THE SWEDISH NATIONAL GRADUATE SCHOOL IN MEDICAL BIOINFORMATICS

2021 APPLIED BIOINFORMATICS

Project: Are well-connected proteins more multi-faceted than others?

Author:
David Lund,
dlund@chalmers.se

June 29, 2021

Results

This project aims to investigate whether proteins that are well-connected with other proteins, i.e. have interactions with many other proteins in a protein interaction network, consist of a larger number of protein domains (functional units in a protein) than other proteins. For this end data from STRINGDB [1] was used to create an interaction network of human proteins, the proteins were separated into two groups based on if their node degree in the network was higher or lower than 100, and the number of protein domains associated with each protein in the two groups was calculated usign data from BioMart from Ensembl [2]. A summary of the results is shown in Table 1.

Table 1: *Summary of the number of protein domains associated with proteins with node degree $>$ 100 versus proteins with node degree $\leq$ 100.*

Node degree	Mean	Median	Min	Max
$>$ 100	2.522	2.0	1	314
$\leq$ 100	2.438	1.0	1	67

It can be seen that the mean number of protein domains was similar between the groups, indicating that the difference between the number of protein domains is small. However, this might me a skewed measure due to e.g. outliers, and therefore might not accurately represent the distributions.

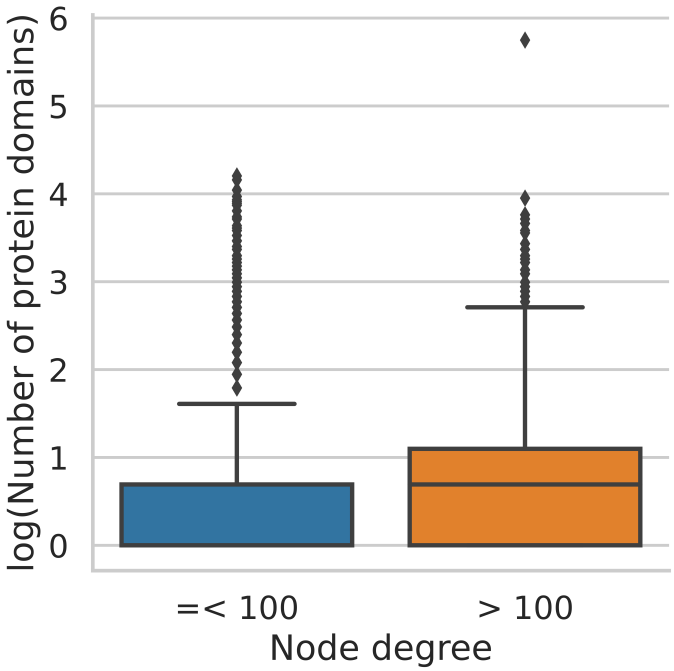

Figure 1: *Boxplot depicting the number of protein domains associated with two groups of proteins: ones with a node degree $>$ 100 and ones with node degree $\leq$ 100 in a protein interaction network. To deal with outliers, the numbers representing protein domains have been log-transform.*

Indeed, there was a difference in the median value, suggesting that the number of domains associated with more well-connected proteins might actually be larger on average. To get a better understanding of the distributions, a boxplot was generated (Figure 1). When generating this plot, it was decided to first normalize the plot by log-transforming the number of protein domains to deal with the large variance that resulted from a few outlying datapoints. From the plot, it becomes clear that there is a difference between the two groups, and that the proteins with node degree > 100 were associated with more protein domains. However, a few things should be pointed out, that make these results less conclusive.

One issue is that the two groups of proteins were of very different sizes, with the less well-connected proteins outnumbering the well-connected proteins by far. This might affect how well the true distributions are reflected in the results, however since even the smaller group contained several thousand proteins this can be argued to not be of much importance. However, it can be seen that both groups in Figure 1 contain several datapoints that are considered outliers, but given more data this might actually not be true as they are very evenly spread out. Finally, there were many proteins for which no information about protein domains was available. This means that the results as they are presented here might not accurately reflect reality, since reality is in this case unknown. However, from the available data, there is at least a clear indication that well-connected proteins are more multi-faceted than others.

Code availability

All code used to produce the presented results can be found at
https://github.com/davidgllund/abi_project.

References

[1] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937, 2016.

[2] Fiona Cunningham, Premanand Achuthan, Wasiu Akanni, James Allen, M Ridwan Amode, Irina M Armean, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, et al. Ensembl 2019. *Nucleic acids research*, 47(D1):D745–D751, 2019.