The Frequency of Feed Forward Loops in Randomized Degree Preserving Networks

John Gelinas
March 17, 2020

The presence of network motifs in transcriptional networks can be a great indicator of evolutionary pressures on organisms, especially when different species converge on the same sets of transcriptional networks. Determining whether a subgraph is in fact a network motif requires the comparison of the frequency of that subgraph in the organism to the expected frequency of that subgraph in a random network. In this project, the frequency of feed forward loops (FFLs), as depicted in Figure 1, has been examined in real networks of *E. Coli* and yeast, random networks with the same number of edges and nodes as the real network (Erdos & Renyi, or ER networks), and random networks with the same degree sequence as the real network as well as the same number of edges and nodes (degree preserving networks). This topic has been based on several scientific papers on the subject. The first paper, "Subgraphs in random networks" [1] discusses the differences in degree distribution, or the distribution of the number of edges connected to each node, in various types of networks. While ER networks have a Poissonian degree distribution, other networks, such as many transcriptional networks, exhibit a power law distribution. These networks are dubbed "scale-free" networks and this paper discusses the expected number of a particular type of subgraph (such as a feed-forward loop) in a scale-free network calculated from the average in-degree and out-degree of the network. This analytical equation will be compared to the numerical solution of the random networks as well as to the real networks. The second paper, "On the uniform generation of random graphs with prescribed degree sequences," [2] describes the algorithms that can be used to generate degree-preserving networks randomly. In addition, the frequency of feed forward loops in these random networks has been measured and matches the results of "Network motifs in the transcriptional regulation network of *Escherichia coli*," [3] closely. This project aims to replicate these results by using the algorithm described to create these random degree-preserving networks and compare the frequency of FFLs to the frequency found in these papers as well as to the analytical equation from the first paper.

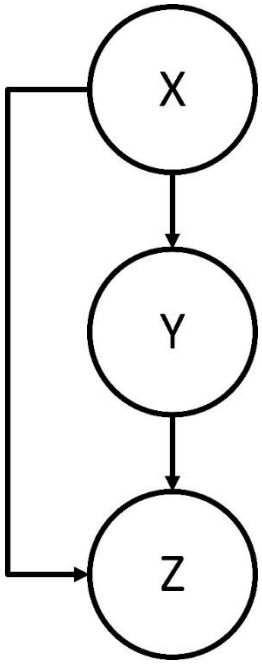

***Figure 1**: A Feed forward loop subgraph, where gene X regulates both gene Y and gene Z, and gene Y regulates gene Z.*

In order to begin to begin analysis of feed forward loops as network motifs, specific transcriptional networks need to be selected. The analytical equation for the expected number of feed forward loops in a network (Eqn. 1) is agnostic to the source of the network, replying only on parameters related to the degree of the network. Therefore, any network chosen can be compared to this equation:

$$Number\ of\ FFLs = \frac{\langle K(K-1)\rangle\langle RK\rangle\langle R(R-1)\rangle}{\langle K\rangle^3} \qquad \text{[Eqn. 1]}$$

where brackets denote average values, K is the out-degree at a node, and R is the in-degree at a node. The random degree-preserving networks, however, were formed from only *E. Coli* and yeast, so this project will also analyze those networks.

For both *E. Coli* and yeast, the real network's adjacency matrix was used to calculate the in-degree and outdegree at each node in addition to the actual number of FFLs in the network using MATLAB (Appendix A). From this information, a power-law fit was used to calculate γ, the parameter of the scale free network. For both real networks, γ falls between 2 and 3, as expected.[1] These results are shown in Table 1. The confirmation that these networks are scale-free networks with typical power degree distributions suggests that using Eqn. 1 is appropriate. The results of Eqn. 1 for these two networks are also shown in Table 1. The number of expected FFLs from this analysis is much lower than the real number of FFLs in the network, however this analysis gives no indication of what the spread of the number of FFLs should be, so it is impossible to tell exactly how likely these numbers of FFLs are to occur by chance alone. To learn more about the distribution, random networks must be generated.

Table 1: *Real Networks*

Results	**E. Coli**	**Yeast**
Actual FFL Number	42	70
Power Fit, γ	2.30	2.43
Predicted FFL Number	20.23	20.37

As previously discussed, random networks can be generated both with and without consideration for the degree of the network. For ER networks, where degree is not preserved, only the number of nodes and edges are matched. The number of FFLs produced in these networks over 1000 repetitions shows a very strong deviation from the real number of FFLs, as seen in Table 2. For degree preserving networks, a "switching" algorithm is used, where edges in a real network are randomly swapped many times to produce a network with random connections between nodes.[2] This preserves the number of edges entering and leaving each node, or the in- and out-degree of the system. The number of switches requires to truly randomize a network is proportional to the number of edges in the network. For good mixing to occur, setting the number of switches to be between 1 and 100 times the number of edges is usually sufficient. For a balance between mixing and program runtime, a switching number of 10 times the number of edges was chosen, which in preliminary tests was just as effective as a switching number 100 times the number of edges. This algorithm was also tested with and without preserving self-edges (SEs). When self-edges are preserved, if a self-edge is picked to be switched, it is disregarded, or if a switch creates a self-edge, then the switch is also disregarded. These degree-preserving networks have been randomly created 1000 times from each real network, and the results are shown in Table 2.

Table 2: *Randomized Networks*

ER Random	**E. Coli**	**Yeast**
FFL Number Mean	2.49	1.67
FFL Number Std. Dev.	1.57	1.28
Z-Score	25	53.5
Degree Preserving, Preserving Self Edges		
FFL Number Mean	7.81	14.5
FFL Number Std. Dev.	3.13	4.32
Z-Score	10.9	12.9

Degree Preserving, Switching Self Edges		
FFL Number Mean	18.5	19.6
FFL Number Std. Dev.	5.13	5.16
Z-Score	4.58	9.78
Degree Preserving, Literature		
FFL Number Mean	7.63	11.0
FFL Number Std. Dev.	3.05	3.71
Z-Score	11.3	15.9

The first result to note is that a completely randomized ER network has a significantly lower number of feed forward loops than the actual system, 25 to 35 standard deviations from the mean. This indicates that the real number of FFLs is certainly not due to random chance if all possible random networks are considered. This result is maintained when looking at degree preserving networks yet is not as extreme. When self-edges are preserved, the real number of FFLs are7.81 and 14.5, which are 10.9 and 12.9 standard deviations above the mean for the two transcriptional networks. Therefore, these number of FFLs are still not due to chance. This result is confirmed by the literature values for the number of FFLs in randomized, degree preserving networks. With 7.63 and 11.0 FFLs on average in *E. Coli* and yeast randomized systems, respectively, literature values also confirm that FFLs in these organisms are statistically significant. Furthermore, the number of FFLs in the literature and in the degree preserving test where self-edges are preserved are nearly the same for *E. Coli*, suggesting that the algorithm is functioning. There is some deviation with the yeast network which is a statistically significant difference. When self-edges are switched, the average numbers of FFLs change. The mean number of FFLs increases to 19.6, which is more similar to the expected number of FFLs calculated using Eqn. 1. The same is true for the degree preserving *E. Coli* networks where self-edges are switched, as the mean number of FFLs increases to 18.5, also more similar to the Eqn. 1 result. This makes sense as the equation does not know the number of self-edges a network has, and therefore would better represent a network that has an average number of self-edges rather than a prescribed number.

While many of the results of this project match up with literature values, the most interesting discrepancy is number of FFLs in self-edge preserving yeast-based random networks deviate from the those in the literature. It is possible that the original adjacency matrix was not in fact the same for these networks. This seems unlikely, as the literature results cited Uri Alon, whose website was used to download the data for this project.[4] Because of this, it is more likely that there is another effect that is not being accounted for in this algorithm, that does not present as much of an issue in the *E. Coli* network. For future work, it may be valuable to try a "matching" algorithm as well, which instead of switching edges of a real network, matched up out- and in-edge stubs randomly.

From this work, several things are very clear. First, degree-preserving networks better represent the real networks by creating a network with power-law degree distribution with the same γ as the real network. Second, the nuances of the algorithms can have a significant effect on the results. Whether certain aspects of the network (such as self-edges) are maintained or randomized, does affect the organization of the network in a noticeable way. Finally, the number of feed forward loops present in actual transcriptional networks of *E. Coli* and yeast are statistically significantly more than in any random network generated from any algorithm discussed in this project. This means that these subgraphs are in fact network motifs and do indicate that they may provide a fitness advantage to the organism.

Works Cited

[1] Itzkovitz, S., Milo, R., Kashtan, N., Ziv, G., & Alon, U. (2003). Subgraphs in random networks. *Physical Review E*, *68*(2). doi: 10.1103/physreve.68.026127

[2] Milo, Ron & Kashtan, N & Itzkovitz, Shalev & Newman, M. & Alon, Uri. (2004). On the uniform generation of random graphs with prescribed degree sequences. Tech rep. 21.

[3] Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, *31*(1), 64–68. doi: 10.1038/ng881

[4] Alon, U. (n.d.). Collection of complex networks. Retrieved from http://www.weizmann.ac.il/mcb/UriAlon/download/collection-complex-networks

Appendices

Appendix A – MATLAB Code