This is an example thesis made with R Markdown.

Thank you for reading and good luck. For more information, see [rosannavanhespenresearch.wordpress.com](http://rosannavanhespenresearch.wordpress.com).

# Introduction

A gene/protein is essential if and only if its removal or disruption results in lethality or infertility of the organism.

There are 4 general problems conserning gene essentiality

1. Evolvability
2. Differential / conditional essentiality
3. Modular essentiality
4. Prediction

All these parameters are part of a bigger theme of biological challenges in the 21th century.

In this work we focused in points 3 and 4.

## Prediction of essentiality

History

Current

Centrality measures use the underlying topology of the network to determine node importance quantitatively.

## Modular essentiality

The basic work of (Ryan et al. 2013).

# Methods

We consider a directed, simple network with a set of nodes and an ordered set of edges . A node denotes a protein and an edge denotes a directed interaction from protein to protein . Each edge has been assigned a signed weight .

## Centralities

The essentiality consensus of a protein in the protein - protein interaction network is most commonly predicted by centrality measures (Jalili et al. 2016). In this work we used the degree, betweenness, weighted betweenness, closeness and information degree centrality. The historically first centrality used for the prediction of the essential proteins is the degree centrality in the influential paper (Jeong et al. 2001) which introduced the **centrality - lethality rule**. The degree centrality (DC) of a node v is defined as :

where deg(v) is the number of neighbors of node v.

Degree centrality predicts that hubs are more likely to be essential than non - hubs. This is a simplified view because there are essential proteins that are not hubs. Another classification of proteins in respect to network topology is to examine whether they are *bottlenecks*. Bottlenecks are the nodes that are located between highly connected clusters and their importance is measured through betweenness centrality (BC) (Freeman 1979; Joy et al. 2005; Yu et al. 2007). Betweenness centrality (BC) of a node v is defined as:

where is the number of all geodesic directed paths between all pairs of nodes, except pairs with v, and is the number of geodesics that pass through node v.

Weighted betweeness centrality (WBC) is defined as :

where the geodesic distance is , that is the minimum distance between nodes s and t is the path with the minimum sum of weights. In this implementation of betweenness, edge weights must be non negative numbers and higher values of weights have negative impact on path distance. Note that this is a crude method of handling weights that in our case isn't biologically appropriate but nevertheless we have included it in the analysis for comparison reasons.

Another centrality index we used is closeness centrality (CC) which is defined as :

And finally we computed the information centrality (IC) defined as :

The computations of the centralities were performed in R using the igraph package (Csardi and Nepusz 2006) except from the information centrality which was calculated manually.

## Decision trees

## Method comparison

In order to evaluate the performance of each method for predicting essentiality we used 3 methods, the precision - recall, the ROC curve and the Jackknife curve. All these methods use the statistical terms :

* True positives (TP) : essential proteins correctly predicted as essential
* False positives (FP) : nonessential proteins falsely predicted as essential
* True negatives (TN) : nonessential proteins correctly predicted as nonessential
* False negatives (FN) : essential proteins falsely predicted as nonessential

## Perron - Frobenius decomposition

## Gene ontology annotation

## Modular essentiality

# Results

## Signed network

## Protein metadata

In the signed network of **D. melanogaster** (Vinayagam et al. 2014) we annotated the essentiality data published in OGEE database (Chen et al. 2012). In table we can see the summary of this annotation.

## Essentiality prediction

The beauty of R Markdown: including your data analysis directly in the thesis, so that you update as you go.

### Centralities

And of course, we can also inlude figures (see figure ).

![The ROC curve for the different prediction methods of protein essentiality. ](data:application/pdf;base64,)

The ROC curve for the different prediction methods of protein essentiality.

AUC score of the different centralities as essentiality predictors.

|  |  |
| --- | --- |
| Centralities | AUC |
| Degree | 0.776 |
| Betweenneess | 0.591 |
| Abs weights betweenness | 0.604 |
| Closeness | 0.658 |
| Information degree | 0.541 |
| Decision tree C5.0 | 0.867 |
| Decision tree rpart | 0.881 |
| Decision tree C4.5 | 0.874 |

Here is the table from table .

### Decision trees

## Essential subgraph

## Modular essentiality

# Discussion

Finally, the discussion. We had a look at headings and citations in the introduction, an equation in the methods and figures, r code and tables in the results. That's it. What will follow now is the list of tables, list of figures, appendices and references.

Relabelling the appendix can be a bit tricky. Here I used the standard syntax '\appendix', and using LaTeX labels for sections instead of R Markdown syntax (i.e. using '\section' and '\subsection' instead of '#' and '##'). This automatically produces labels 'A.1', 'A.2', etc. I came up with doing it this way during the last two days before handing in my thesis, so I'm sure this is not the most elegant way of dealing with the appendices, but it works.

By setting eval = FALSE and echo = TRUE,   
the actual code will be displayed but not run.

# References

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. doi:[10.1093/nar/gkr986](https://doi.org/10.1093/nar/gkr986).

Csardi, Gabor, and Tamas Nepusz. 2006. “The Igraph Software Package for Complex Network Research.” *InterJournal* Complex Systems: 1695. <http://igraph.org>.

Freeman, Linton C. 1979. “Centrality in social networks conceptual clarification.” *Soc. Networks* 1 (3): 215–39. doi:[10.1016/0378-8733(78)90021-7](https://doi.org/10.1016/0378-8733(78)90021-7).

Jalili, Mahdi, Ali Salehzadeh-Yazdi, Shailendra Gupta, Olaf Wolkenhauer, Marjan Yaghmaie, Osbaldo Resendis-Antonio, and Kamran Alimoghaddam. 2016. “Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks.” *Front. Physiol.* 7 (August): 375. doi:[10.3389/fphys.2016.00375](https://doi.org/10.3389/fphys.2016.00375).

Jeong, H, S P Mason, a L Barabási, and Z N Oltvai. 2001. “Lethality and centrality in protein networks.” *Nature* 411 (6833): 41–42. doi:[10.1038/35075138](https://doi.org/10.1038/35075138).

Joy, Maliackal Poulo, Amy Brock, Donald E. Ingber, and Sui Huang. 2005. “High-betweenness proteins in the yeast protein interaction network.” *J. Biomed. Biotechnol.* 2005 (2): 96–103. doi:[10.1155/JBB.2005.96](https://doi.org/10.1155/JBB.2005.96).

Ryan, Colm J., Nevan J. Krogan, Pádraig Cunningham, and Gerard Cagney. 2013. “All or nothing: Protein complexes flip essentiality between distantly related eukaryotes.” *Genome Biol. Evol.* 5 (6): 1049–59. doi:[10.1093/gbe/evt074](https://doi.org/10.1093/gbe/evt074).

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](https://doi.org/doi:10.1038/nmeth.2733).

Yu, Haiyuan, Philip M. Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. 2007. “The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics.” *PLoS Comput. Biol.* 3 (4): 713–20. doi:[10.1371/journal.pcbi.0030059](https://doi.org/10.1371/journal.pcbi.0030059).