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 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

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

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 FBgn0283685, poor gastrulation, from our list of regulators, as 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 Breadth-First Search:

BFS was used as a second method for ranking potential regulators—in our case examining known regulators and uncategorized nodes to determine the distance between them. Each non-regulator node was scored (Eq 1), ranked from highest to lowest, and normalized to a value of one. The top 100 uncategorized nodes from this algorithm were identified as potential regulators because of their proximity to known regulators in the network.

$$\sum_{P \subset V} \frac{1}{BFS(P, v)}$$

Equation 1: sum of the reciprocal distance from a positive regulators, P, to node v

Results

The Steiner tree approximation produced a list with few likely non-regulators, with only 2 out of 31 Steiner tree genes also located in the list of likely non-regulators. By contrast, the BFS analysis performed relatively less well, with 7 of 30 genes tagged as non-regulators. Four genes were present in candidate sets from the Steiner tree approximation and BFS: Ubiquitin 63E (fb0003943), Spinophilin (fb0010905), Netrin B (fb0015774), and Casein Kinase IIα (fb0264492). Due to their presence in both lists, as well as the modest degree of of overlap between lists, we believe these genes and their associated proteins may represent higher priority targets of investigation.

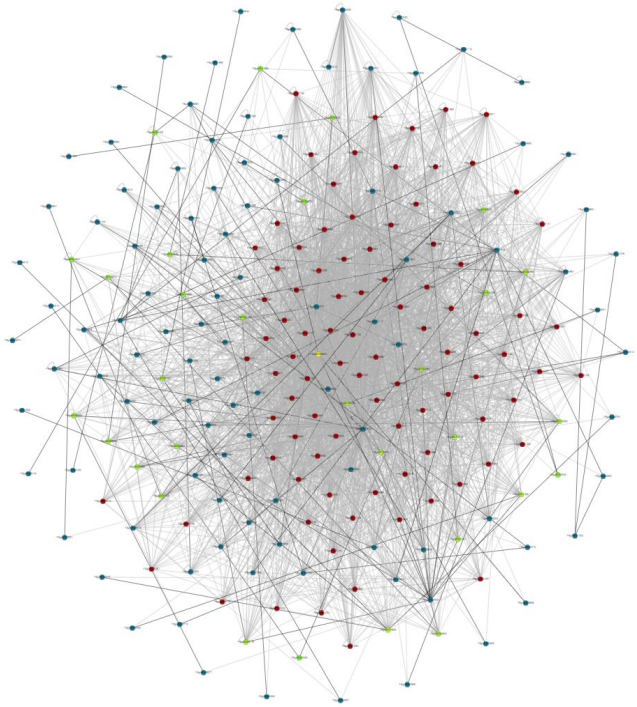

Figure 1. Composite network containing proteins identified in the Steiner tree approximation (green), BFS (red), and in both methods (yellow).

Result Number	FlyBase ID	Gene Name	Duplicate?	Non-regulator?
1.	‘FBgn0000579'	Enolase	No	No
2.	FBgn0001308'	Kinesin heavy chain	No	No
3.	FBgn0014010'	Rab5	No	No
4.	FBgn0031946'	CG7164	No	No
5.	FBgn0011661’	Moesin	No	No
6.	FBgn0004587'	B52	No	No
7.	FBgn0052479'	Ubiquitin specific protein 10	No	No
8.	FBgn0033593'	Listericin	No	No
9.	FBgn0030873'	CG15814	No	No
10.	FBgn0003943'	Ubiquitin 63E	BFS Rank 1	Yes
11.	FBgn0040505'	Anaplastic lymphoma linase	No	No
12.	FBgn0015283'	Regulatory particle non-ATPase 10	No	No
13.	FBgn0030540'	CG11581	No	No
14.	FBgn0042178'	CG32162	No	No
15.	FBgn0032906'	Replication Protein A2	No	No
16.	FBgn0032391'	escl	No	No
17.	FBgn0004644'	hedgehog	No	Yes
18.	FBgn0014868'	Oligosaccharyltransferase 48kD subunit	No	No
19.	FBgn0050178'	CG30178	No	No
20.	FBgn0015774'	Netrin B	BFS Rank 11	No
21.	FBgn0039635'	CG11876	No	No

22.	FBgn0003892'	patched	No	No
23.	FBgn0028408'	DNA fragmentation factor-related protein 2	No	No
24.	FBgn0015614'	Calcineurin B	No	No
25.	FBgn0264492'	Casein Kinase IIα	BFS Rank 26	No
26.	FBgn0032640'	Small glutamine-rich tetratricopeptide containing protein	No	No
27.	FBgn0031089'	CG9572	No	No
28.	FBgn0030342'	CG10347	No	No
29.	FBgn0027084'	Lysyl-tRNA synthetase	No	No
30.	FBgn0010905'	Spinophilin	BFS Rank 17	No
31.	FBgn0022764'	Sin3A	No	No

Table 1. List of candidate genes from application of the Steiner Tree approximation algorithm. Included in the table is the gene's FlyBase reference number, common name if available (or CG number if no common name given), indication of duplication, and the gene's status as a presumptive negative. Four results overlap results from the BFS algorithm (see Table 2), and two of the resulting genes of interest were listed among the possible presumptive negatives. A reference number for each result is also provided for convenience, but this number does not convey any meaningful information.

Ordinal Rank	FlyBase ID	Gene Name	Normalized BFS Rank	Duplicate?	Non-regulator?
1	FBgn0003943'	Ubiquitin 63E	1	ST #10	Yes
2	FBgn0001233'	Heat Shock Protein 83	0.82	No	Yes
3	FBgn0004638'	Downstream of receptor kinase	0.78	No	No
4	FBgn0000108'	β amyloid protein precursor-like	0.78	No	No
5	FBgn0259174'	Nedd4	0.75	No	No
6	FBgn0004907'	14-3-3ζ	0.74	No	No
7	FBgn0020622'	Pi3K21B	0.74	No	No
8	FBgn0010379'	Akt1	0.73	No	Yes
9	FBgn0001139'	groucho	0.73	No	No
10	FBgn0004177'	Microtubule star	0.73	No	Yes
11	FBgn0015774'	Netrin-B	0.73	ST #20	No
12	FBgn0003091'	Protein C kinase 53E	0.73	No	No
13	FBgn0004647'	Notch	0.73	No	No
14	FBgn0025800'	Smad on X	0.73	No	Yes
15	FBgn0020238'	14-3-3ε	0.73	No	No
16	FBgn0005672'	spitz	0.73	No	No
17	FBgn0010905'	Spinophilin	0.72	ST #30	No
18	FBgn0010263'	RNA-binding protein 9	0.72	No	No
19	FBgn0003557'	Suppressor of deltex	0.72	No	No
20	FBgn0086675'	found in neurons	0.72	No	No
21	FBgn0003321'	small bristles	0.72	No	No

22	FBgn0029006'	SMAD specific E3 ubiquitin protein ligase	0.72	No	No
23	FBgn0015520'	nonA-like	0.72	No	No
24	FBgn0000043'	Actin 42A	0.72	No	No
25	FBgn0086558'	Ubiquitin-5E	0.72	No	No
26	FBgn0264492'	casein kinase IIα	0.72	ST #25	No
27	FBgn0266599'	HSP cognate 4	0.72	No	Yes
28	FBgn0000042'	Actin 5C	0.71	No	No
29	FBgn0000499'	dishevelled	0.71	No	Yes
30	FBgn0040068'	Vav GEF	0.71	No	No

Table 2. List of candidate genes from application of the Breadth First Search algorithm. Included in the table is the gene's FlyBase reference number, common name or designation, normalized ranking of distance from all positives, ordinal ranking relative to other genes found through BFS, indication of duplication, and the gene's status as a presumptive negative. Four results overlap results from the Steiner Tree approximation algorithm (see Table 1), and seven of the resulting genes of interest were listed among the possible presumptive negatives.

Discussion

In our analysis, we propose the application of two well-known graph optimization algorithms (Steiner Tree approximation and Breadth-First Search) 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.

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 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.
- Average Neighbor Degree and other summary statistics are useful pieces of information which will also help with data interpretation.
- More thorough analysis of node relationships via Community Detection may allow us to better assess when genes of interest are clustered/interrelated.

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 (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.