Survey of Potential Regulators in the Fog/Non-Muscle Myosin-II Pathway

Abstract

Determining proteins that are involved in signaling pathways is essential for understanding complex cellular and developmental processes. However, wet lab approaches for identifying unknown members of a pathway can be costly. Using a computational approach to identify potential regulators from known members of a pathway and previously compiled protein-protein interactome networks saves time and money, and provides potential proteins that can be validated in a laboratory setting. We use Steiner tree approximation and breadth-first search (BFS) algorithms to identify potential regulators of apical constriction that participate in the Fog pathway. Overlap between our applications of both Steiner Tree approximation and BFS indicate the genes coding for Ubiquitin 63E (fb0003943), Spinophilin (fb0010905), Netrin B (fb0015774), and Casein Kinase IIα (fb0264492) as our primary candidates of interest. The extended lists produced by these algorithms also include 27 (Steiner Tree approximation) and 26 (BFS) additional candidate genes for a total list of 57 genes of interest.

Motivation

Cellular contractility is important for cell motility, cell division, and morphogenesis. One example of cellular contractility is apical constriction. During apical constriction, the apical side of a cell narrows following movement of non-muscle myosin-II motor proteins (NM-II). NM-II ‘walks’ along actin filaments and performs work in the cell using ATP. Much is known about the biochemistry of non-muscle myosin-II, including how it works in groups to bind to actin and move in the cell. In addition, there are some known regulatory molecules in Fog signaling pathway, the pathway that controls apical constriction. However, there are many other proteins that are important to the Fog signaling pathway that have not yet been identified.

Traditionally, proteins of interest have been found through manual review of the available scientific literature and tested through experimental manipulations in a laboratory setting. Because of this, it is extremely time-intensive and financially expensive to find and test proteins as they relate to a given pathway. Using a computational approach to solve this problem is substantially less expensive than manual investigation of these protein-pathway interactions. Network analysis is useful because possible related proteins can be determined from pre-existing interaction data and known members of the pathway, meaning that fewer proteins must be tested *in vivo*. Network analysis can be used to generate and rank lists of proteins and their likely relationship to a given pathway - this saves resources by narrowing the scope of the proteins considered to those which are likely to be relevant to the target pathway.

The challenge of using a computational method is that it is costly to determine known regulators, thus the input of known regulators that is used to find new regulators is limited. However, using the available known regulators as a starting point, algorithms can be used to identify other genes that are close to known regulators and far from known non-regulators. Our method aims to produce a list of candidate genes that is ranked according to likelihood of being involved in Fog signaling. Using a fly protein-protein interactome and known members of the Fog pathway, we suggest candidate Fog pathway genes for further testing in the laboratory.

Methods

Annotated Drosophila Interactome:

We used an undirected, unweighted *Drosophila* interactome provided by Dr. Ritz which is a composite of six other interactomes that specify protein-protein interactions.[1] FlyBase IDs have been preserved where possible, although a subset of edges cannot be mapped from UniProtKB IDs. The resulting dataset contains 17,736 proteins (nodes) and 364,157 protein-protein interactions (edges). Our computational approach works on a graphical model, $G = (V, E)$, where the graph is comprised of a set of nodes V, which encodes the proteins, and set of edges E, which encodes protein-protein interactions.

Labeled Regulators and Non-regulators:

Known regulators are those that have been shown to affect the signaling pathway through wet lab techniques. Additional node labels have been provided for known regulators and likely non-regulators. Twenty-one known, or positive, regulators are described in Manning et al. (2014), with additional regulators coming from the Gene Ontology database entries for apical constriction and gastrulation. In total, 113 nodes are identified as known regulators. An additional set of 501 likely non-regulators has also been derived from Gene Ontology, including proteins from pathways involved in autophagy, pigmentation, cell growth, circadian rhythm, and immune response. These pathways are assumed to be unassociated with apical constriction.

Pre-processing:

One of the features of protein-protein interactomes is that they contain a huge amount of data that is not necessarily relevant to a single signaling pathway. We have

[1] Information on these resources, DroID, FlyMine, myProteinNet, Mentha, SignaLink, and FlyReactome, can be found at https://github.com/annaritz/fly-interactome.

chosen to limit the size of the network based on proximity to known regulators prior to implementing our algorithms, thus decreasing the running time of our algorithms while hopefully increasing accuracy. Using BFS to identify maximally connected subgraphs, we identified the largest connected component of our composite interactome, keeping labeled nodes in either positive or negative sets that were included in this subgraph. “Connected” means that there is a way to get to every node included in the subgraph from every other node. The goal of this was to minimize the amount of edges and nodes to run through the rest of our analyses, because of efficiency concerns. We removed the protein *poor gastrulation* (FBgn0283685) for example from our list of regulators, because it does not appear in the composite interactome data. Removing nodes located more than four protein-protein interactions away from known regulators, we reduced our network to 14,326 proteins.

Ranking by Steiner Tree Approximation:

We refined our known regulators to a set of 104 proteins and used these to construct a Steiner tree approximation. A Steiner tree is a minimized version of the full network that incorporates the smallest path length between the positive nodes, while adding nodes that might make this path smaller. One real world application of steiner trees would be to evaluate travel between some set of locations via bus stations, with “edge weights” being a way to rank bus routes, shorter or smaller being preferable. Our algorithm assumed an edge weight of one for all protein-protein interactions. Removing the set of known regulators from the nodes found in the Steiner tree returned one set of potential regulators.

Ranking by Dijkstra’s Algorithm:

Dijkstra’s algorithm provides a set of distances from a source node to all other nodes. Each node was evaluated in terms of its distance from each positive in the positive set. The score is proportional to the sum of these distances, for a given node, in this way all nodes get scored and ranked according to scores. In this way, an arbitrary cut off of either a the top ranked or those nodes which scored within a desirable range get reported. Higher scores indicate greater confidence as possible positive regulators.

Ranking by Shortest Paths:

In this new formulation, we calculated the shortest path to non-muscle myosin II from all positive regulators. From this, we gained a list of non-terminal nodes that were used to obtain the shortest path from a positive, and compiled a list of positive nodes that used that non-terminal node to get to non-muscle myosin II in our network. We

believe the nodes that are used in this way are both “closer” to the positives in our network and to the protein of interest.

Results

Summary Statistics

Steiner Tree Approximation

The Steiner Tree approximation algorithm returned a graph containing 25 proteins that were not previously identified as positives (Table 1). Shortest Paths to NM-II identified 48 candidate proteins between positive nodes and the NM-II.

Table 1. Protein Identified in Steiner Tree (25)				
CG10347	CG7164	His3:CG33839	Nup54	Spn
CG11581	Chi	Khc	Pkn	Ten-m
CG17666	CtBP	Lcp4	Rac1	Ubi-p63E
CG34168	DCTN3-p24	Lsd-2	Sgt	eIF4G2
CG34227	Drep2	NetB	Sin3A	gro

Table 2. Proteins Identified in Dijkstra (37)				
14-3-3epsilon	Hsc70-4	Pkc53E	Ubi-p63E	nonA-l
14-3-3zeta	Hsp83	Rbp9	drk	sbr
Act42A	Myc	Smox	dsh	sgg
Act5C	N	Smurf	fne	spi
Akt1	Nab2	Spn	gro	
Appl	Nedd4	Stat92E	gskt	
Cdc5	NetB	Su(dx)	mts	

CkIIalpha	Pi3K21B	Ubi-p5E	nej	

Table 3. Proteins Identified in Shortest Paths to NM-II (48)				
14-3-3epsilon	CkIIalpha	N	Spps	ems
Acam	CycB	Patj	Ten-m	esc
Act57B	Cyp4g1	RPA2	Tl	eya
Alk	Df31	Rad51D	Ubi-p63E	flw
CG10347	Diap2	RfC3	alc	p53
CG17666	Drep2	RpA-70	cep290	qkr58E-1
CG7164	Hsp83	Sema-1a	cher	sgg
Cam	Khc	Sin3A	csw	wbl
CanB2	Lsd-2	Snapin	dally	
Chi	Moe	Spn	eIF4G2	

Table 4. Proteins Identified in Steiner Tree and Shortest Paths (12)			
CG10347	Chi	Lsd-2	Ten-m
CG17666	Drep2	Sin3A	Ubi-p63E
CG7164	Khc	Spn	eIF4G2

Table 5. Proteins Identified in Dijkstra and Shortest Paths (11)			
14-3-3zeta	Pi3K21B	Ubi-p63E	mts
CkIIalpha	'Pkc53E	drk	sgg
Hsp83	Spn	dsh	

Table 6. Proteins Identified in Steiner Tree and Dijkstra (4)

CkIIalpha	NetB	Spn	Ubi-p63E

Table 7. Proteins Identified in All Methods (3)		
CkIIalpha	Spn	Ubi-p63E

Discussion

In our analysis, we propose the application of a well-known graph optimization algorithm (Steiner Tree approximation and Dijkstra's algorithm) to graphs derived from Drosophila protein-protein interactomes. Each method produced a separate list of possible candidate genes, and these lists may be considered independently or jointly. This analysis represents an important first step in determining future targets of investigation. However, considering certain limitations to our analysis and the existence of feasible extensions for our work, we believe these results may be further refined for better accuracy prior to experimental assessment of these candidates.

EGFR comparison

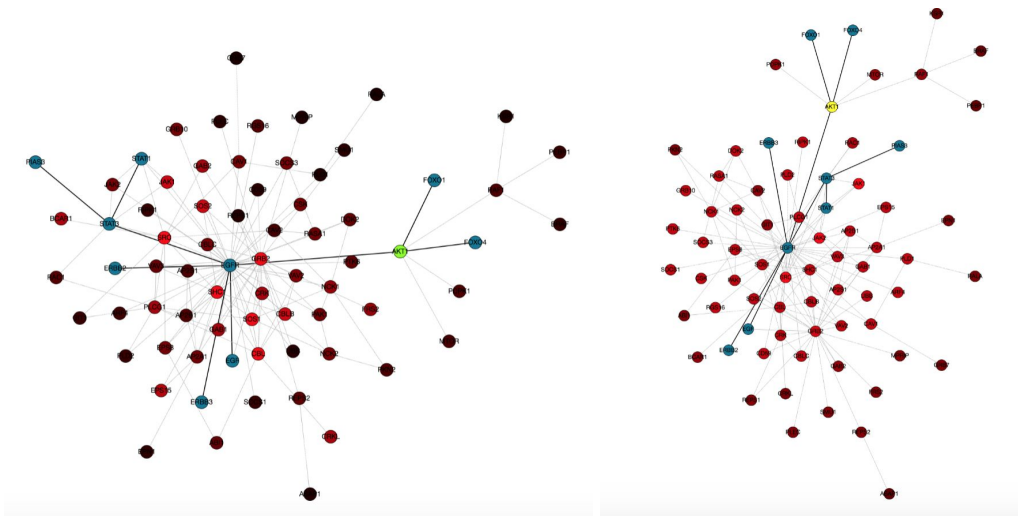

Figure 1. On the left, a visual representation of our algorithm output in a non-uniform edge weight scheme, while outputs are represented at right of an assumed equal edge weight scheme.

Our methods relied on a weighting scheme, which can either be evaluated by cross listing empirical data or made on an assumption of equal edge weights. Weights may be thought of as the 'distance' separating two nodes. We were not able to arrive at a system of assigned weights, such as those provided by HIPPIE or IntScore. To evaluate our assumption of uniform edge weights throughout the graph, we used EGFR as a toy data set, where assigned weights were available. By running our algorithm on the EGFR graph with assigned weights derived from empirical measures, and again on the EGFR graph where edge weights were assumed to be uniform, we found that the steiner tree and shortest paths algorithms return the same set of potential positives in both cases. In the case of the dijkstra rank algorithm, we found the comparison to have two of five top ranked nodes common to both sets, and four out of ten common in the top ten-new steiner tree. This indicates that the Dijkstra rank method is subject to greater variability when edge weights are assigned or assumed to be of equal value. However, we have no indication that one method is more successful at returning results of greater accuracy--the EGFR comparison does not necessarily change our confidence between steiner tree, shortest paths, and dijkstra based methods.

Role of Negatives

It is important to note that both analyses were performed with non-regulators included, and the status of these nodes was not taken into consideration by the algorithms. Additionally, the method for generating non-regulators should be considered when interpreting these results. Genes that were listed as likely non-regulators were involved in another pathway that was considered to be unrelated to apical constriction. However, it is not possible to entirely discredit all listed negatives because there may be cross-talk between seemingly unrelated pathways during development. Investigating genes originally indicated as non-regulators may encourage revision of classifications used to build the initial network. This consideration may be important for future iterations of this project.

Project Limitations and Extensions

In order to produce a functional output for our algorithms, certain assumptions were made about our data during pre-processing.

- Implementation of our Steiner tree algorithm utilized a placeholder weight of '1' for all edges which may have a significant impact on the corresponding results. Using a network based on the EGFR network as a toy dataset, there is an immediately evident loss of accuracy when setting all edge weights to 1 vs. empirically derived edge weights. This probe suggests edge weight will be an important aspect of our analysis to update.
- Experiment type was not an evaluated factor for interactions/edges. Distinguishing between binary interactions and complexes may inform the

analysis of our results, or may help us establish a basis for more accurate edge weights.

- More thorough analysis of node relationships via Community Detection may allow us to better assess when genes of interest are clustered/interrelated.
- Shortest-paths returns a dictionary with keys of non-terminal nodes and values as sets of positives that link to that node. These values vary between nodes, and could be used as a further statistic to help characterize potential regulators in further research.

Future Directions

We believe these results indicate a promising first attempt at determining novel candidates for genes affecting the Fog signaling pathway. However, we recognize that the items identified above reflect meaningful areas for further development. Adjusting our algorithms following examination of our initial results will allow us to determine what network characteristics (such as Negative node status and edge weight) most strongly influence our outputs. Implementing additional algorithms could give us additional dimensions for comparison between lists of genes, which may either identify additional nodes or more strongly emphasize nodes within our present candidate list. Lastly, feedback from wet lab experiments will be critical as it add new data and evaluate the strength of these initial predictions.

Contributions
Code from previous homeworks from Miriam, Kathy, Wyatt
All participated in code writing during class
Abstract(Elaine and Logan)
Motivation (Elaine and Logan)
Methods Wyatt, Kathy, Miriam, Elaine
Results (Logan, Wyatt, Elaine)
Discussion (Wyatt, Logan)

References

Manning, A.J., and Rogers, S.L. (2014). The Fog signaling pathway: Insights into signaling in morphogenesis. Developmental Biology *394*, 6–14.

Ritz, Fly Interactome, (2017), GitHub repository, https://github.com/annaritz/fly-interactome.