**Identification of Core Proteins in Endocytosis and Exocytosis**

**Abstract**

Exocytosis and endocytosis are very important for the cell activity. There are many unsolved questions in the secretory pathway regulation, especially the regulation between exocytosis and endocytosis. In this paper, we want to find the core proteins in the exocytosis and endocytosis. We set up four different networks, which are exocytosis network, endocytosis network, exocytosis or endocytosis network and vesicle network. We rank all the proteins in the network with both FPC score and core/periphery scoring and analyzed the top five proteins. We can find the core proteins in the integrity of the network (exocytosis or endocytosis) can be easily omitted, which means these proteins are not core proteins in exocytosis network and endocytosis network separately.

**Introduction**

Within and around cells, materials are constantly being shipped one way or another across membranes. At any given moment, a particular substance may transport from one side of the membrane to the other side. This transport is very important to cellular function, and the disorder of transport system in the cell causes many diseases, such as Wilson Disease. Cellular transport contains many different kinds of process, such as diffusion, active transport and electron carriers.1 However, exocytosis and endocytosis are the most essential part of the cellular transport system.

Exocytosis is the durable, energy-consuming process by which a cell directs the contents of secretory vesicles out of the cell membrane and into the extracellular space. The membrane-bound vesicles contain soluble proteins will be secreted to extracellular environment, membrane proteins and lipids will be sent to the cell membrane. There are basically five steps for the exocytosis, which are Vesicle trafficking, Vesicle tethering, Vesicle docking, Vesicle priming and Vesicle fusion.2

Endocytosis is an energy-using process by which cells absorb molecules (such as proteins) by engulfing them. It is used by all cells of the body because most substances large polar molecules that cannot pass through the cell membrane. Endocytosis is more widely used than exocytosis in the cells. Endocytosis can be divided into four different subtypes: clathrin-mediated endocytosis, non-clathrin-mediated endocytosis, macropinocytosis and Phagocytosis. Clathrin-mediated and non-clathrin-mediated endocytosis are involved in the receptors on the cell membrane pathway. Macropinocytosis is the process of forming a pocket in the cellular membrane which then fills with extracellular fluid and molecules. Phagocytosis is the process by which cells bind and internalize particulate matter larger than around 0.75 µm in diameter via the usage of vesicles.3 Therefore, endocytosis is a widely used process within the cell.

Exocytosis and endocytosis can be involved in the same cell in the same pathway. One example is in the synaptic signaling pathway. Synapses are very important for the human body, especially for the signal transmission in the neuron system. Exocytosis and endocytosis are the most important processes in the vesicles containing neuron-transmitters release in the pre-synapse, which are responsible for the signal transmission in the synaptic junction. Specifically, during the neuron-transmitters release process, the first step is the fusion of the vesicle to the membrane in the pre-synapse, while the second step is the re-sealing of the vesicle after releasing the signaling molecule, and the third and final step is the return of the vesicle to the cytoplasm.3 Thus, endocytosis occurs after exocytosis in the same pathway. Therefore, exocytosis and endocytosis are tightly related to each other. However, the full extent of this relationship is not understood. We therefore propose to utilize new methods to more fully investigate this relationship.

Exocytosis and endocytosis can take place in many different parts of cell, such as the endoplasmic reticulum, the Golgi apparatus, the endosome, the lysosome and the cell membrane. However, it is not known what the most important location for the connection between exocytosis and endocytosis is in the cell.4 Therefore, we also would like to investigate the most important part of the cell for the regulation between exocytosis and endocytosis.

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.

**Methods**

*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).5 The rich-club coefficient can be given 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 . 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.6,7

*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,

(3)

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 .8 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.10 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,

(4)

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.11 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.12,13

*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.

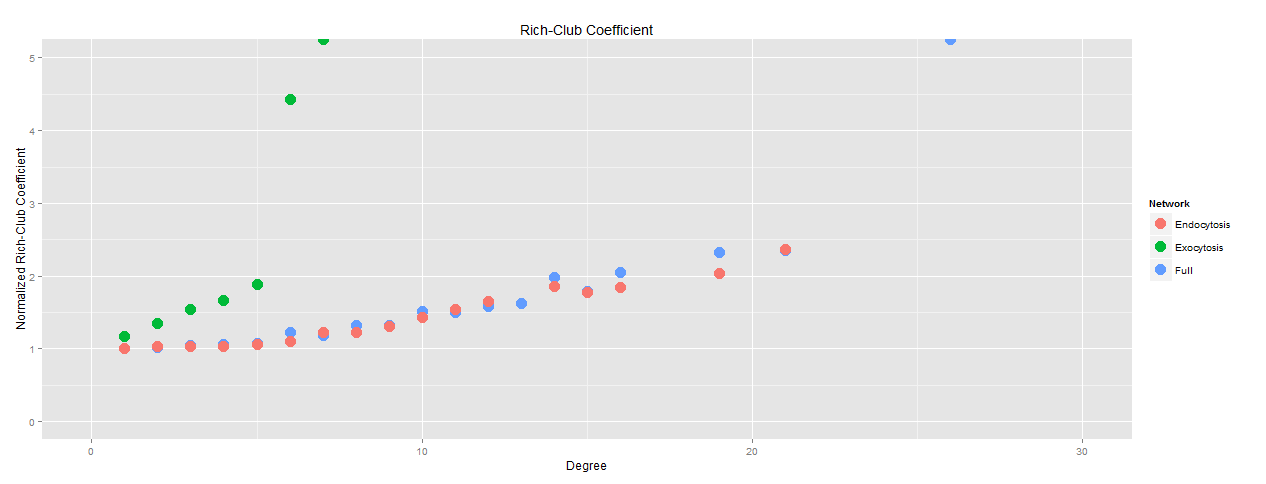
A final validation procedure was based around testing if the dataset utilized included all of the proteins that could potentially be involved in both endocytosis and exocytosis. This is actually a major issue, as manual checking of the exocytosis and endocytosis networks revealed that several proteins that were missing from these networks, despite being listed in several papers on their function in these processes. While utilizing the full network probably resolved much of 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.

**Results**

*Data Generation*

In order to generate our datasets, we utilized the human BIOGRID database.14,15 First, we searched the entire GeneOntology database to find proteins annotated to have function in exocytosis and endocytosis.16 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).

*Endocytosis network*

 We first analyzed the endocytosis PPI network, as identified by BIOGRID and Gene Ontology. In order to show the existence of a rich-core structure in the networks, the calculated normalized rich club coefficient should be greater than 1 for all degrees as well as nearly always increasing as degree increases.5,7 The endocytosis network that was identified shows this, as is shown in Figure 1. Both the FPC score and core/periphery analysis rankings were utilized in order to rank the “structurally dominant” nodes. The top five proteins, as determined by the FPC score, are as follows: TNK2 (activation of CDC42), CDC42 (GTPase of the Rho family), SRC (cell growth), GRB2 (EGF receptor), EGFR (EGF receptor). Figure 1 shows the full endocytosis network with the top FPC scoring proteins shown in red. Similar results to the FPC were found from the core/periphery analysis scoring.

*Figure 1: Normalized rich-club coefficient plotted against degree*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Full network** | | EGFR | GRB2 | SRC | LYN | COPS5 |
| **Location** |  | membrane | membrane | membrane | membrane | ER |
| **Exocytosis network** | | RAB27B | SYTL4 | STX1A | STXBP1 | RPH3AL |
| **Location** |  | membrane | membrane | membrane | membrane | membrane |
| **Endocytosis network** | | EGFR | GRB2 | SRC | CDC42 | TNK2 |
| **Location** |  | membrane | membrane | membrane | membrane | membrane |

*Table 1 Top 5 nodes as identified by FPC score and their location in different networks*

We notice the top 5 FPC score proteins in the endocytosis network are all located in the membrane. And all the proteins are associated with membrane signal receptor, such as EGF. These results are somewhat different from our expectation.

*Exocytosis network*

Next, we set up the exocytosis network from the Gene ontology and BioGRID database, then we analyze the database by all the same methods as in the endocytosis network. First, we can show that there is a rich-core to the network via the features of the rich-club coefficient as a function of degree (monotonically increasing and greater than 1). The normalized rich-club coefficient is plotted as a function of degree in Figure 1. Again, we rank the top 5 FPC scoring proteins in the exocytosis network. The top five proteins from this method are RPH3AL (Rab GTPase effector), STXBP1 (Syntaxin-binding protein 1), STX1A (vesicle fusion process), SYTL4 (Synaptotagmin-like protein 4), RAB27B (vesicle fusion process). These five nodes are shown in red in Figure 2, which shows the entire exocytosis protein network, as derived from Gene Ontology and BIOGRID. Again, similar results are found from the core/periphery analysis procedure. We also notice the top 5 proteins are all belong to SNARE complex, which is responsible for the vesicle fusion to the cell membrane.

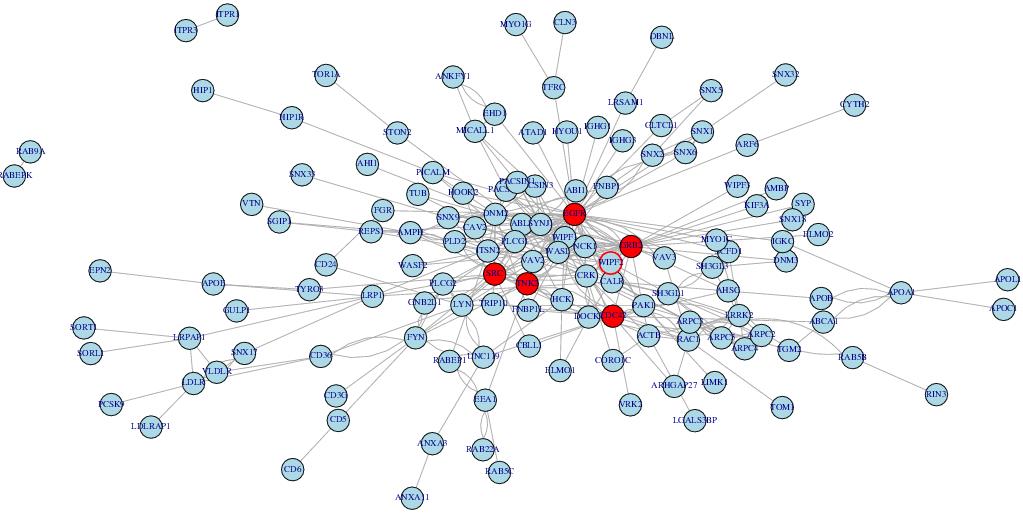


Figure 2: Endocytosis network with top 5 nodes as identified by FPC score marked in red

*Exocytosis or Endocytosis network*

We want to find the overlap of the endocytosis and endocytosis network. However, the identified overlap network is very small, and attempting to find a core for this network would be useless. Another issue is that the proteins derived from Gene Ontology is very limited, as there are many famous regulatory proteins, such as dynamin and sydapin, that are absent from either one or both endocytosis and exocytosis. Thus, we broadened our database to explore proteins implicated in either exocytosis or endocytosis. Surprisingly, the top scoring proteins are most similar to the endocytosis network, possibly showing the effects of large degree nodes on these methods.

We set up the exocytosis or endocytosis network from the Gene ontology and BioGRID database, and then we analyzed the network by our methods. First, we also determine the existence of a rich-core in this network as well via the normalized rich-club coefficient, also shown in Figure 1. Again, we rank the top scoring proteins in the full network. The top five proteins from the FPC scoring method are COPS5 (regulator of E3 ligase), LYN (association with receptor), SRC (cell growth), GRB2 (EGF receptor), EGFR (EGF receptor). Similar results are also seen from the core/periphery analysis, which is shown in Table (?).It is also worth noticing that COPS5 is a regulator of E3 ligase and located in ER, which is different from other proteins, as all the other proteins are located in the cellular membrane. From this point, we can say our method really work, because when we connect both exocytosis and endocytosis network, we can find some proteins can be easily omitted.

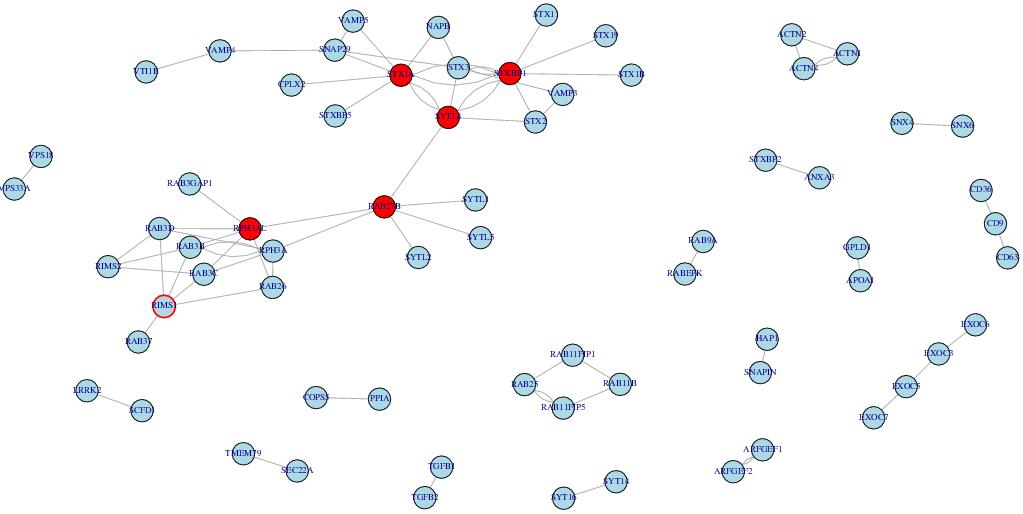


Figure 3: Exocytosis network

*Vesicle network*

Since we find that Gene Ontology is incomplete for exocytosis and endocytosis, it is necessary to look for other data sources. Unfortunately, a full database consisting of exocytosis and endocytosis pathways and protein interactions is not available. However, it should be possible to utilize the vesicle network to show that Gene Ontology is unreliable, as it contains more proteins that may be relevant to both exocytosis and endocytosis. We set up the vesicle network from the Reactome Pathway Database17,18 and get all the protein-protein interactions for this network from the BioGRID.14 Then we analyzed the network by FPC scoring, to rank the top 5 FPC score proteins in the vesicle network. The top five proteins from the FPC scoring are ALB (binding to water, Ca(2+), Na(+), K(+)), CLTC (Clathrin heavy chain 1), ARRB1 (Arrestin beta 1), YWHAQ (mediate signal transduction), UBC (ubiquitin).

We can see the top 5 proteins in the vesicle network make more sense than the network derived from Gene ontology database, and that these five proteins are highly involved in the exocytosis and endocytosis. For example, UBC is ubiquitin being involved in all the proteins degradation, which is widely used in the cell activity.

*Validation of Results*

We validated our results via two different methods. The first involved finding the correlation between the FPC score and the core/periphery analysis score. For each different network, we also show the relationship between the two scoring methods as well in Figures 5-7. The correlations between the two scoring methods are shown in Table 2. It is worth noting that there are significant correlations between the two measures – FPC scoring and core/periphery scoring.

The second validation method involved comparing the resampled scores against the scores of the complete network. For each different network, the correlation was recorded for the resampled scores; these correlations are also shown in Table 2. Judging from the correlations, it seems like the methods were implemented correctly and the results are likely to be biologically valid as there are very high correlations for all of the completed correlations.

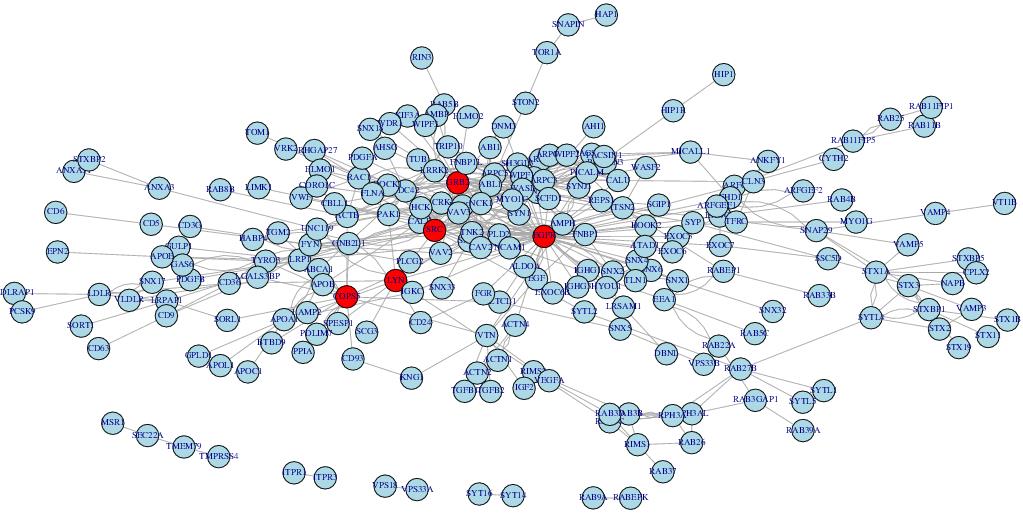


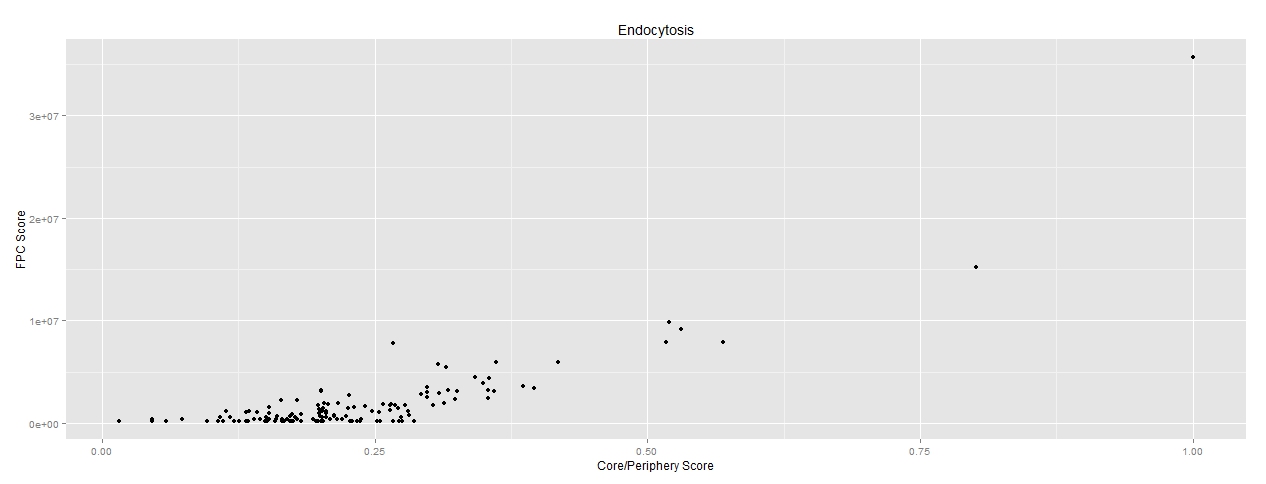
Figure 4: Exocytosis or endocytosis (full) network; top 5 FPC scored nodes are shown in red

**Discussion**

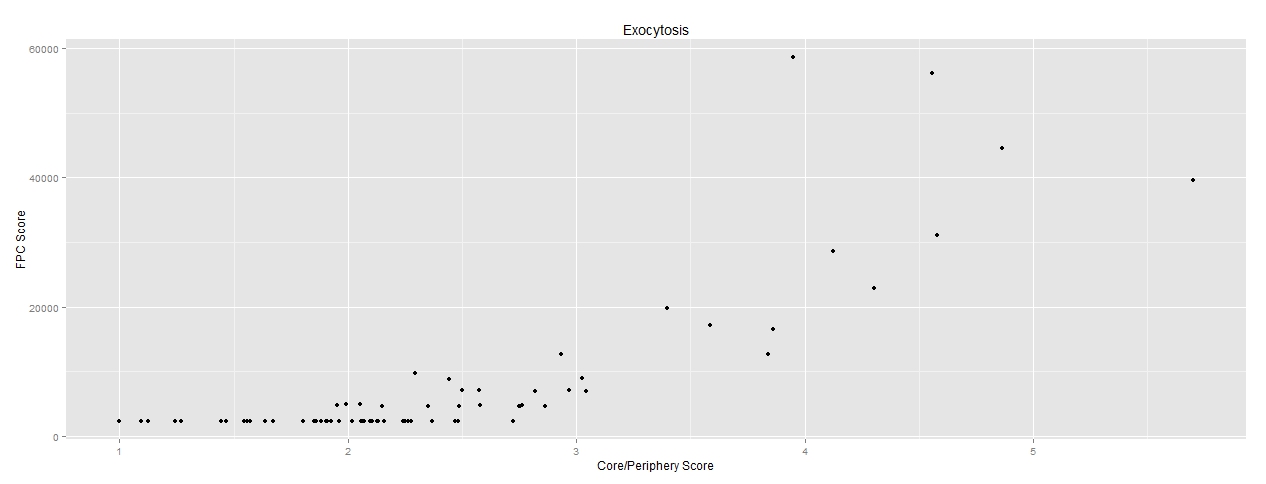
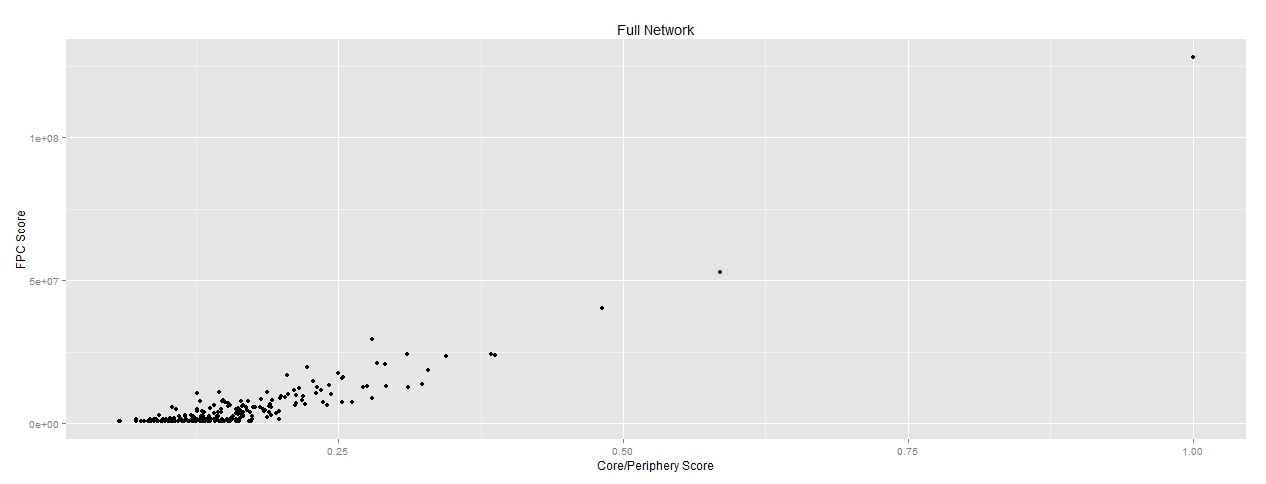
We set up the exocytosis network from the Gene ontology and BioGRID database, and then we analyze the database by our methods. We rank our proteins in the network with FPC score. The top 5 proteins are all belong to SNARE complex, which are responsible for the vesicle fusion to the cell membrane. The SNARE complex is very important to the exocytosis, which makes our analysis meaningful.

Table 2: Correlation coefficients showing the association for each different method and each network. Also see Footnote 2.

|  |  |  |
| --- | --- | --- |
| Correlation Coefficient | Comparison | Network |
| 0.924508103 | FPC, C/P | Full |
| 0.852576907 | FPC, C/P | Endocytosis |
| 0.820836959 | FPC, C/P | Exocytosis |
| 0.999999322 | FPC, FPC Validation | Full |
| 0.999997786 | FPC, FPC Validation | Endocytosis |
| 0.999994891 | FPC, FPC Valation | Exocytosis |
| 0.8290775 | C/P, C/P Validation | Full |
| 0.810518922 | C/P, C/P Validation | Exocytosis |
| 0.693445 | C/P, C/P Validation | Endocytosis |



*Figure 6: Endocytosis network FPC score vs Core/Periphery score*

 We notice that for the endocytosis network, the top five FPC score proteins are all involved in the signal receptors, such as EGFR. These results are outside of our original expectation, because these signal pathways don’t take part in endocytosis? other than in regulation. However, looking closely at the structure of the network as well as what is known about endocytosis resolves this issue. It is well-known that endocytosis has four different subtypes: clathrin-mediated endocytosis, non-clathrin-mediated endocytosis, macropinocytosis and phagocytosis. Clathrin-mediated endocytosis has been well studied for many years, so most of the proteins in the endocytosis network come from this pathway. Similarly, signaling cascades have been highly implicated in clathrin-mediated endocytosis pathway. Therefore, clathrin-mediated endocytosis accounts for the most important part in our endocytosis network, as well as the difference in the results from the expection

*Figure 7: Exocytosis Network FPC Score vs Core/Periphery Score*

*Figure 7: Full Network FPC Score vs Core/Periphery Score*

We have only 20 proteins both in exocytosis network and endocytosis network, which is far less than our expectation. This is clearly due to the limitations of existing pathway databases. For example, the roles of dynamin and synapin in both endocytosis and exocytosis are well-studied. However, neither protein is in the exocytosis pathway in the dataset that comes from Gene Ontology. There does not seem to be any method to resolving this, as there is no exocytosis pathway in the KEGG database (<http://www.genome.jp/kegg/>). One possible method of attempting to resolve this issue is to combine the networks of endocytosis and exocytosis.

Utilizing this combined network gives some interesting results. Interestingly, we find the COPS5 in the top 5 score from the FPC method. COPS5 is a regulator of E3 ligase, which can regulate many proteins expression. COPS5 FPC score ranks very low in exocytosis network and didn’t appear in endocytosis network. However, when combine these two network, COPS5 becomes to be the most core proteins in the network. In another word, COPS5 have potential to be a very important proteins in the regulation of exocytosis and endocytosis. Here we can see the efficacy of the methods used to find this easily omitted protein.

Finally, we set up the vesicle network, trying to find the most complete database. Even though we couldn’t distinguish the exocytosis and endocytosis in the vesicle network, we can prove the limitation of the network we set up before. We also have the top 5 rank proteins, which are ALB (binding to water, Ca(2+), Na(+), K(+)), CLTC (Clathrin heavy chain 1), ARRB1 (Arrestin beta 1), YWHAQ (mediate signal transduction), UBC (ubiquitin). These five proteins are most widely used in the cell. For example UBC is the necessary for all the proteins degradation in the cell, which is the most essential protein in the cell activity.

As a final conclusion, we find that the methods utilized in this project worked well. However, the data itself, despite massive advances, may limit the practicality and usefulness of projects such as these. Despite this limiting factor, it does seem as if the proteins that we identified do play a large role in controlling both endocytosis and exocytosis. Further research is needed in all of these areas in order to increase the accuracy of the results.

**References**

1. Jiang, C. & Salton, S. The role of neurotrophins in major depressive disorder. *Transl. Neurosci.* **4,** (2013).

2. Taguchi, T. Emerging roles of recycling endosomes. *J. Biochem. (Tokyo)* **153,** 505–510 (2013).

3. Haucke, V., Neher, E. & Sigrist, S. J. Protein scaffolds in the coupling of synaptic exocytosis and endocytosis. *Nat. Rev. Neurosci.* **12,** 127–138 (2011).

4. Ivanov, A. I. in *Exocytosis and Endocytosis* (ed. Ivanov, A. I.) **1174,** 3–18 (Springer New York, 2014).

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

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

7. 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).

8. 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).

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

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

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

12. 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).

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

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

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

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

17. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42,** D472–477 (2014).

18. Milacic, M. *et al.* Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers* **4,** 1180–1211 (2012).

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.8,9 [↑](#footnote-ref-1)