Miriam Bern
January 11, 2018

Project Addendum Write Up

In our original project, we set out to find potential regulators of non-muscle myosin-II in the Fog signaling pathway. We did so by implementing three different methods, each of which involve finding shortest paths. One major assumption made at the start of our project was to assign each interaction in the unweighted interactome a weight of 1, which no doubt decreases accuracy.

In this addendum, I set out to see what effect scoring the interactions has on the results. The FlyBase interactome used in this project was put together using 6 different databases, with the majority of interactions coming from the DroID database; the interactome contains 17,763 nodes and 364,157 edges, while the DroID proteins in the interactome constitute 11,442 nodes and 262,179 edges. The DroID database contains a search function that provides all interactions that include a given protein. Furthermore, it also provides confidence scores for most of these interactions. The team that created the system[1] for scoring the interactions intended for it to be used to "annotate a PPI network so that interactions become weighted or probabilistic links useful for a variety of downstream analyses."[2] Given that this was a very convenient and reliable way to apply weights to our interactome, I took all interactions from the FlyBase interactome that came from the DroID database and searched for their weighted interactions. I then found the DroID FlyBase interactions in the weighted interactions downloaded from the database

[1] http://www.droidb.org/references.jsp
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2638943/

and ended up with a much smaller, weighted interactome to which I applied our three methods.

According to the DroID website, the confidence scores applied to each interaction represent how "biologically relevant" the interaction is. The closer to 1 the score is, the more likely it is to be biologically relevant than interactions with lower values[3]. Without accurate weights, our methods simply relied on how many edges there were between nodes to find the shortest paths. When trying to find regulators in a pathway, this will not produce incredibly accurate results because it treats each interaction as equally biologically significant and does not consider that there are interactions with lower probabilities of being biologically relevant. In addition, the methods used in our project rely on weights to find shortest paths, and therefore, very different results will be produced when edges are given unique weights. Thus, I expected the weighted interactome to produce different results.

Before applying the methods to the weighted interactome, I ran the code on the unweighted DroID portion of the FlyBase interactome to see what sort of effect discarding the interactions from the other 5 databases would have on the results. Overall, the results were more or less the same. About the same number of protein candidates were found from each method, though for the Steiner and shortest paths to NM-II methods, most of the candidates were different from the ones produced in the results of the original project. There were 16 fewer candidates derived from the Dijkstra method, but all proteins except 1 were the exact same as the original candidates. 3 candidates were also present in all outputs and 2 out of 3 were the same as the original 3 candidates; Spn

[3] http://www.droidb.org/ConfidenceScores.html

and Ubi-p63E. Because this interactome was also unweighted, the difference in outputs is most likely due to the fact that 6,321 nodes and 101, 978 edges were missing. It is no surprise that when two-thirds of the interactome is used, two-thirds of the results are the same, though it is worth checking if the different results found in the original candidates are actually from the other 5 databases.

The results from the weighted interactome were surprising. Not only were the results overall completely different, but each method produced tens if not hundreds more candidates than the original results. The Dijkstra method produced 1,459 candidates versus 37, though most of these 37 were found in the list of 1,459. However, the Steiner and shortest paths to NM-II methods did not contain most of the original candidates produced from these methods. The weighted interactome produced 40 candidates that were present in all outputs versus 3 from the original project, and none of these 3 were found in the list of 40 candidates.

Though the different results were expected, the number of results was unexpected. It is possible that more candidates were produced because the shortest paths happened to be the ones made up of lots of "lower" weight edges (keeping in mind that the weights would be inverted since lower weights are favored in shortest path algorithms but higher weights are more biologically relevant.) When edge weights must be considered, the paths with the smallest number of edges are not necessarily the lowest weight paths. This would make sense given that the majority of the edges have weights 0.5 and lower, with more than half of the total edges having weights around 0.2-0.4. In all 27,432 edges, only 2 have weights that are above 0.8 and those are the highest weights in the entire list of edges. If this is in fact the reason why so many candidates are produced, it is possible that

our method in the future needs to be refined in order to produce a smaller list of better, more rigorously selected candidates. In addition, it is worth noting that the weighted interactome was a much smaller portion of the original interactome, containing 27,432 edges and 7,920 nodes instead of 364,157 edges and 17,763 nodes. A future step would be to find weights for all interactions and see how the candidates change when the full interactome is used.