Discussion

In this study, we applied centrality indices to predict essential proteins in the signed protein interaction network of *D.melanogaster*. We found that degree centrality had the best performance with $AUC = 0.804$ and betweenness scored low with performance of $AUC = 0.591$. Both methods performed better that degree, $AUC = 0.435$, in the original PPI network of *D.melanogaster*. Actually, the latter result suggest that the centrality - lethality rule is very weak in the PPI network of *D.melanogaster*. To our knowledge, this result hasn't been made clear in literature. For example, in (Peng et al. 2015), the authors test the prediction power of their novel method and some centralities for essentiality consensus. They used the same network data from BioGRID (Stark et al. 2006) that we used. In Figure 6a of their manuscript they present the same results with ours. Their results indicate that the top 80 hub proteins contained only 5% of essential proteins and the 25% of proteins with the highest degree don't even contain 7.5% of essential proteins. Their novel method also never surpasses the prediction of 10% of the essential proteins of *D.melanogaster*. These prediction results suggest that degree and the other centrality methods are poor predictors of essentiality and question the applicability of the centrality - lethality rule in the protein interaction networks of *D.melanogaster*.

The centrality - lethality rule has been extensively tested in more than 10 organisms, most of them unicellular (Zhang, Acencio, and Lemke 2016), and this weak performance in *D.melanogaster's* PPI's hasn't been stated directly. There is a gap in the literature about *D.melanogaster* and the centrality-lethality rule since most studies focus on *S.cerevisiae*. Despite these limitations in the PPI network of *D.melanogaster* we found that centralities are good predictors in the signed network. Further tests are needed to decipher the reasons of these differences between these networks. The success of centralities in the signed network may appear due to biases towards essential proteins since it is a sub-network of the PPI of *D.melanogaster* by construction.

The best centrality measure for essential protein prediction is the local interaction density centrality (LIDC) which incorporates protein complex data and topological features (Luo and Qi 2015). This finding further validates the general argument that the integration of diverse data provides better results. Apart from data integration, it is generally accepted that method integration provides better results that individual methods (Cheng et al. 2014; Zhang, Acencio, and Lemke 2016). In this study, the best results for the prediction of essential proteins were from the decision trees when applied on centralities. As seen in Figure **??**, positive degree and weighted positive degree were used as decision rules to predict essential proteins. This is a property that characterizes essential proteins in the signed protein network of *D.melanogaster* which hasn't been reported before. As new signed networks will be created in the future this property shall be re-examined.

Besides high positive degree, we found that 36 essential proteins form a cluster with only positive interactions between them. Perhaps the result of the decision trees was driven by this essential cluster. The authors of (Zotenko et al. 2008) discovered that essential proteins tend to cluster together and that this behavior is due to their function in complexes. The result that these interactions are activation interactions is revealed in the signed protein network of *D.melanogaster* which is the first of its kind. Also the creators of the signed network (Vinayagam et al. 2014) and (Lin et al. 2013) have found that positive interactions are mostly found between proteins of the same complexes and negative interactions between proteins of different complexes. We also found that the essential cluster is reducible, and we reconstructed the network with its irreducible components. This result indicates that information flow is directed which decomposes the essential cluster in 3 parts, input, processing and output. Further analysis must be done to decipher the biological implications of this finding. In order to study further the essential cluster we enriched its proteins with terms of the biological process component of gene ontology. We found 4 types of processes but mostly the proteins participate in the proteasome and in the ATP biosynthesis. So there are activation interactions between essential proteins of these processes which indicates that the essential cluster is neither a component of a specific process nor a single protein complex. In spite of these previous findings, we ought to be careful because the positive interactions between essential proteins maybe because of bias, focus on specific proteins (Edwards et al. 2011), and missing interactions.

Nevertheless, we want to point that a lot of important theoretical work has recognized the positive - activation interactions for the emergence of self-organization (Corning 1995). In theoretical network population models, with activation-inhibition interactions, it has been observed that natural selection favors positive interactions (Mehrotra, Soni, and Jain 2009; S. Jain and Krishna 2001). Also using a more abstract model, the hyper-cycles, Eigen (Manfred Eigen 1971) suggested that cooperation was an essential step toward the emergence of complex and self-organized chemical systems (Sole 2011). And since essential genes are more conserved than nonessential (E. V. Koonin 2003; Mushegian and Koonin 1996) there might be an evolutionary explanation of their positive interactions. The connection of the essential cluster and the aforementioned theoretical work is very vague at the moment and more investigation is needed.

Although prediction of essentiality focuses on individual proteins, the work of (Hart, Lee, and Marcotte 2007) suggested that essentiality isn't a protein property. They indicated that protein essentiality is a byproduct of protein complex essentiality. This means that the lethality of the organism after the disruption of a protein is due to the malfunction of a complex that this proteins participates in. In 2013, (Ryan et al. 2013) referred to this hypothesis as *"All or Nothing"* which means that a complex will either contain mostly nonessential proteins or mostly essential proteins. They tested this by bootstrapping the proteins of complexes to create a null distribution of essentiality fraction (equation **??**). Their tests were performed on unicellular organisms. We followed their methodology for *D.melanogaster's* complexes and found similar results (Figure **??**). Specifically we found that in *D.melanogaster* there is only one complex with essentiality fraction above 0.8 which led to slightly different result. There are more complexes than expected with essentiality fraction in the intervals $[0, 0.2]$ and $(0.6, 0.8]$. In the other intervals the observed complexes are less than expected, also all the results were substantial (Figure **??**). In our analysis, we used ≈ 4 times more complexes that the previous studies which might be the reason of the slightly weaker results. The *"All or Nothing"* hypothesis should be rechecked when more reliable and rich data about complexes appear. Because at the moment identifying complexes, both experimentally and computationally[1], remains a huge challenge (Hartwell et al. 1999; Koch 2012).

In this work, we used centralities for the evaluation of the centrality - lethality rule in the signed PPI of *D.melanogaster*. Signed protein interaction networks are more relevant biologically because they contain activation - inhibition information which is a big step towards the understanding of cellular processes (Mitra et al. 2013; Ward, Sali, and Wilson 2013). In the future, it is important that more signed physical interaction networks will be constructed for other organisms (for example *S.cerevisiae*). With signed networks also comes the need to generalize the tools to analyse them in order to incorporate signs. For example, we found that positive weighted degree was an important predictor of essential proteins but more complex centralities like betweenness and closeness can't use signed weights. Another challenge is to detect experimentally the temporal nature of physical interactions in the cell (Gavin, Maeda, and Kühner 2011). Apart from the dynamic nature of protein interactions, they are also spatial which means that proteins function in specific locations in the cell (Aebersold and Mann 2016). In addition with the spatiotemporal information of protein interactions there is the differential nature of interactions. It is a fact that different enviromental conditions lead to radically different processes in organisms and consequently in

[1] During the analysis we discovered a bias towards small sized complexes in the COMPLEAT database (see Appendix **??**)

different network interactions (Ideker and Krogan 2012). Another challenge is to decipher the modular function of proteins because it has been discovered that proteins function by forming complexes. This finding adds a new scale of interactions, between comlexes, which creates a need for experimental advances as well as new network analysis tools to handle different network scales (Koch 2012; Coronges, Barabási, and Vespignani 2016). As a conclusion, the goal of devoloping of a complete representation of biological modules underlying cellular architecture and function is undoubtedly far. But the recent advances in experimental and analysis tools and more importantly the advances in the conceptual thinking are important steps towards this goal.

Aebersold, Ruedi, and Matthias Mann. 2016. "Mass-spectrometric exploration of proteome structure and function." *Nature* 537 (7620): 347–55. doi:10.1038/nature19949.

Cheng, Jian, Zhao Xu, Wenwu Wu, Li Zhao, Xiangchen Li, Yanlin Liu, and Shiheng Tao. 2014. "Training set selection for the prediction of essential genes." *PLoS One* 9 (1). doi:10.1371/journal.pone.0086805.

Corning, Peter A. 1995. "Synergy and Self-Organization." *Syst. Res.* 12 (2): 89–121. doi:10.1002/sres.3850120204.

Coronges, Kate, Albert-László Barabási, and Alessandro Vespignani. 2016. "Future directions of network science." Arlington, VA.

Edwards, Aled M., Ruth Isserlin, Gary D. Bader, Stephen V. Frye, Timothy M. Willson, and Frank H. Yu. 2011. "Too many roads not taken." *Nature* 470 (7333): 163–65. doi:10.1038/470163a.

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.

Hart, G Traver, Insuk Lee, and Edward R Marcotte. 2007. "A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality." *BMC Bioinformatics* 8: 236. doi:10.1186/1471-2105-8-236.

Hartwell, L H, J J Hopfield, S Leibler, and A W Murray. 1999. "From molecular to modular cell biology." *Nature* 402 (6761 Suppl): C47–C52. doi:10.1038/35011540.

Ideker, Trey, and Nevan J Krogan. 2012. "Differential network biology." *Mol. Syst. Biol.* 8 (565). Nature Publishing Group: 1–9. doi:10.1038/msb.2011.99.

Jain, S, and S Krishna. 2001. "A model for the emergence of cooperation, interdependence, and structure in evolving networks." *Pnas* 98 (2): 543–7. doi:10.1073/pnas.021545098.

Koch, C. 2012. "Modular Biological Complexity." *Science (80-.).* 337 (6094): 531–32. doi:10.1126/science.1218616.

Koonin, Eugene V. 2003. "Comparative genomics, minimal gene-sets and the last universal common ancestor." *Nat. Rev. Microbiol.* 1 (2): 127–36. doi:10.1038/nrmicro751.

Lin, Chen Ching, Chia Hsien Lee, Chiou Shann Fuh, Hsueh Fen Juan, and Hsuan Cheng Huang. 2013. "Link Clustering Reveals Structural Characteristics and Biological Contexts in Signed Molecular Networks." *PLoS One* 8 (6). doi:10.1371/journal.pone.0067089.

Luo, Jiawei, and Yi Qi. 2015. "Identification of essential proteins based on a new combination of local interaction density and protein complexes." *PLoS One* 10 (6): 1–27. doi:10.1371/journal.pone.0131418.

Manfred Eigen. 1971. "Self organization of matter and the evolution of biological macromolecules." *Naturwis-*

senschaften 58: 465–523. doi:10.1007/BF00623322.

Mehrotra, Ravi, Vikram Soni, and Sanjay Jain. 2009. "Diversity sustains an evolving network." *J. R. Soc. Interface* 6 (38): 793–9. doi:10.1098/rsif.2008.0412.

Mitra, Koyel, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. 2013. "Integrative approaches for finding modular structure in biological networks." *Nat. Rev. Genet.* 14 (10). Nature Publishing Group: 719–32. doi:10.1038/nrg3552.

Mushegian, A R, and E V Koonin. 1996. "A minimal gene set for cellular life derived by comparison of complete bacterial genomes." *Proc. Natl. Acad. Sci. U. S. A.* 93 (19): 10268–73. doi:10.1073/pnas.93.19.10268.

Peng, Xiaoqing, Jianxin Wang, Jun Wang, Fang Xiang Wu, and Yi Pan. 2015. "Rechecking the centrality-lethality rule in the scope of protein subcellular localization interaction networks." *PLoS One* 10 (6): 1–22. doi:10.1371/journal.pone.0130743.

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.

Sole, R. V. 2011. *Phase Transitions*. Princeton: Princeton University Press.

Stark, Chris, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. "BioGRID: a general repository for interaction datasets." *Nucleic Acids Res.* 34 (Database issue): D535–9. doi:10.1093/nar/gkj109.

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.

Ward, A. B., A. Sali, and I. A. Wilson. 2013. "Integrative Structural Biology." *Science (80-.).* 339 (6122): 913–15. doi:10.1126/science.1228565.

Zhang, Xue, Marcio Luis Acencio, and Ney Lemke. 2016. "Predicting essential genes and proteins based on machine learning and network topological features: A comprehensive review." *Front. Physiol.* 7 (MAR): 1–11. doi:10.3389/fphys.2016.00075.

Zotenko, Elena, Julian Mestre, Dianne P. O'Leary, and Teresa M. Przytycka. 2008. "Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality." *PLoS Comput. Biol.* 4 (8). doi:10.1371/journal.pcbi.1000140.