Core proteins in exocytosis and endocytosis

Dong Liu and Wesley Maddox

Project Aims

The secretory mechanism is a very important cell function, especially for the secretory cells, such as chromaffin cell and goblet cells. Besides, secretory pathway plays a very important role in the neuron signal transport. There are also many diseases related to dysfunction in secretory pathway (Salton SR et al, 2013). Overall, secretory mechanisms are very important for cell, organism and human body. However, there are many unsolved questions in the secretory pathway regulation, especially the regulation between exocytosis and endocytosis. Exocytosis is a process in which an intracellular vesicle moves to the plasma membrane and subsequent fusion of the vesicular membrane and plasma membrane ensues. Endocytosis is an energy-using process by which cells absorb molecules by engulfing them. Endocytosis is used by all cells of the body because most substances important to them are large polar molecules that cannot pass through the hydrophobic plasma or cell membrane (Taguchi T, 2013). Per the definition of exocytosis and endocytosis, we can say endocytosis is the opposite process of exocytosis. There are many cross linkages between exocytosis and endocytosis. First, exocytosis and endocytosis share many proteins and molecules during their regulation pathways. Second, both exocytosis and endocytosis are involved in secretory pathway, their main function is transport, these two process transport many same or associated molecules. Third, many organelles are involved in both exocytosis and endocytosis, such as ER, Golgi, endosome and lysosome. However, the regulation between exocytosis and endocytosis is very complicated. There are more than 200 different proteins in the regulation of exocytosis pathway. Therefore, we want to find the core proteins in exocytosis and endocytosis, trying to give a better understand of regulation between exocytosis and endocytosis.

We aim to find the core proteins in the exocytosis and endocytosis. We will search all the proteins involved in exocytosis and endocytosis pathway in KKEG website and search every protein's interaction from protein-protein interaction website, then build up the protein-protein interaction database in exocytosis and endocytosis. First, we will utilize the rich-club coefficient to determine if our network has a "rich-core". Second, we will use multiple methods to determine the core proteins. Besides, we will build up the spatial connection among these core proteins to achieve a complete picture of core proteins in the exocytosis and endocytosis.

Exocytosis and endocytosis are responsible for the regulation of synapse function, which is the main fundamental basics for the neuron signal transport in body. However, many questions remain unsolved in the regulation between exocytosis and endocytosis. There are more than 200 different kinds of proteins in the exocytosis and endocytosis pathway, which make the regulation of exocytosis and endocytosis very complicated. Therefore finding the core proteins in the exocytosis and endocytosis provides us a systematic picture to illustrate the regulation between exocytosis and endocytosis, which will lead us to find the important proteins in the regulation between exocytosis and endocytosis. We will use the rich-club coefficient to determine if our network has a "rich-core". This method is first time employed for the analyze core proteins in the exocytosis and endocytosis, which will provide us a better understanding of the regulation between exocytosis and endocytosis.

Significance

We aim to find the core proteins in the exocytosis and endocytosis from this project. We notes that since the first exocytosis event was found in 1963, more than 50 years has passed (Guillery RW et al, 1963). However, there are many questions remain unsolved, especially the regulation between the exocytosis and endocytosis. We will attempt to find the core proteins in the exocytosis and endocytosis to provide a better understanding of the secretory regulation picture.

There are many papers report the important proteins in the regulation in the exocytosis and endocytosis. For example, Protein scaffolds in the coupling of synaptic exocytosis and endocytosis (Haucke et al, 2011). We can obtain some proteins related to the exocytosis and endocytosis from some literature, but we cannot obtain all the important proteins from the literature. We can obtain most of the proteins related to exocytosis and endocytosis from KKEG website. However, we cannot obtain the core proteins from the website. Therefore, our project will set up the protein-protein interaction related to the exocytosis and endocytosis from KKEG website and systematically analyze the database, trying to find the core proteins in the regulation of the exocytosis and endocytosis.

We will use rich-club coefficient method to determine if our network has a “rich-core”. This method is first time to use in determine the core proteins in exocytosis and endocytosis. Besides, we will use multiple methods to determine the core proteins and set up the spatial relation between these core proteins. All these methods we will use above will provide us a convincing list of the core proteins in the exocytosis and endocytosis. Our project will give a better understanding on the regulation of the exocytosis and endocytosis.

Innovation

We read lots of papers related to the analyze method of finding the core proteins in the exocytosis and endocytosis. The paper “Models of core/periphery structures” provide us the new sight of analyzing database. “rich-club” is a powerful method to analyze the core proteins (Dennis, E.L., et al 2014; Zhou, W.-X, et al 2008; Caetano, T.S, et al, 2007). However, no paper use the “rich-club” concept to analyze the core proteins in exocytosis and endocytosis. Besides, we will use multiple measurements to find the core proteins and compare these proteins in spatial. We will use a new method to test our results, which is that we use another secretory pathway proteins other than exocytosis and endocytosis to see if there are strong connection between these proteins and the core proteins in the exocytosis and endocytosis. All above, we use very new and meaningful methods to analyze our data.

We are the first one to use rich-club coefficient method to find the core proteins in the exocytosis and endocytosis. We are the first one to try to use multiple methods to analyze the whole exocytosis and endocytosis pathway regulation, from ER, Golgi, Endosome and plasma membrane. We are the first one to search the spatial connection between the core proteins we will find from the exocytosis and endocytosis database.

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,

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k} - 1)} \quad (1)$$

where $\phi(k)$ describes the value of the rich-club coefficient at the specified value of k, $N_{>k}$ is the number of nodes with degree greater than k, and $E_{>k}$ is the total number of edges of the nodes in the set $N_{>k}$. 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 k(Colizza et al. 2006). 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,

$$\rho(k) = \frac{\phi(k)}{\phi_{ran}(k)} \quad (2)$$

where $\phi(k)$ is described in Eq. 1, and $\phi_{ran}(k)$ 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, $\phi_{ran}(k)$ is found by taking the mean of the values of $\phi(k)$ for all 100 random networks. This allows calculation of the normalized rich-club coefficient $\rho(k)$(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 $\rho(k)$ 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 k 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,

$$FPC = \sum_{i=1}^{8} w_i C_i \tag{3}$$

where FPC is the total score for the node, C_i is one of the eight measures involved in the score and w_i is that specific measure's weight. In Eq. 3, the weights, w_i, are generated through maximizing the weights against the covariance matrix of all measures that make up C. 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, $CS(i)$, is given by the equation,

$$CS(i) = Z \sum_{\alpha,\beta} (\delta_i^*(\alpha,\beta) \sum_{j \in N(i)} \delta_i^*(\alpha,\beta)) \tag{4}$$

where Z is a normalization score assigned after the fact in order to make the maximum core score equal to 1, α is a parameter that describes the sharpness (or fuzziness) of the difference between core and

periphery ($\alpha = 1$ is the sharpest), β is a parameter that describes the size of the core as a percentage of the total nodes in the network, and

$$\delta_i^*(\alpha, \beta) = \frac{1}{1 + \exp\left\{-(i - N\beta)\tan\left\{\frac{\pi\alpha}{2}\right\}\right\}} \quad (5)$$

Finally, $N(i)$ is the set of all neighbors of i, so that we are also calculating the score for all of the neighbors of i as well as for i 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.

References

Borgatti, S.P., and Everett, M.G. (2000). Models of core/periphery structures. Social Networks *21*, 375–395.

Boycott BB, Gray EG, Guillery RW. Synaptic structure and its alteration with environmental temperature: A study by light and electron microscopy of the central nervous system of lizards. Proc R Soc B 1961; 154:151-172

Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., et al. (2015). The BioGRID interaction database: 2015 update. Nucleic Acids Res. *43*, D470–D478.

Colizza, V., Flammini, M., Serrano, A., and Vespignani, A. Detecting rich-club ordering in complex networks. Nature Physics *2*, 110–115.

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

Dennis, E.L., Zhan, L., Jahanshad, N., Mueller, B.A., Jin, Y., Lenglet, C., Yacoub, E., Sapiro, G., Ugurbil, K., Harel, N., et al. (2014). Rich Club Analysis of Structural Brain Connectivity at 7 Tesla Versus 3 Tesla. In Computational Diffusion MRI and Brain Connectivity, (Springer), pp. 209–218.

Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (New York: Springer).

Haucke, V., Neher, E., and Sigrist, S.J. (2011). Protein scaffolds in the coupling of synaptic exocytosis and endocytosis. Nature Reviews Neuroscience *12*, 127–138.

Jiang, Z.-Q., and Zhou, W.-X. (2008). Statistical significance of rich-club phenomena in complex networks. New Journal of Physics *10*.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

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

Lin WJ, Salton SR. The regulated secretory pathway and human disease: insights from gene variants and single nucleotide polymorphisms. Front. Endocrinol: 10.3389/fendo.2013.00096

McAuley, J.J., da Fontoura Costa, L., and Caetano, T.S. (2007). Rich-club phenomenon across complex network hierarchies. Applied Physics Letters *91*, 084103.

Pei Wang, Xinghuo Yu, and Jinhu Lu (2014). Identification and Evolution of Structurally Dominant Nodes in Protein-Protein Interaction Networks. IEEE Transactions on Biomedical Circuits and Systems *8*, 87–97.

Rombach, M.P., Porter, M.A., Fowler, J.H., and Mucha, P.J. (2014). Core-Periphery Structure in Networks. SIAM Journal on Applied Mathematics *74*, 167–190.

Taguchi T. Emerging roles of recycling endosomes. J Biochem. 2013 Jun; 153(6):505-10

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