Results

Data

Signed network

Signed networks are very important in systems biology because they include more information than "bare" networks, hence they are better representations of the real systems. Signed protein networks include the physical interactions between proteins as well as signs, activation - inhibition interactions. The first large scale signed protein interaction network was constructed in 2014 for *D.melanogaster's* proteome by (Arunachalam Vinayagam et al. 2014). At the time of writing and to the author's knowledge no other signed protein interaction network exists. The data from (Arunachalam Vinayagam et al. 2014) are freely available for everyone to download.

Table 0.1: Summary of the signed PPI network

Type	All *D.melanogaster*	Complete network	Giant component
Proteins	9107	3352	3058
Interactions	47239	6094	5930
Positive	0	4109	3998
Negative	0	1985	1932

The authors of (Arunachalam Vinayagam et al. 2014) integrated protein-protein interaction data, that are available in many databases, with data from RNAi screens to reveal activation-inhibition relationships. Their approach was validated with some already known activation-inhibition relationships derived from small scale experiments (literature). Also some previously unknown relationships were unraveled that were later confirmed experimentally, a result that showed the high predicted power of the approach.

Table 0.2: Sources of interactions of the signed PPI network comparison and summary

Type	Positive	Negative	NA	Different	Total
Sign score - All interactions	4109	1985	0	-	6094
Sign score - Predicted	3826	1865	0	-	5691
Sign score - Literature	309	125	0	-	434
Sign score - Duplicates	-	-	0	-	31
Co-express development correlation	4127	1873	94	-	6094
Comparison of Co-express development correlation & Sign score interactions	3008	834	94	2158	6094

The integration of signs in the protein interaction network of *D.melanogaster* didn't come without a cost. As seen in the table 0.1 only $\approx 33\%$ of the original proteins are included and even less, $\approx 13\%$, of their original interactions. The original protein interactions which are experimentally detected are estimated to represent only $\approx 20\%$ of the real interactions (Gavin, Maeda, and Kühner 2011; Yu et al. 2008). So the signed network contains about $\approx 3\%$ of the expected real protein interactions of *D.melanogaster*.

The interactions between proteins of the signed protein interaction network are both directed and signed. The

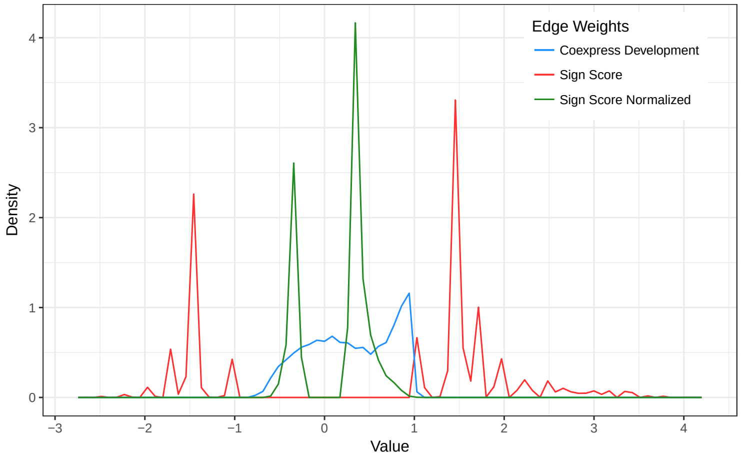


Figure 0.1: Density of signed weights and gene expression correlation. Also we normalized the original signs by dividing all values with the maximum absolute value so the distribution lies in the $[-1, 1]$.

signs take scores in the interval $[-2.645751, 4.123106]$ as seen in the density plot (Figure 0.1). It is noticeable that values in the interval $(-1, 1)$ are missing. This is due to the cutoff values in the interval $(-1, 1)$ which was applied to reduce possible errors. Also we found that there were 31 duplicated interactions which is due to the inclusion of signs from literature. From these interactions we kept the ones from literature.

Another approach to add signs to a protein interaction network is to use gene expression data and then correlate the levels of expression between genes (Ou-Yang, Dai, and Zhang 2015). These correlations are sometimes used as signs of interactions though this approach is not widely accepted and in not consider a good practice. Nevertheless the authors compared gene expression time-course data with the signs and found big differences, 2158 signs have the opposite score (table 0.2).

The signed network is not connected but has a giant component of 3058 proteins and 5930 interactions. The degree distribution of the network is scale free, following a power law-like distribution (Figure 0.2). For the rest of this article when we refer to the network we will mean its giant component.

Protein essentiality

To annotate the proteins of the signed network proteins with their essentiality consensus we used the freely available database: *Online GEne Essentiality database (OGEE)* (Chen et al. 2012). OGEE has 3 distinct labels for genes, essential, conditional and nonessential. In table 0.3 we can see that from all the 13373 genes of *D.melanogaster* only $\approx 2\%$ are essential. Essential genes in OGEE are those who were identified as essential consistently in all distinct experiments. On the hand, conditional are the genes that have been identified as essential in at least one experiment and nonessential in other experiments.

From the annotation of OGEE data to the signed network we found 156 proteins that are not included in the database (NA values in table 0.3). In all analyses we considered the conditionally essential proteins to be nonessential. Also for the decision trees inference we excluded the NA proteins, although we kept them when calculating the centrality indices.

Table 0.3: Gene essentiality consensus from OGEE database

Consensus	All *D.melanogaster*	Complete network	Giant component
Nonessential	13373	3009	2737
Essential	267	154 + 33 conditional	146 + 29 conditional
Conditional	141	33	29
NA	0	156	146
Total	13781	3352	3058

Protein complexes

Protein complexes are functional molecular units that consist of physically interacting proteins. In order to learn more about the proteins of the signed network we downloaded protein complex data from the COMPLEAT database (A Vinayagam et al. 2013). COMPLEAT database has freely available data and also provides a platform for analyses for various types of data. We downloaded the protein complexes of *D.melanogaster* and their proteins. There are 2 types of complexes in COMPLEAT, those collected from individual experiments referred as *literature* and those inferred from 2 algorithms, *CFinder* and *NetworkBlast*. When we plotted the distribution of complexes size in terms of number of containing proteins we saw a pattern (Figures **??** and **??**). NetworkBlast, which predicted ≈ 50%(2893) of the complexes (table **??**), has a upper limit of 16 proteins in complex size (Figures **??** and **??**). This has an impact in analyses so for the rest of the article we will distinguish the complexes in 2 categories, All complexes and Literature complexes. See more about the COMPLEAT database and the bias we discovered in Appendix **??**.

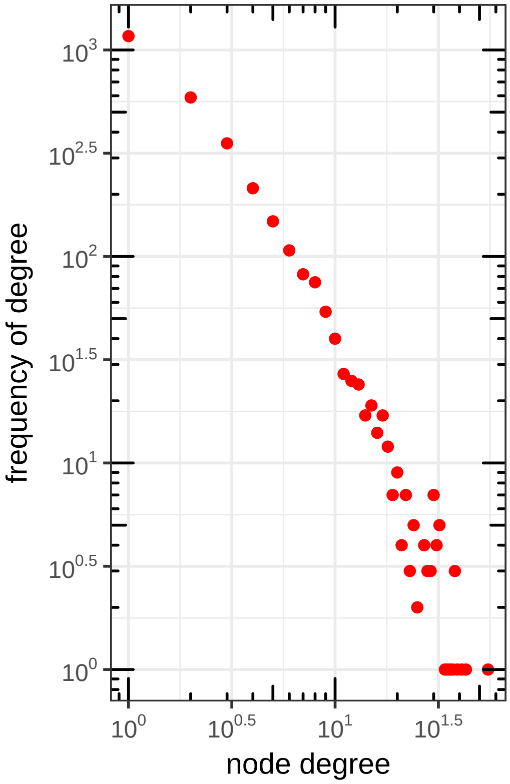


Figure 0.2: The degree distribution of the network is scale-free.

Evaluation of essentiality prediction methods

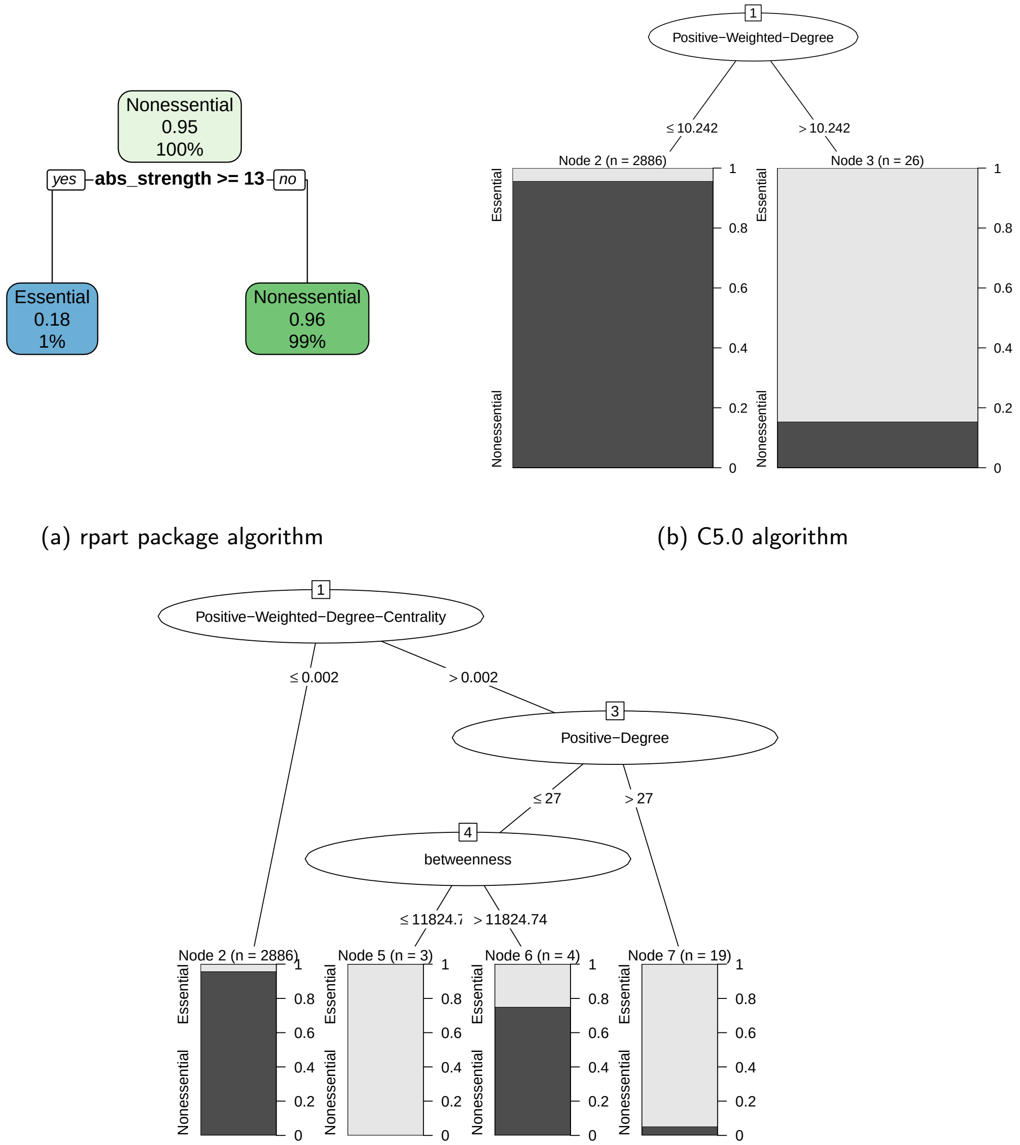
After the calculation of centrality indices for all network proteins we created decision trees for essentiality consensus prediction. We chose the centralities as variables for decision rules from which we constructed three trees using the algorithms from rpart package, C4.5 and C5.0. The C4.5 algorithm created a tree with higher complexity, more branches, than the rpart and C5.0 algorithms (Figure 0.3c). Also the C4.5 algorithm had better precision, because it had less false positives but lower recall than the other algorithms (table 0.4, equations **??** and **??**). In addition, rpart algorithm used the weighted degree but with the absolute values of signs and the algorithms C4.5 and C5.0 used positive weighted degree (equation **??**), positive degree (equation **??**) and betweenness (equation **??**) as decision rules. The latter is a new and interesting result because it may represent a new property of essential proteins in signed networks.

Table 0.4: Confusion matrix for the 3 different algorithms of decision trees.

Type	C5.0	rpart	C4.5
True Positives	22	23	21
False Negatives	124	123	125
True Negatives	2762	2761	2765
False Positives	4	5	1
Precision	0.846	0.821	0.955
Recall	0.151	0.158	0.144

We used ROC curve, Precision Recall curve and Jackknife curve to compare the predictability power of centralities and decision trees (Figure 0.4).

In Figure 0.4 we see that the best methods for predicting protein essentiality consensus from the signed network are the decision trees. The rpart algorithm surpasses all centralities in all methods ($AUC = 0.881$). Quite similar performance is delivered from the C4.5 algoritm ($AUC = 0.874$). Degree centrality is the best performed centrality with $AUC = 0.776$. Worth mentioning is the low performance of betweenness centrality ($AUC = 0.591$) and closeness centrality ($AUC = 0.658$). In the Jackknife curve (Figure 0.4b) we see that after the 25 proteins there is sudden decrease in the essential protein accumulation from all best methods. Degree centrality accomplished the highest retrieval of essential proteins. Decision trees even though had faster essential protein accumulation (i.e higher precision) reached a plateau in 23 essential proteins. Also closeness centrality eventually and gradually reached the top methods in correct essential protein

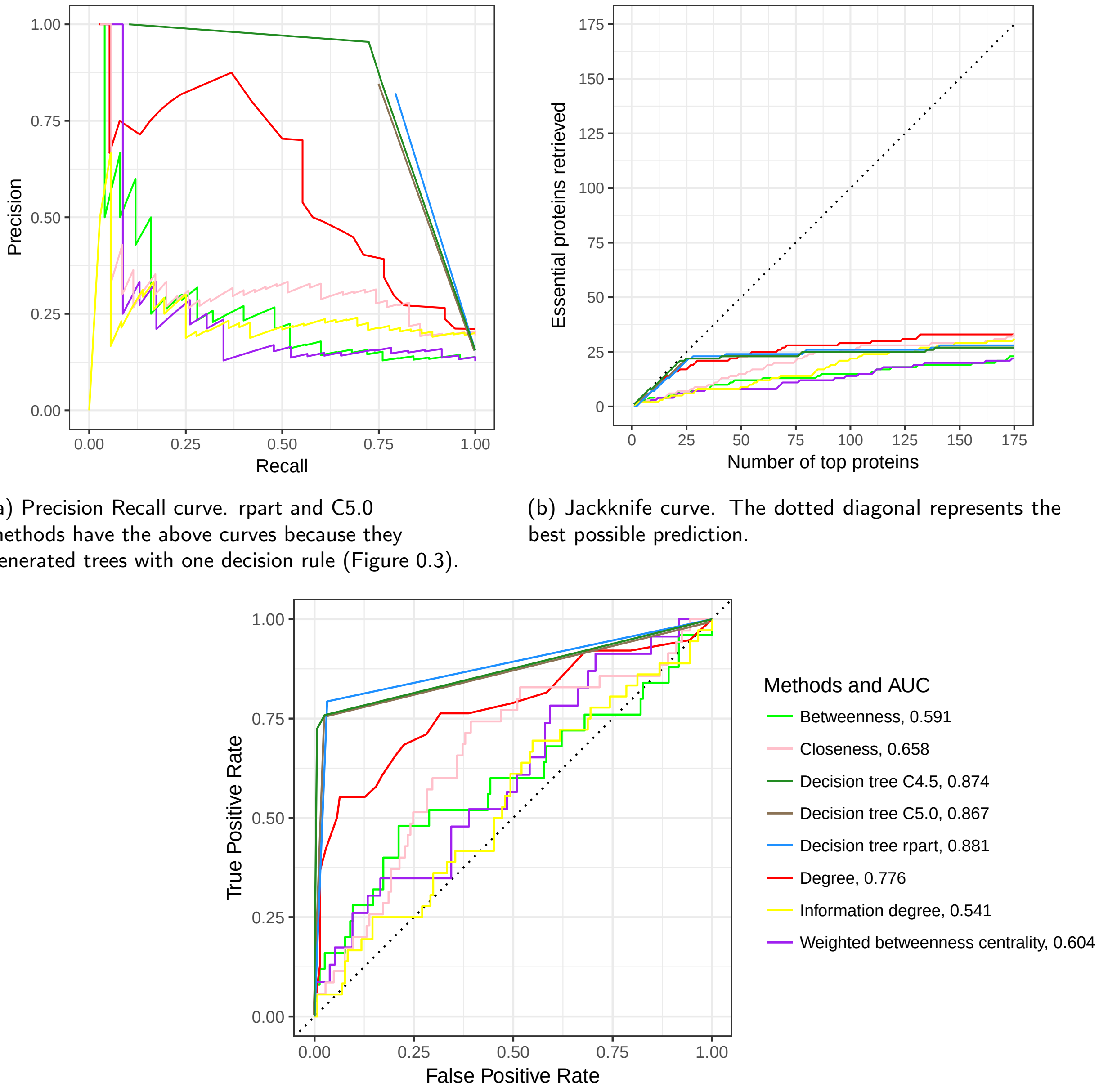


(a) rpart package algorithm

(b) C5.0 algorithm

(c) C4.5 algorithm

Figure 0.3: Trees from different algorithms. **(a)** and **(b)** generated oversimplified trees but **(c)** generated a little more complex and more precise tree.



(a) Precision Recall curve. rpart and C5.0 methods have the above curves because they generated trees with one decision rule (Figure 0.3).

(b) Jackknife curve. The dotted diagonal represents the best possible prediction.

(c) ROC curve. The dotted diagonal represents the random predictions.

Figure 0.4: Evaluation methods for the different prediction methods of protein essentiality.

prediction.

Essential subgraph

We investigated the subgraph of essential proteins of the signed network which contains only essential proteins and their interactions (Figure 0.5). What we found was that essential proteins form a cluster which contains only positive - activation interactions. There are 3 negative interactions but they are from conditionally essential proteins. To investigate further this unexpected result we studied the inner structure of the essential cluster with graph theory tools and we performed Gene Ontology annotation.

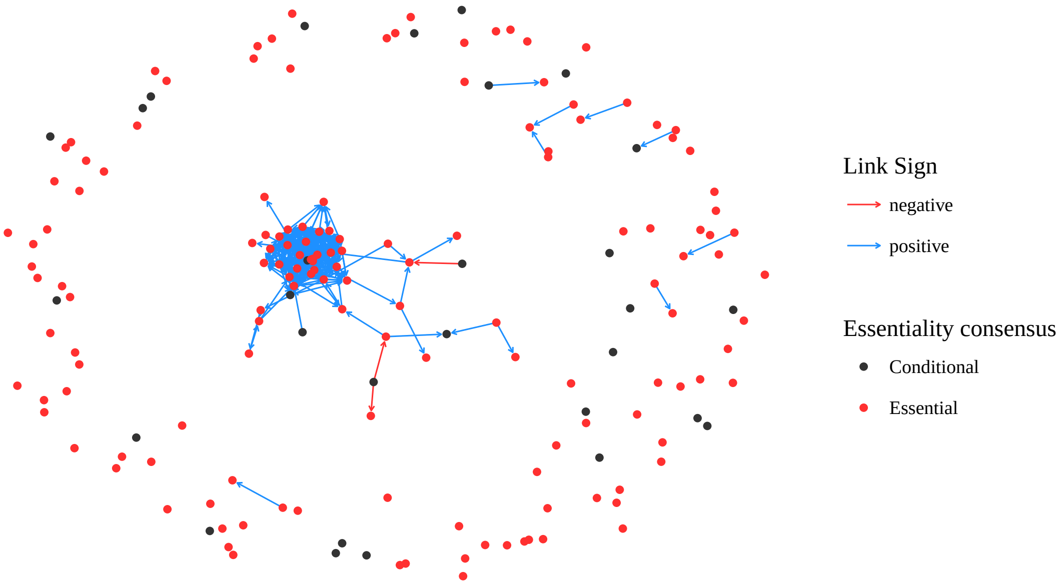


Figure 0.5: Interactions between essential proteins are positive. The only negative interactions are from conditionally essential proteins.

Decomposition of essential cluster

The goal of the Perron - Decomposition is to decipher how the information flows in the network. Using Tarjan's algorithm (Tarjan 1971) we found the strongly connected components (or equivalently the irreducible components, Theorem **??**) of the essential cluster (table 0.6). There is only one strongly connected component with 20 proteins, all the other proteins are singular components (Figure 0.6). Information in the strongly component can reach all proteins from any protein in the component (Definition **??**).

Table 0.5: Essential cluster information

Type	Values
Essential proteins in network	146 + 29 conditional
Connected essential cluster	36 proteins, 243 positive interactions
Strongly connected components	17 components (16 singular)
Irreducible component	20 proteins, 118 positive interactions
Perron–Frobenius eigenvalue of irreducible component	4.0210
Number of equal maximum eigenvalues	$k = 1$

The essential cluster is weakly connected so it is reducible which by Definition **??** means that its adjacency matrix can take an upper triangular form. Ultimatly this means that some proteins have only outgoing interactions and some only incoming interactions. So information in the essential cluster has direction. In Figure 0.7 we reconstructed the network using the components (16 singular and 1 with 20 proteins) to present the direction of the essential cluster. Information can move only from top to bottom. That way we can divide the proteins into 3 categories, input, processing and output. In input are the proteins 14,15,16,17, in processing are the proteins 2,3,4,5,11,13 and the irreducible component 8 and finally in the output are the proteins 1,6,7,9,10 and 12 (Figure 0.7).

Next we further explored the structure of the irreducible component (Figure 0.6). Using the strong theorem from Perron and Frobenius (Theorem **??**) we calculated the eigenvalues of the component and we found there is only one positive real eigenvalue equal to the spectral radius of the graph (table 0.6). We conclude that the essential strongly connected component has primitive adjacency matrix (Definition **??**) and it doesn't have cycles of the form of matrix **??** (Theorem **??**).

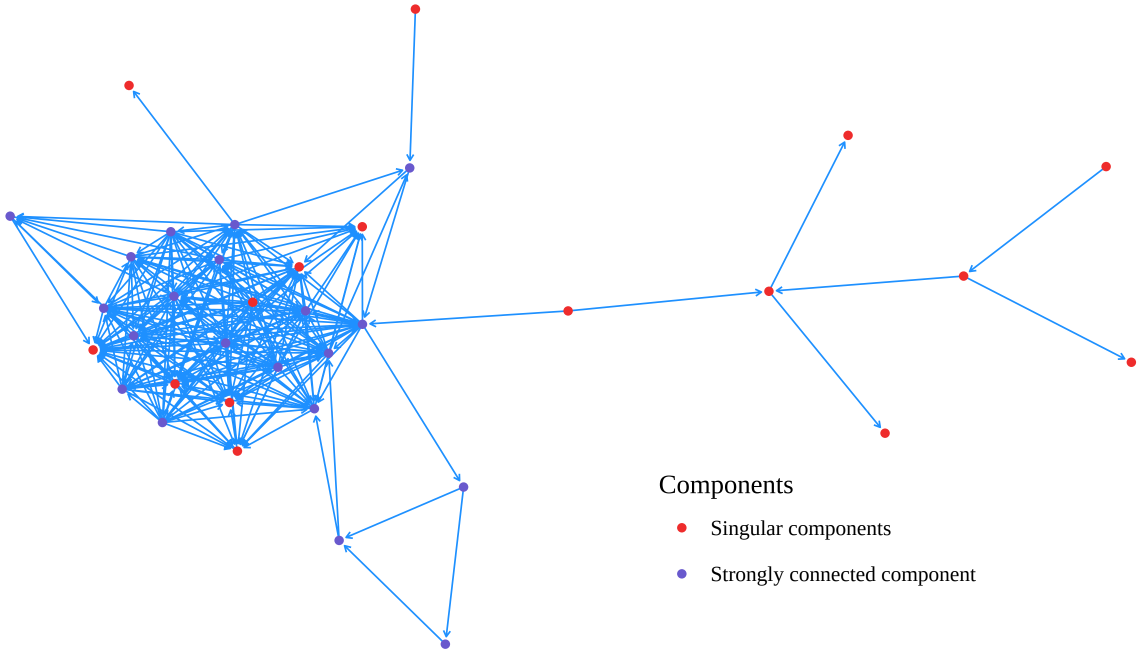


Figure 0.6: Essential proteins strongly connected component.

Gene ontology annotation

To examine the functions of proteins in the essential cluster we annotated them to Gene Ontology. More specifically we used the Biological Process ontology of Gene Ontology and we performed Fisher's Exact Test to find statistically significant terms. As a background protein pool we used all the proteins in the signed protein network. The test resulted in 58 significant GO terms with $p - value < 0.01$

Table 0.6: Essential cluster information

Ontology	Network Proteins
Biological Process	2858
Molecular Function	2721
Cellular Component	2655
None	317

Modular essentiality

Chen, Wei Hua, Pablo Minguez, Martin J. Lercher, and Peer Bork. 2012. "OGEE: An online gene essentiality database." *Nucleic Acids Res.* 40 (D1): 901–6.

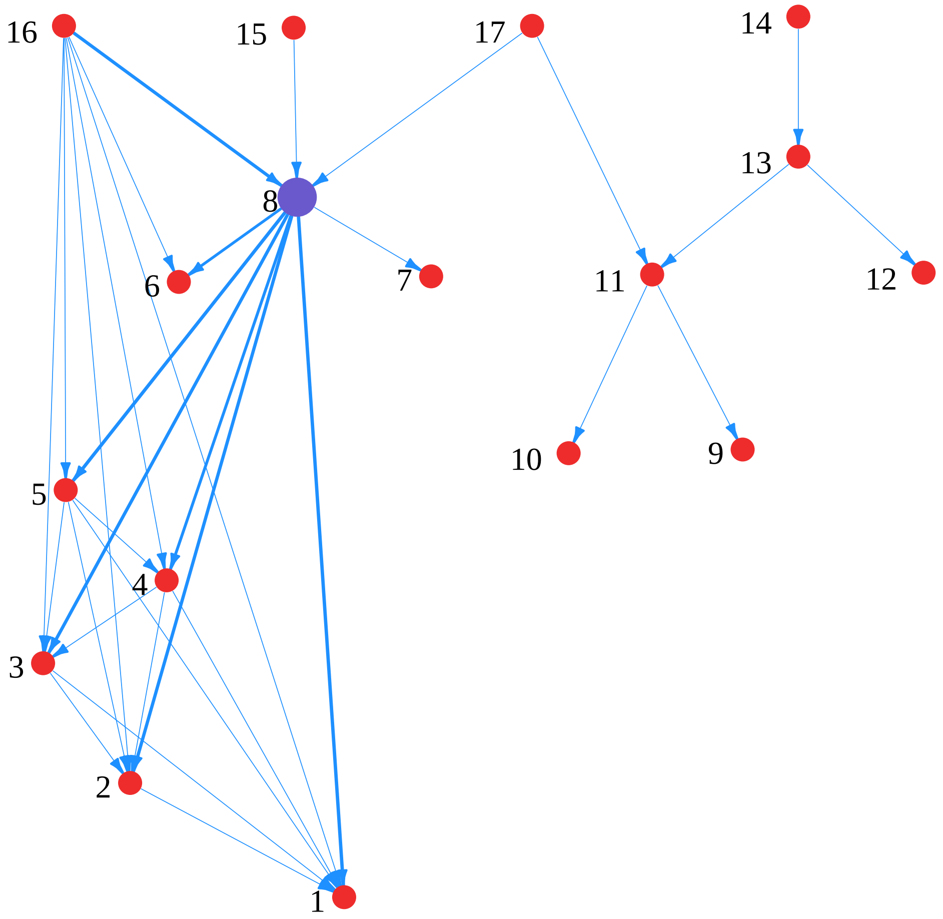


Figure 0.7: Perron - Frobenius decomposition. The purple node (8) is the strongly connected component. The links thickeness increases with the number of interactions between the components.

doi:10.1093/nar/gkr986.

Gavin, Anne Claude, Kenji Maeda, and Sebastian Kühner. 2011. "Recent advances in charting protein-protein interaction: Mass spectrometry-based approaches." *Curr. Opin. Biotechnol.* 22 (1): 42–49. doi:10.1016/j.copbio.2010.09.007.

Ou-Yang, Le, Dao Qing Dai, and Xiao Fei Zhang. 2015. "Detecting Protein Complexes from Signed Protein-Protein Interaction Networks." *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 12 (6): 1333–44. doi:10.1109/TCBB.2015.2401014.

Tarjan, Robert. 1971. "Depth-first search and linear graph algorithms." *12th Annu. Symp. Switch. Autom. Theory (Swat 1971)* 1 (2): 146–60.

doi:10.1109/SWAT.1971.10.

Vinayagam, A, Y Hu, M Kulkarni, C Roesel, R Sopko, S E Mohr, and N Perrimon. 2013. "Protein complex-based analysis framework for high-throughput data sets." *Sci Signal* 6 (264): rs5. doi:10.1126/scisignal.2003629.

Vinayagam, Arunachalam, Jonathan Zirin, Charles Roesel, Yanhui Hu, Bahar Yilmazel, Anastasia A. Samsonova, Ralph A. Neumüller, Stephanie E. Mohr, and Norbert Perrimon. 2014. "Integrating protein-protein interaction networks with phenotypes reveals signs of interactions." *Nat Methods* 11 (1): 94–99. doi:doi:10.1038/nmeth.2733.

Yu, H., P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, et al. 2008. "High-Quality Binary Protein Interaction Map of the Yeast Interactome Network." *Science (80-.).* 322 (5898): 104–10. doi:10.1126/science.1158684.

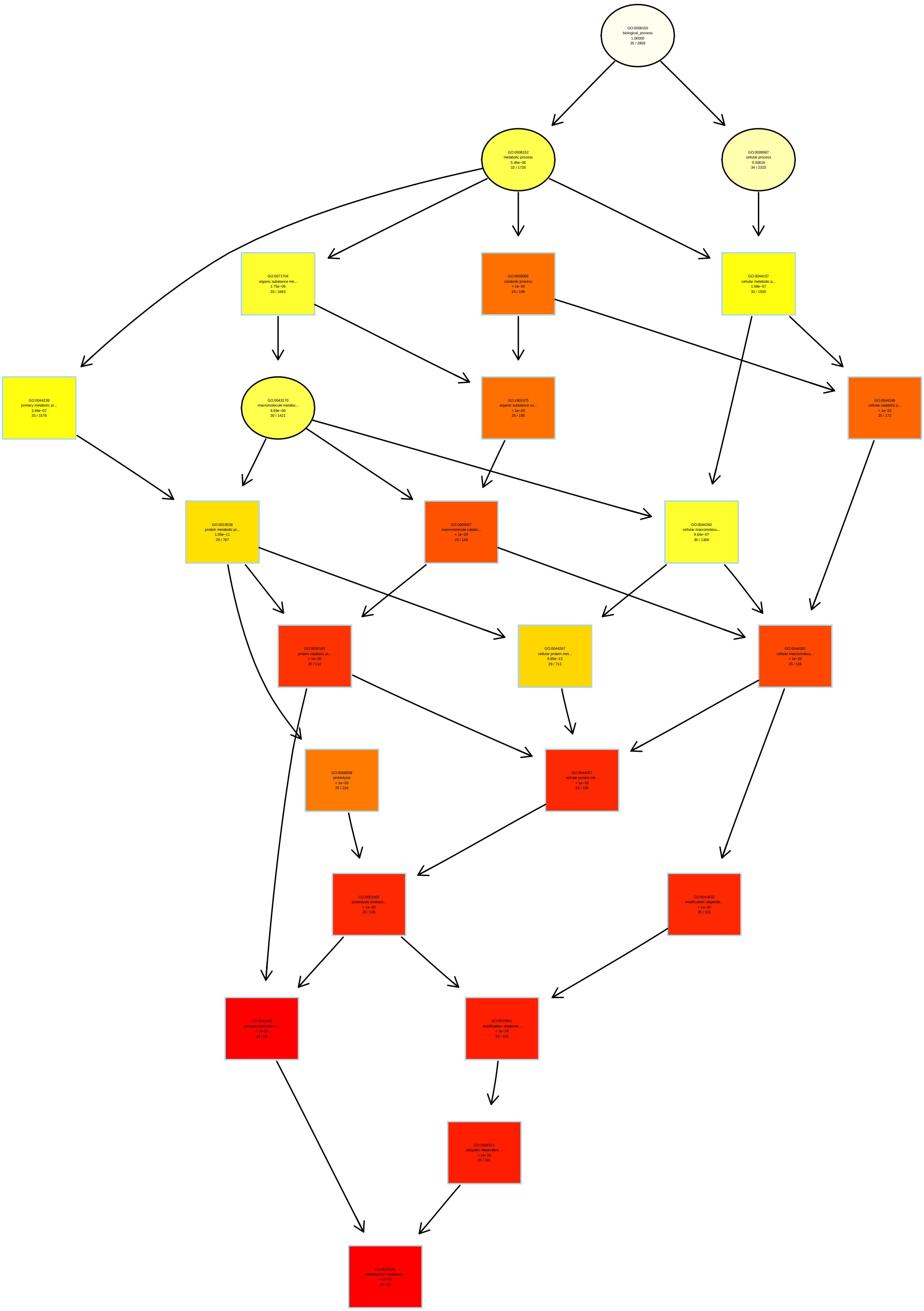


Figure 0.8: Singular enrichment analysis.

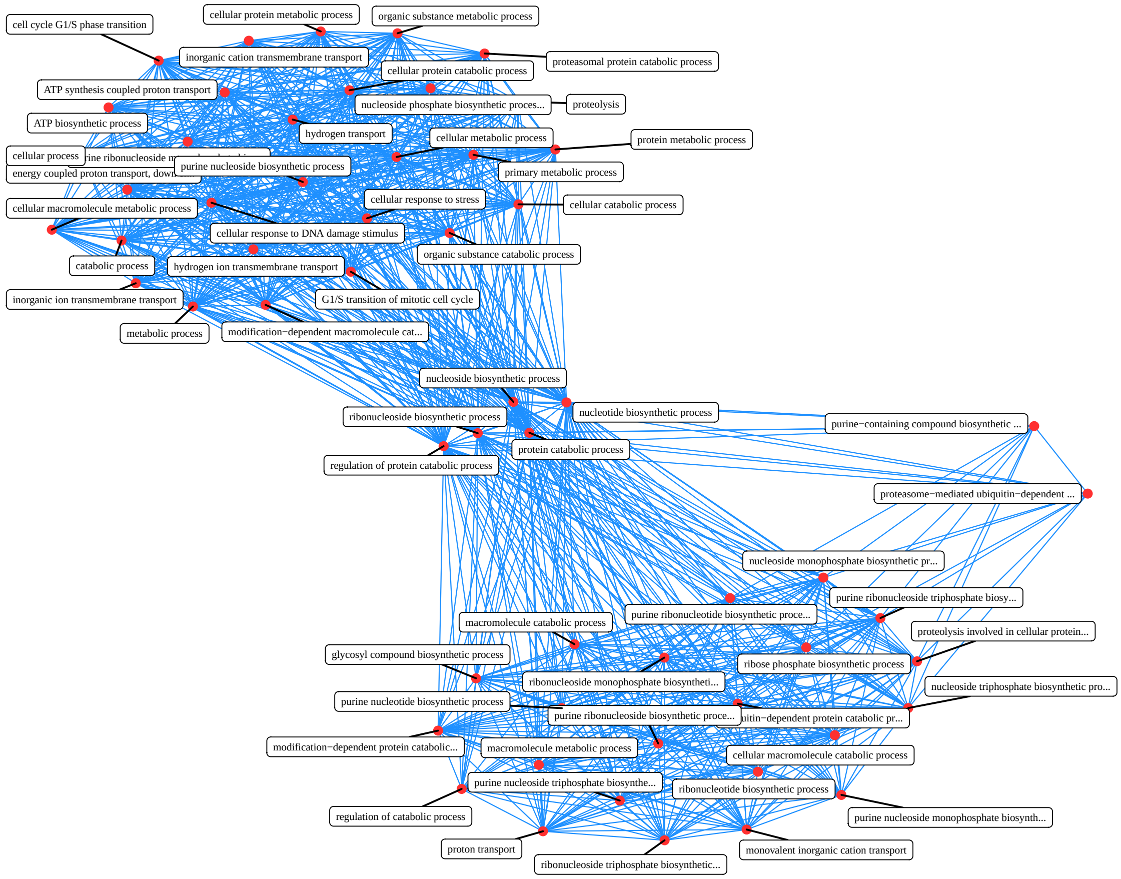
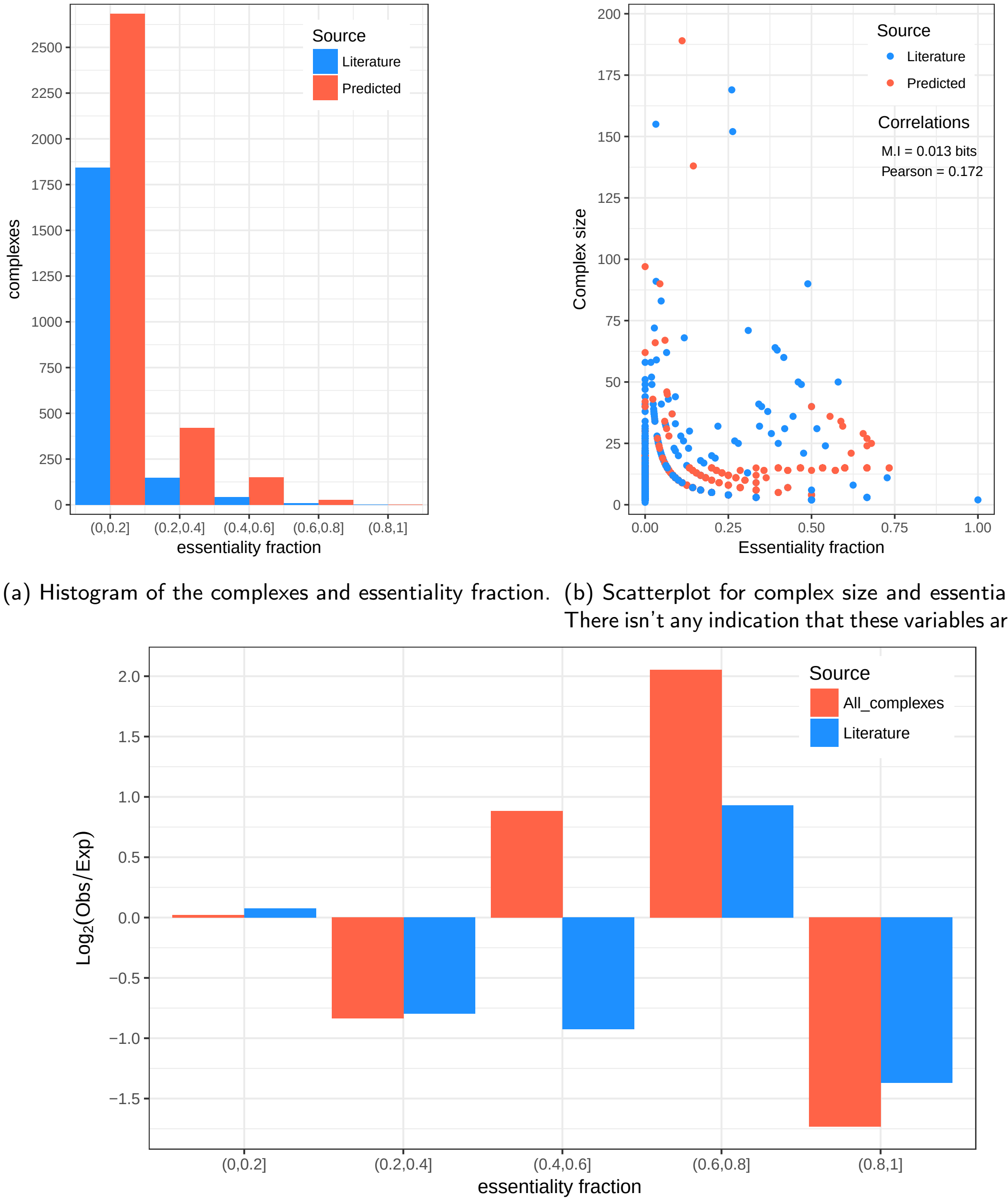

cell cycle G1/S phase transition
cellular protein metabolic process
organic substance metabolic process
inorganic cation transmembrane transport
proteasomal protein catabolic process
cellular protein catabolic process
ATP synthesis coupled proton transport
nucleoside phosphate biosynthetic proces...
proteolysis
ATP biosynthetic process
hydrogen transport
cellular metabolic process
protein metabolic process
cellular process
rine ribonucleoside m
energy coupled proton transport, down
purine nucleoside biosynthetic process
primary metabolic process
cellular response to stress
cellular catabolic process
cellular macromolecule metabolic process
cellular response to DNA damage stimulus
organic substance catabolic process
catabolic process
hydrogen ion transmembrane transport
inorganic ion transmembrane transport
G1/S transition of mitotic cell cycle
metabolic process
modification−dependent macromolecule cat...
nucleoside biosynthetic process
ribonucleoside biosynthetic process
nucleotide biosynthetic process
purine−containing compound biosynthetic ...
protein catabolic process
regulation of protein catabolic process
proteasome−mediated ubiquitin−dependent ...
nucleoside monophosphate biosynthetic pr...
purine ribonucleoside triphosphate biosy...
macromolecule catabolic process
purine ribonucleotide biosynthetic proce...
proteolysis involved in cellular protein...
glycosyl compound biosynthetic process
ribose phosphate biosynthetic process
ribonucleoside monophosphate biosyntheti...
nucleoside triphosphate biosynthetic pro...
purine nucleotide biosynthetic process
purine ribonucleoside biosynthetic proce...
uitin−dependent protein catabolic pr...
modification−dependent protein catabolic...
macromolecule metabolic process
cellular macromolecule catabolic process
purine nucleoside triphosphate biosynthe...
ribonucleotide biosynthetic process
regulation of catabolic process
purine nucleoside monophosphate biosynth...
proton transport
monovalent inorganic cation transport
ribonucleoside triphosphate biosynthetic...

Figure 0.9: Functional enrichment analysis.

Table 0.7: Comparison of the observed abundance of complexes in respect to essentiality fraction with a bootstrapped distribution

Type	Essentiality fraction	Number of complexes	Expected complexes	$Log_2 \frac{observed}{expected}$
All complexes	[0, 0.2]	4893	4823.73	0.021
	(0.2, 0.4]	222	396.383	-0.836
	(0.4, 0.6]	173	93.825	0.883
	(0.6, 0.8]	37	8.902	2.055
	(0.8, 1]	1	3.323	-1.732
Literature complexes	[0, 0.2]	1907	1810.119	0.075
	(0.2, 0.4]	90	156.544	-0.799
	(0.4, 0.6]	37	70.278	-0.926
	(0.6, 0.8]	9	4.7286432160804	0.928
	(0.8, 1]	1	2.58397365532382	-1.370



(a) Histogram of the complexes and essentiality fraction.

(b) Scatterplot for complex size and essentiality fraction. There isn't any indication that these variables are correlated.

(c) Log ratio .

Figure 0.10: sss.

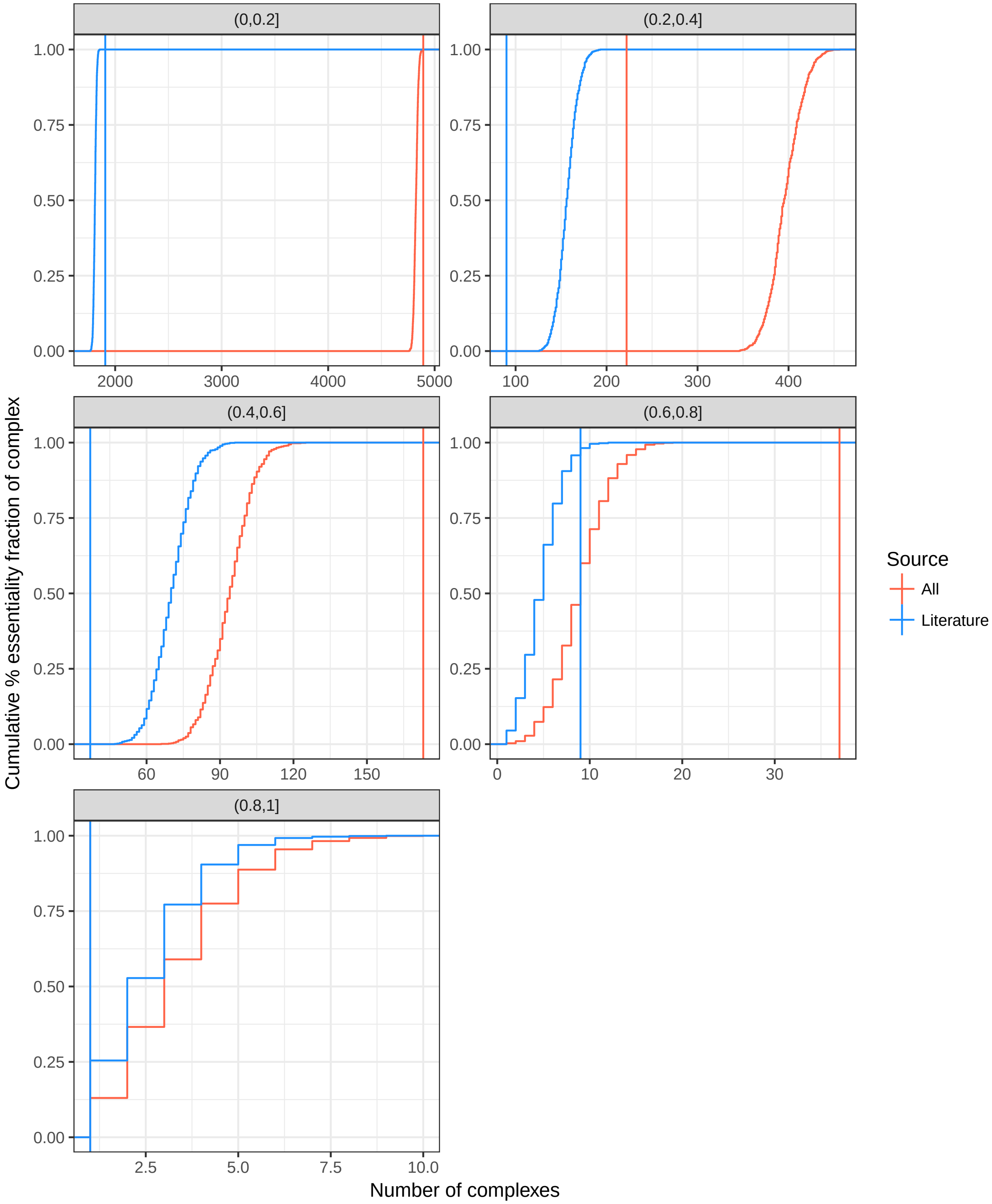


Figure 0.11: Cumulative distributions of bootstraped essentiality of the complexes. The vertical lines are the observed number of complexes belonging to the relavant bin.