**Methods**

**Underlying Principles**

The primary premise of this study is that there are a set of core driver proteins that are highly influential in the processes of exocytosis and endocytosis. These core driver proteins should be found in the pathways of both exocytosis and endocytosis, thus showing that their importance in both pathways. Similarly, it should be possible to identify these proteins given a protein-protein interaction (PPI) network by quantifying the strength of the protein’s connections.

**Data Generation**

In order to generate our datasets, we utilized the human BIOGRID database.12 First, we searched the entire GeneOntology database to find proteins annotated to have function in exocytosis and endocytosis.3 Then, we identified the interactions of these proteins from BIOGRID and used those proteins and all proteins that interact with the original set of proteins to generate the full exocytosis and endocytosis network (full). Next, we filtered this network to include only the exocytosis PPIs (exocytosis). Finally, we filtered the full network to include only the endocytosis PPIs (endocytosis).

**Rich-Club Coefficient Analysis**

We first wished to test the broad hypothesis to determine the existence of a “rich-club” in the full (combined exocytosis and endocytosis PPI network). The primary method to test the existence is that of the rich-club coefficient, which accurately describes the “rich-club” phenomenon in complex networks (central and dominant nodes form interconnected communities with each other).4 The rich-club coefficient can be given by the equation

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 . In order to appropriately normalize the rich-club coefficient, it is necessary to divide this measure by a set of random graphs with the same degree distribution, such that

(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 . Rich-clubs are detected when the ratio , showing a tendency of large degree nodes in the network to be well-connected to each other.56

**Determination of Structurally Dominant Nodes**

After determining the existence of the rich-club in our PPI networks (full, exocytosis, and endocytosis), we wished to determine the structurally dominant (i.e. in the rich club) nodes in the PPI networks. After conducting a comprehensive literature search, it seems like there is no appropriate consensus method. Thus, we decided to utilize two proposed methods and compare them: First Principal Component (FPC) scoring and core/periphery analysis scoring.Our primary workflow takes as input a protein-protein interaction network and then runs either FPC scoring or core/periphery analysis in order to rank the nodes in the network.

*First Principal Component Scoring*

The FPC scoring method is a relatively new method of identifying structurally dominant nodes in complex networks that acts in an integrative fashion bringing together eight different network description measures. These eight measures are as follows: degree, betweenness, clustering coefficient, closeness, k-shell decomposition, eigenvector centrality, semi-local centrality, and network motif centrality. These are all well-defined and are often evaluated and used often for identification of centrality of nodes in networks.[[1]](#footnote-1) Next, these methods are summed into the FPC score which is described by,

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 .7 In our implementation of this method, we ran this measure on all three PPI networks that were generated.

*Core/Periphery Analysis Scoring*

Core/periphery analysis is a more established method of identifying the “core” proteins in both directed and undirected networks.9 Instead of being integrative, it identifies the probabilities of being in the core solely from the size of the predicted core as well as the number of neighbors (degree of the node). The total core score (probability of node being in the core) 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 the score is also calculated for all of the neighbors of as well as for itself. This allows somewhat efficient implementation of the method.

(5)

However, the procedure for this method actually begins by maximizing the core quality for the given parameters and via simulated annealing.10 The starting point for the optimization is a random shuffle of the core defining function . This optimization allows determination of what nodes are most likely to be in the network for the parameters given. Then the values are summed together amongst all parameters for both the node given and all neighbors of the node in order to create a score. Finally, this score is normalized so that the highest score is equal to one (100% probability of being in the core of the network). This score can then be reported as a ranked list of nodes in the network.1112

**Validation Procedure**

In this project, the validation of the results are built into essentially every stage of the analysis. First, we compared the results for each network (full, endocytosis, and exocytosis) for consistency of results. Another useful validation procedure was done by comparing the results of the two different analysis procedures. This was done via finding a simple correlation coefficient between the results of each analysis for each network. Utilizing the correlation coefficient in this manner allows determination of the similarity of the results even if the FPC score and the core/periphery analysis score have very different ranges.

However, the primary validation procedure was based around resampling, and was the same for each network and for both analysis types. First, one node was removed from the network and the same analysis (FPC or core/periphery) was repeated on the network missing that node. This step was repeated across all nodes in the network. The second step averages together all of the scores for the repeated observations and ranks the nodes by averaged score. Finally, this new score was compared against the original scored list via finding the correlation between the two scores.

A final validation procedure was based around testing if the dataset utilized included all of the proteins that could potentially be involved in both endocytosis and exocytosis. This is actually a major issue, as manual checking of the exocytosis and endocytosis networks revealed that several proteins that were missing from these networks, despite being listed in several papers on their function in these processes. Chief among these was dynamin-2, a protein that has been shown to act in both exocytosis and endocytosis. However, the Gene Ontology database includes dynamin-2 as acting in only endocytosis, and not exocytosis. While utilizing the full network probably resolved this issue, we believed that there might be proteins absent from the full network that actually do play a role in either exocytosis or endocytosis. Thus, a larger network based on the entire process of vesicle-mediated transport was generated and was utilized to compare the identified structurally dominant nodes as identified via FPC analysis against the full PPI network.

References

1. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* **43,** D470–478 (2015).

2. Stark, C. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34,** D535–D539 (2006).

3. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

4. Colizza, V., Flammini, M., Serrano, A. & Vespignani, A. Detecting rich-club ordering in complex networks. *Nat. Phys.* **2,** 110–115

5. McAuley, J. J., da Fontoura Costa, L. & Caetano, T. S. Rich-club phenomenon across complex network hierarchies. *Appl. Phys. Lett.* **91,** 084103 (2007).

6. Wuchty, S., Adams, J. H. & Ferdig, M. T. A comprehensive Plasmodium falciparum protein interaction map reveals a distinct architecture of a core interactome. *PROTEOMICS* **9,** 1841–1849 (2009).

7. Pei Wang, Xinghuo Yu & Jinhu Lu. Identification and Evolution of Structurally Dominant Nodes in Protein-Protein Interaction Networks. *IEEE Trans. Biomed. Circuits Syst.* **8,** 87–97 (2014).

8. Wang, P., Lü, J. & Yu, X. Identification of Important Nodes in Directed Biological Networks: A Network Motif Approach. *PLoS ONE* **9,** e106132 (2014).

9. Borgatti, S. P. & Everett, M. G. Models of core/periphery structures. *Soc. Netw.* **21,** 375–395 (2000).

10. Kirkpatrick, S., Vecchi, M. P. & others. Optimization by simmulated annealing. *science* **220,** 671–680 (1983).

11. Rombach, M. P., Porter, M. A., Fowler, J. H. & Mucha, P. J. Core-Periphery Structure in Networks. *SIAM J. Appl. Math.* **74,** 167–190 (2014).

12. Csermely, P., London, A., Wu, L.-Y. & Uzzi, B. Structure and dynamics of core/periphery networks. *J. Complex Netw.* **1,** 93–123 (2013).

1. These measures can be defined as follows: Degree: number of edges connected to a node ; Betweenness centrality: , where is the number of shortest paths between and , is the number of those paths through ; Clustering Coefficient: , where is the number of subgraphs with 3 edges and 3 vertices that include and is the number of subgraphs that can be formed with two edges and 3 vertices that include ; Closeness: , where is the distance between nodes and ; K-Shell Decomposition: the ordered maximal subgraph in which each vertex has at least degree , such that the coreness of belongs to the -core but not to the -core; Eigenvector Centrality: for each node defined as its component of the maximal eigenvalue’s eigenvector of the adjacency matrix; Semi-Local Centrality: , where , where is the number of nearest and next nearest neighbors of and is the set of nearest neighbors of ; Network Motif Centrality: a network motif based measure for calculating node importance.7 8 [↑](#footnote-ref-1)