s weight. In Eq. 3, the weights, , are generated through maximizing the weights against the covariance matrix of all measures that make up . We will then rank all nodes in our PPI network with this score, and measure the top 10 nodes and generate a subgraph consisting of just their interactions(Pei Wang et al., 2014).

Core/Periphery Analysis

Core/periphery analysis is an older and somewhat more established method of identifying “core” in both directed and undirected networks (Borgatti and Everett, 2000). We will use it primarily to verify the results from the FPC scoring method. Even though the core/periphery score is not necessarily an integrative method of different network descriptors, we expect the score (and resulting ordering) to be similar to the FPC score. The total core score of a node, , is given by the equation,

where is a normalization score assigned after the fact in order to make the maximum core score equal to 1, is a parameter that describes the sharpness (or fuzziness) of the difference between core and periphery (is the sharpest), is a parameter that describes the size of the core as a percentage of the total nodes in the network, and

Finally, is the set of all neighbors of , so that we are also calculating the score for all of the neighbors of as well as for itself(Csermely et al., 2013; Rombach et al., 2014). After completing this score for all nodes in the network, we will then rank the nodes by increasing score, and measure the top 10 nodes in the network. Finally, we will compare the scores between the core/periphery score and the FPC score to serve as a quick validation measure of our methodology.

Validation

Since this section focuses solely on computationally determining what the most significant proteins in the combined endocytosis-exocytosis PPI network, there is a large need for validating our procedure in a more biological fashion. Thus, we have multiple validation checks on our analysis.

To serve as our primary validation measure, we will utilize a method similar in concept to leave one-out cross-validation (LOOCV). In our implementation of LOOCV, we will iterate through each node in the network, removing it while performing both FPC and core/periphery scoring (James et al, 2013). We will note the ranking of the nodes via each method while iterating through the different nodes. Finally, we will sum and normalize the together back into a single ordered list of the scores for each node for both of our measures of structural dominance in the network.

Next, we will separate the combined network of endocytosis and exocytosis into networks consisting of proteins that are annotated as only acting in endocytosis or in exocytosis. We should note that proteins that are involved in both networks will still be involved in both networks, but that proteins only functioning in exocytosis would only be in the exocytosis network. We will perform the same analysis that is described above in an attempt to test if nodes that are listed as structurally dominant in the combined network are actually structurally dominant in the segregated networks or if they are merely considered dominant because they act in both networks. A similar cross-validation procedure as is performed in the main analysis will also be done in order to assess the significance of these results. Finally, the calculated structurally dominant nodes in the combined network will be compared to the top structurally dominant nodes in the combined network.

Finally, we will also test if our method privileges proteins that act over a wide area in the cell, or if proteins that function in a certain organelle (ie the endoplasmic reticulum) are privileged over proteins in another location. Since the locations of where proteins act are also available on KEGG, we will use this data to annotate and then separate the PPI network into separate subnetworks for each different location within the cell. Then, we will perform the same analysis as is performed on the combined network. A similar cross-validation procedure will be done in order to determine the significance of our results. Finally, the nodes from the combined network will be compared with the structurally dominant nodes in the subnetwork in order to test if our analysis privileges proteins that act solely over a wide are in the cell.