Inter-faculty Master Program on
Complex Systems and Networks

SCHOOL of MATHEMATICS
SCHOOL of BIOLOGY
SCHOOL of GEOLOGY
SCHOOL of ECONOMICS
ARISTOTLE UNIVERSITY of THESSALONIKI

Master Thesis

Title

The centrality lethality rule in signed protein interaction networks

Savas Paragamian

Supervisor: Stefanos Sgardelis, Professor of Biology, AUTH
Co-Supervisor: Ioannis Antoniou, Professor of Mathematics, AUTH
Co-Supervisor: Christoforos Nikolaou, Associate Professor of Biology, UOC

Thessaloniki, June 2017

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ
ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ στα
ΠΟΛΥΠΛΟΚΑ ΣΥΣΤΗΜΑΤΑ και
ΔΙΚΤΥΑ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΓΕΩΛΟΓΙΑΣ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τίτλος Εργασίας

Ο κανόνας κεντρικότητας – θνησιμότητας σε προσημασμένα πρωτεϊνικά δίκτυα

Σάββας Παραγκαμιάν

ΕΠΙΒΛΕΠΩΝ: Στέφανος Σγαρδέλης, Καθηγητής, Τμήμα Βιολογίας ΑΠΘ
ΣΥΝΕΠΙΒΛΕΠΩΝ: Ιωάννης Αντωνίου, Καθηγητής, Τμήμα Μαθηματικών ΑΠΘ
ΣΥΝΕΠΙΒΛΕΠΩΝ: Χριστόφορος Νικολάου, Επίκουρος Καθηγητής, Τμήμα Βιολογίας, Πανεπιστήμιο Κρήτης

Εγκρίθηκε από την Τριμελή Εξεταστική Επιτροπή την 6η Ιουνίου 2017.

.....	.....	.....
Στέφανος Σγαρδέλης Καθηγητής, Τμήμα Βιολογίας ΑΠΘ	Ιωάννης Αντωνίου Καθηγητής, Τμήμα Μαθηματικών ΑΠΘ	Χριστόφορος Νικολάου Επίκουρος Καθηγητής, Τμήμα Βιολογίας, Πανεπιστήμιο Κρήτης

Θεσσαλονίκη, Ιούνιος 2017

Copyright

. . . . . .

Σάββας Παραγκαμιάν Πτυχιούχος Βιολόγος Π.Κ.

Abstract

Essential are the genes/proteins which are indispensable for the organisms. A lot of research has focused on the identification of essential genes/proteins because they are considered part of the *minimal gene set*; they are possible drug targets for pathogens and more knowledge about them will contribute to the improvement of therapeutic strategies for human diseases. The experimental procedures for the detection of essential genes are expensive, laborious and in most cases unfeasible. Hence scientists have created tools for their prediction from other data using computational approaches. The most important results have come from centrality indices in protein interaction networks which formed the *centrality - lethality rule*. According to *centrality - lethality rule* the higher the interactions of a protein the more likely for it to be essential. Since its introduction, this rule has been expanded to other centralities and many novel methods have been developed that integrate a variety of data. Despite all these advancements, the protein - protein interaction network has largely remained the same. For a better representation of protein interactions, additional information should be taken into consideration like activation/inhibition, direction and molecular function. In this work, we used the first large scale signed protein interaction network, which was constructed using protein interaction and RNAi screen data for *D.melanogaster*, to predict essential protein using centrality indices. This revealed that when a protein has many activation interactions it is more likely to be essential.

Keywords: essential gene/protein, centrality lethality rule, signed protein networks, systems biology, protein complex

Contents

1 Introduction

1.1 Gene essentiality

A gene/protein is essential if and only if its removal or disruption results in lethality or infertility of the organisms. With the development of knock-out technics (Tatum and Lederberg 1947) scientists started studying the phenotypes of organisms after the removal of a gene (Gluecksohn-Waelsch 1963). These experiments are part of the genotype - phenotype problem and one of the strongest phenotypes to connect to a genotype is death or infertility. Until the late 1990's these experiments were performed on small scale so testing all possible gene deletions of an organism was incredibly laborious and in most cases impossible. In 1999, a large scale experiment was conducted and tested all genes of *S.cerevisiae* for essentiality consensus (Winzeler 1999). The large scale detection of essential genes was later performed for *D.melanogaster* using RNAi screens (Boutros et al. 2004). Currently there are protocols for the small and large scale exploration of essentiality in many organisms (Lu 2015). These large scale studies showed that in *S.cerevisiae* about $\approx$ 17% genes are essential whereas in *D.melanogaster* this number is only $\approx$ 2% (Chen et al. 2012). Generally, more complex organisms have higher proportion of essential genes when similar techniques are used. In the previous example with *S.cerevisiae* and *D.melanogaster* the techniques used were Genetic screens and RNAi screens, respectively.

The research interests for essential genes span across many disciplines. Since essential genes are indispensable for the organisms researchers study them in order to find the least possible number of genes to sustain life. These genes constitute what is called the minimal gene set Fraser et al. (1995); Mushegian and Koonin (1996)]. The research of the minimal gene set of an organism, in specific environmental conditions, has implications in the origins of life problem as well as synthetic biology (Koonin 2003; Koonin 2000). Minimal genome design has great biotechnology application prospects and one of the latest advancements in the field is the reduction of the genome of *Mycaplasma mycoides* from $1079kb$ pairs to $531kb$ pairs (Hutchison et al. 2016). Apart from the study of early life and synthetic biology, the study of essential genes is important for medicine. The essential genes of human pathogens are possible drug targets. More specifically the essential genes of pathogens that don't have orhtologs in humans are studied for the design of new drugs. Moreover, the knowledge of human essential genes and their functions will provide valuable information for the origin of diseases, like cancer, and novel therapeutic strategies (Zhan and Boutros 2016).

Even though so much research has focused on essential genes, the definition of essentiality has some caveats. From the beginning of the study of essentiality, the generality of the term was questioned, mainly because of the limited conditions tested in experiments (Gluecksohn-Waelsch 1963). The environmental conditions under which experiments are performed are crucial for the discovery of essential genes because in different conditions the essentiality is very likely to change (Zhang and Ren 2015). Nevertheless, the vast majority of essentiality data available today is derived from experiments in optimum conditions (D'Elia, Pereira, and Brown 2009). In 2015 a new parameter to essential genes was introduced, evolvability (Liu et al. 2015). The authors conducted evolutionary experiments in *S.cerevisiae* and discovered that the organism could overcome the lethal phenotype of the deletion of some of the essential genes by adaptive evolution. They found that 88 essential genes from the $\approx$ 1000 of *S.cerevisiae* can be dispensable through adaptive evolution

and thus they have to be distinguished from the other essential genes (Lieben 2015).

Another challenge of gene essentiality is to find its origins. In 2007, a large scale study was conducted to identify protein complexes in *S.cerevisiae*, (Hart, Lee, and Marcotte 2007). Protein complexes are modules formed by the gathering/assembly of proteins and are the functional machines of the cells. The finding of the above mentioned study was that complexes are composed of mostly non essential or mostly essential proteins. This finding suggests that essentiality is modular which means that the deletion of gene is lethal because it results to malfunction of an essential protein complex. This result further supports the concept of modularity in biological functions which states that the vast majority of biological functions are facilitated from modules, collections of different molecules that interact (Hartwell et al. 1999; Koch 2012). This is a conceptual transition of molecular biology which studies molecules to what Hartwell called *modular biology* which studies ensembles of molecules and their interactions, the modules (Hartwell et al. 1999). The identification of complexes requires first the identification of protein interactions with tandem affinity purification (TAP) of affinity-tagged proteins followed by mass spectrometry and then computational analysis (Krogan et al. 2006).

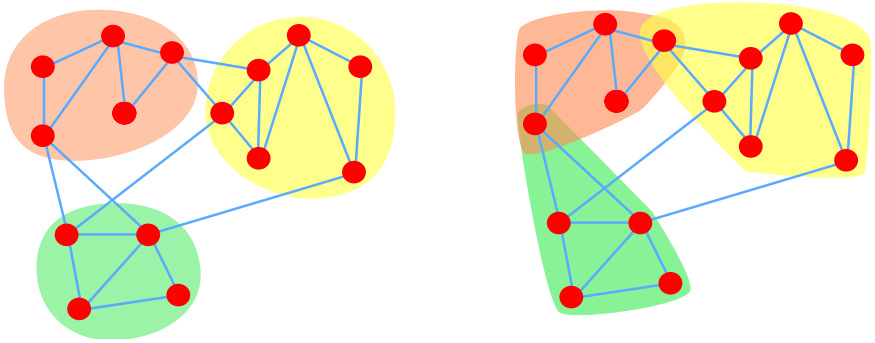


Figure 1.1: Different approaches to network communities inference. Left: modularity - based methods. Right: overlapping communities.

The computational problem of complex identification is similar to the community detection problem in networks. The difference is that protein complexes are represented as overlapping communities so modularity based methods are not suitable (see Figure 1.1) (Brohée and Helden 2006; Newman 2006). With the accumulation of the aforementioned protein interactions and protein complexes data for other organisms, (Ryan et al. 2013) showed that modular essentiality appears to other unicellular organisms and that some complexes between organisms appear to change completely their essentiality status, which further supports the *"All or Nothing"* hypothesis. In this work we tested whether the *"All or Nothing"* hypothesis holds in a more complex organism, *D.melanogaster*.

1.2 Prediction of essentiality

The experimental discovery of essential genes remains a very laborious procedure even though much progress has been made. In most occasions it's actually impossible to conduct these experiments, so methods for the prediction of essential genes have been developed. After the first genome sequencing projects, comparative

genomic methods were used to compare essential genes. These methods use homologous genes to predict essential genes because highly conserved genes are more likely to be essential (Jordan et al. 2002). Although these methods are generally reliable, two limitations have been observed. First, conserved orthologs between species account for a small portion of a genome. Second, orthologs in distantly related species often exhibit differences in gene regulation, function and complexes, leading to a potential diversity of gene essentiality. To circumvent these limitations, researchers have developed feature-based methods that can be used to distinguish essential genes from non-essential ones based on the presence of features similar to those of essential genes (Cheng et al. 2014; Zhang, Acencio, and Lemke 2016).

The feature-based methods can be divided in two categories, machine learning and network analysis methods. These methods use data from large scale experiments like gene expression, RNAi screens, flux balance analysis, protein - protein interactions experiments (Zhang, Acencio, and Lemke 2016). The most striking result came in 2001 when the authors of (Jeong et al. 2001) discovered that hubs (proteins with the highest number of neighbors) are more likely to be essential. This seminal research was done on the protein interaction network of *S.cerevisiae* and introduced the *centrality - lethality rule*. Centrality indices are quantitative measures which use the underlying topology of the network to determine node importance (Freeman 1979). Degree (number of neighbors of a node) even though is a very simple centrality opened the research of other centralities for the detection of essential proteins. Later it was recognized that not only protein hubs are essential but also proteins with high betweenness (Joy et al. 2005). Betweenness highlights the nodes of a network that act as *"intersections"* of information flow, like bottlenecks (Shaw 1954; Freeman, Borgatti, and White 1991; Newman 2005). High betweenness can detect essential proteins that act like *bottlenecks* as seen in Figure 1.2 (Yu et al. 2007). Researchers afterwards began to integrate data, for example gene expression, protein subcellular location and other data, into novel centrality indices to improve the performance of the prediction as well as the diverse applicability to different organisms (Jalili et al. 2016).

The integration of diverse data to centrality indices has proved to be effective. Proteins function in specific sub-locations in the cells and this creates a specific interaction environment. Sub-location information of proteins from various databases was used to create a novel centrality index to predict protein essentiality on protein interaction networks (Peng et al. 2015). To our knowledge, the most effective method yet for predicting essential proteins is the combination of protein complex information with centralities (Luo and Qi 2015) which further supports the modular nature of essentiality. In addition to data integration, method integration resulted to better predictions of essentiality. More specifically, application of machine learning approaches on centralities showed even better results than individual methods (Zhang, Acencio, and Lemke 2016).

The network approach methods for essentiality prediction have been tested, combined and expanded the last decade. The underlying networks are protein networks which have been updated with more reliable interactions. But only $\approx 20\%$ of the total protein interactions have been found for the organisms that protein networks have been constructed (Aebersold and Mann 2016). Protein interaction networks are important because across the species, proteins constitute about 50% of the dry mass of a cell and reach a remarkable total concentration of 2–4 million proteins per cubic micrometer or 100–300 mg per ml (Aebersold and Mann 2016). However, little progress has been made to decipher the function of these

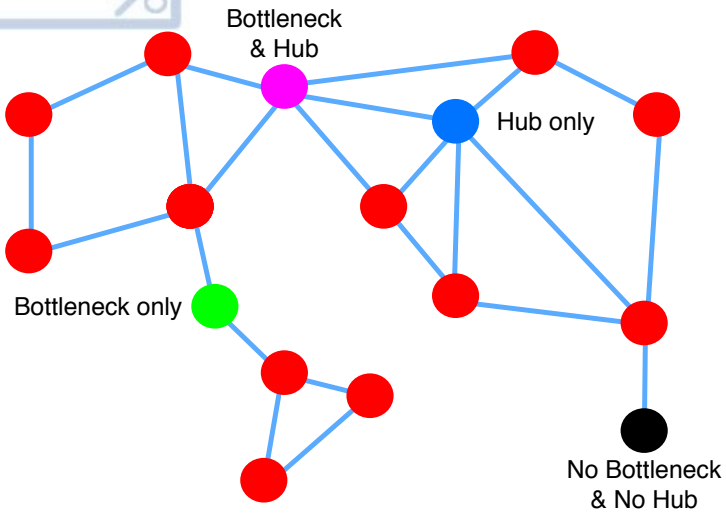


Figure 1.2: A schematic representation of the difference between hubs and bottlenecks.

physical protein interactions.

It's important to infer causal relationships between interacting proteins in large scale. These relationships have direction (edge direction), sign (activation/inhibition), weight (strength of activation/inhibition) and mode (e.g. phosphorylation, ubiquitination). That way the protein interaction network would contain signal flow information. A lot of progress has been made to computational methods that predict the direction of the protein interactions. Manually curated pathways from KEGG (Kanehisa and Goto 2000) and REACTOME (Croft et al. 2014) databases is rich. In addition many computation methods have been developed to predict the directions of interaction in *S.cerevisiae* (Gitter et al. 2011), in *H.sapiens* (Vinayagam et al. 2011) and other organisms. Despite all this work for the directionality prediction little progress has been made for the sign prediction of protein interactions. The first protein interaction network with signs and weight was created for *D.melanogaster* using RNAi phenotypes and protein interactions (Vinayagam et al. 2014). Although smaller that the original protein interaction network of *D.melanogaster*, the signed protein network contains activation / inhibition information which is an important step towards the better representation of cellular processes.

1.3 Aim of this study

The aim of this study was to improve the prediction of essential genes / proteins by extending the centrality – lethality rule to signed protein interaction networks.

2 Methods

2.1 Centralities

We consider a undirected, simple network $G(V, E)$ with a set of N nodes V and an ordered set of edges E. A node $v \epsilon V$ denotes a protein and an edge $e(v, u) \in E$ denotes a interaction from protein v to protein u. Each edge has been assigned to a signed weight $w_{v,u} \in [-1, 1]$.

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 :

$$DC(v) = deg(v), \tag{1}$$

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.

Because the network is signed we can further distinguish the degree to positive and negative degree. Positive degree (PD):

$$PD(v) = deg^{+}(v), \tag{2}$$

where $deg^{+}(v)$ is number of nodes the have positive interactions with node v.

Also because the network is weighted we define Positive Weighted Degree Centrality (PWDC) as:

$$PWDC(v) = \sum_{u}^{N} (w^{+}_{v,u} + w^{+}_{u,v}), \tag{3}$$

where $w^{+}_{v,u}$ are the positive weights from node v to its u neighbors and $w^{+}_{u,v}$ is the reverse.

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:

$$BC(v) = \sum_{s \neq t \neq v \in V} \frac{g_{st}(v)}{g_{st}}, \tag{4}$$

where g_{st} is the number of all geodesic paths between all pairs of nodes, except pairs with v, and $g_{st}(v)$ is the number of geodesics that pass through node v.

Edge betweenness is defined similarly to node betweenness.

$$BC(e) = \sum_{s \neq t \neq v \in V} \frac{g_{st}(e)}{g_{st}}, \tag{5}$$

where $g_{st}(e)$ is the number of shortest paths that pass through the edge named e.

Weighted betweenness centrality (WBC) is defined as :

$$WBC(v) = \sum_{s \neq t \neq v \in V} \frac{g^w_{st}(v)}{g^w_{st}}, \tag{6}$$

where the geodesic distance is $g^w_{st} = min(\sum w_{st})$, 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. Hence, we used the absolute values of edge E weights of G. 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 :

$$CC(v) = \sum_{v \neq t \in V} \frac{1}{g_{v,t}}, \tag{7}$$

Finally, we computed the information centrality (IC) defined as :

$$IC(v) = -log_2 \frac{n(DC(v))}{N}, \tag{8}$$

where $n(DC(v))$ is the number of nodes that have $degree = DC(v)$.

The computations of centralities were performed in R using the igraph package (Csardi and Nepusz 2006) apart from the information centrality which was calculated manually on the original network as well as its giant component.

2.2 Decision trees

Decision trees are supervised machine learning tools used to build classification models (Kotsiantis 2013; Quinlan 1986; Kabacoff 2011). We implemented decision trees on the centrality measures mentioned above to test if the integration of centralities provides better results than single centrality indices for the prediction of essential proteins. Furthermore, we excluded the proteins from the giant component that weren't annotated with essentiality consensus (NA consensus). We used three algorithms, the algorithm in

the rpart package (Therneau, Atkinson, and Ripley 2017), the C4.5 algorithm from the J48 function in RWeka package (Hornik, Buchta, and Zeileis 2009) and the latest algorithm C5.0 from the C5.0 package (Kuhn et al. 2015). After the tree creation each protein was assigned probabilities of essentiality from the 3 different algorithms.

2.3 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 (Holman et al. 2009; Manning, Prabhakar, and Schutze 2008). 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

These terms form the confusion matrix of a binary classifier which in our case is essentiality consensus and are used to calculate the following fractions :

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$False\ Positive\ Rate = \frac{FP}{TP + FN} \tag{11}$$

Precision (Equation 9) is the ratio of the number of correct predictions to the total number of predictions. On the other hand recall (Equation 10) is the ratio of the number of correct predictions to the total number of possible correct predictions. Using these measures we can plot the Precision - Recall curve through an iterative process. In the first iteration k top ranked proteins (in terms of a variable, i.e degree) are retrieved and the precision and recall are measured. In the next iteration $k+1$ proteins are retrieved, if the protein is nonessential then recall remains the same but precision decreases. If the protein is essential then both recall and precision increase.

False positive rate (Equation 11) is the ratio of the wrong predictions to the total number of possible correct predictions. This measure and the recall measure, also called true positive rate, are plotted to create the receiver operating characteristic curve (ROC curve). The ROC curve of a random predictor is the $y = x$ line, any predictor above this line is considered better. The area under ROC curve is called AUC. The ROC curve is plotted with similar way as the precision - recall curve. Both methods were computed using the ROCR package (Sing et al. 2005).

The Jackknife curve was first presented in (Holman et al. 2009) and is a simple alternative method to

evaluate predicting tools for binary classifiers. In our case it expresses the relationship between the number of essential proteins in respect to the number of top ranked proteins retrieved based on a variable. This curve is created by incrementally increasing the number of retrieved proteins and the theoretical 100% successful model is plotted in the $y = x$ line.

2.4 Predict edge orientation

To infer the direction of the interactions we followed a methods similar to (Vinayagam et al. 2011). We used the naive Bayes algorithm included in the e1071 r package (Meyer et al. 2017). The training set that we created was the combination of manually curated directed protein interactions (Fazekas et al. 2013) with the *D.melanogaster* interlogs of the *H.sapiens* directed protein interaction network (Murali et al. 2011; Vinayagam et al. 2011). We duplicated the edge list of these directed interactions and used them as training data with also the edge betweenness, signed score and co express coefficient are classification variables. Then we applied the model to our network. to predict the interactions orientation.

2.5 Frobenius decomposition theory

The topology structure of a network is possible to reflect its function. The work of Frobenius and Perron on matrices can provide some useful insights when implemented on directed graphs. The following definitions and theorems are well documented with proofs and further details in the books of (Varga 2000) and (Gantmacher 1987). The Frobenius graph decomposition can illustrate the flow of information in a directed network. If a network is strongly connected as defined in Definition 1 then the information can reach all nodes from all nodes. This means that there is no distinction between nodes or clusters in the network, in terms of information distribution. In addition, we can explore further the inner structure of a strongly connected component using the paragraph 5 of Theorem 2[1]. After the calculations of the eigenvalues we can evaluate if there are more than one eigenvalues that equal to spectral radius of the component. If this is true, then a cycle in the component exists and its permuted adjacency matrix takes the form of matrix 13.

If the network is weakly connected, or equally its adjacency matrix is reducible as stated in Theorem 1, then we have to find its strongly connected components. The most efficient algorithm to perform this task was developed by (Tarjan 1971) and is included in the *Graph BOOST Library* (Siek, Lee, and Lumsdaine 2001) which has an interface in R (Carey, Long, and Gentleman 2016). By implementing Tarjan's algorithm we identify the network's strongly connected components and single nodes that aren't participating in any strongly connected component. After we can partition the network into tree components:

1. Input: Nodes that have only out edges
2. Processing: Nodes that have incoming and outgoing edges
3. Output: Nodes that have only incoming edges

This structure indicates that information flow is directed in the network.

[1]Theorem 2 is the famous theorem that was proved independently from Perron in 1907 for positive matrices and from Frobenious in 1912 for non-negative matrices.

Definition 1 (Strongly connected) *A directed graph with n nodes is strongly connected if, for any ordered pair (P_i, P_j) of nodes with $1 \leq i, j \leq n$, a direct path connecting P_i to P_j exists.*

Theorem 1 *An $n \times n$ complex matrix A is irreducible if and only if its directed graph $G(A)$ is strongly connected.*

Definition 2 (Reducibility) *A $n \times n$ complex matrix A, is reducible if there exists a $n \times n$ permutation* [2] *matrix such that A takes an upper triangular form:*

$$PAP^T = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}, \tag{12}$$

where B and D are square matrices. If there is not such permutation then A is irreducible. In case A is reducible and B or D are also reducible then they are further permutated to components. This process is repeated for as many times needed for all the upper triangular components of A to be irreducible.

Theorem 2 (Frobenius, 1912) *When A is a square, nonnegative and irreducible matrix then :*

1. *A has a positive real eigenvalue equal to its spectral radius, r.*
2. *To r there corresponds an eigenvector $x > 0$.*
3. *r increases when any entry of A increases*
4. *r is a simple eigenvalue of A*
5. *if A has has h eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_h$ equal to its spectral radius r ($|\lambda_1| = |\lambda_2| = \cdots = |\lambda_h| = r$) and $h > 0$, then A can be permuted to the following "cyclic" form:*

$$PAP^T = \begin{pmatrix} O & A_{1,2} & O & \cdots & O \\ O & O & A_{2,3} & \cdots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \cdots & A_{h-1,h} \\ A_{h,1} & O & O & \cdots & O \end{pmatrix}, \tag{13}$$

where there are square blocks along the main diagonal.

Definition 3 (Primitive matrix) *If a irreducible matrix $A \geq 0$ has h eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_h$ equal to its spectral radius r ($|\lambda_1| = |\lambda_2| = \cdots = |\lambda_h| = r$), then A is called **primitive** if $h = 1$ and **imprimitive** if $h > 1$. In the latter case h is called index of impimitivity of A.*

2.6 Enrichment analyses

We performed gene ontology singular enrichment analysis in order to decipher the biological processes that are over-represented in our protein set (Ashburner et al. 2000; Rhee et al. 2008). We used R bioconductor packages AnnotationDbi (Pagès et al. 2017) and org.Dm.eg.db (Carlson 2016b) for *D.melanogaster's*

[2] Permutation matrix is a square matrix that has one entry unity in each row and column and zeros elsewhere.

protein ID conversion, GO.db (Carlson 2016a) for protein ID mapping on gene ontology terms and topGO (Alexa and Rahnenfuhrer 2016) to facilitate Fisher's exact test for over-representation of biological process terms. Fisher's exact test uses a background distribution of GO terms and occurrences that is compared with a specific test. In our case, we used all protein IDs of the signed network of *D.melanogaster* (Vinayagam et al. 2014) as a background to test a subset of this network with essential interacting proteins (Figure 3.5). From the statistical test we obtained the biological process terms associated with a p-value, a Bonferroni correction and FDR. We choose to use the simple p-value at a=0.5 significance level. It is worth noticing that Fisher's exact test as well as other similar tests (i.e hypergeometric test) share the same assumption for the null hypothesis, that the probabilities for the selection of each gene are equal (Rivals et al. 2007). However, it turns out that the probabilities for the selection are not equal because the structure of gene's ontology bipartite network of genes and gene terms has a heavy tail degree distribution and hence these tests are biased to high degree terms (Glass and Girvan 2014). The authors of (Glass and Girvan 2014) created an algorithm that escaped this bias.

Singular enrichment analysis results in a long format table with one column representing the statistically significant GO terms and another column with the protein IDs. This can be considered as a bipartite network with the 2 sets of nodes being GO terms and the protein IDs belonging to them. By projecting the bipartite network to the one-mode network of GO terms we investigate the functional relationships between GO terms. This analysis is called functional enrichment analysis.

We also performed KEGG pathway annotation (Kanehisa and Goto 2000). We used the r bioconductor packages KEGGREST (Tenenbaum 2017) for the annotation and pathview (Luo et al. 2013) for the visualisation of the pathways.

2.7 Modular essentiality

Each protein complex has many proteins and each protein can participate in many complexes. How are the essential proteins distributed amongst complexes? In order to answer this question we have to do a statistical test with the hypothesis claiming that the distribution of essential proteins in complexes is random. The null distribution was created using the bootstrap procedure. We performed sampling with replacement to the essentiality consensus of the proteins of complexes for 1000 rounds using the sample() function of base R. That way complexes had always the same size. After we calculated the essentiality fraction (EC) of a complex c_i which is defined as:

$$EC(c_i) = \frac{number\ of\ essential\ proteins\ in\ c_i}{total\ proteins\ of\ c_i} \in [0,1] \tag{14}$$

$EC(c_i)$ was calculated for the original data and for each one of the 1000 permutations. We then sorted the complexes in 5 equally sized bins according to their essentiality fraction. Afterwards, for each bin of the original data and the 1000 permutations we counted the included complexes. Hence, for each bin we had a null distribution for hypothesis testing and one-tailed p-value calculation. We next calculated the mean number of complexes in each bin of the permutations in order to compare the expected with the observed number of complexes. The comparison was made with the log ratio:

$$Log - ratio(bin(EC)) = log_2(\frac{number\ of\ complexes \in bin(EC)}{mean\ estimated\ number\ of\ complexes \in bin(EC)}) \quad (15)$$

2.8 Tools

All the calculations and analyses were done in R (R Core Team 2016) using the R Studio (RStudio Team 2016) interface. Data handling and manipulation were performed with the packages dplyr (Wickham and Francois 2016), tidyr (Wickham 2017) and readr (Wickham, Hester, and Francois 2017). Data visualization was done with the packages ggplot2 (Wickham 2009) and ggraph (Pedersen 2017) and graphic design of Figures 1.2 and 1.1 was done with AUTODESK® GRAPHIC application. In addition, all scripts were written in rmarkdown (Allaire et al. 2017) with text alongside the code so all results are easily reproducible (Peng 2011; Piccolo and Frampton 2016). The machine used is a late 2013 model Macbook Pro with 13" retina screen, 2.4GHr Intel Core i5 processor, 8GB RAM memory and macOS Sierra operating system. The thesis was conducted in R Studio using rmarkdown and LaTeX.

3 Results

3.1 Data

3.1.1 Networks

From the BioGRID database (Chatr-Aryamontri et al. 2015; Stark et al. 2006), version 3.4.148, we downloaded *D.melanogaster's* protein-protein interaction (PPI) network. All physical interactions were selected (Table 3.1) and the giant component of the network was used as a benchmark.

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 (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 (Vinayagam et al. 2014) are freely available to download.

Table 3.1: Summary of the PPI network and the signed PPI network of *D.melanogaster*

Type	All *D.melanogaster* network	Giant component of *D.melanogaster* network	Complete signed network	Giant component
Proteins	8103	8006	3352	3058
Interactions	38364	37011	6094	5930
Positive	0	0	4109	3998
Negative	0	0	1985	1932

The authors of (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). Some previously unknown relationships were also unraveled that were later confirmed experimentally, a result that showed the high predictive power of the approach.

The integration of signs to the protein interaction network of *D.melanogaster* didn't come without a cost (Vinayagam et al. 2014). As seen in the Table 3.1 only $\approx 40\%$ of the original proteins are included and even less, $\approx 16\%$, 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* (Vinayagam et al. 2014).

The interactions between proteins of the signed protein interaction network are both directed and signed. The signs take scores in the interval $[-2.645751, 4.123106]$ as seen in the density plot (Figure 3.1). 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 (Vinayagam et al. 2014). We also found that there were

Table 3.2: 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

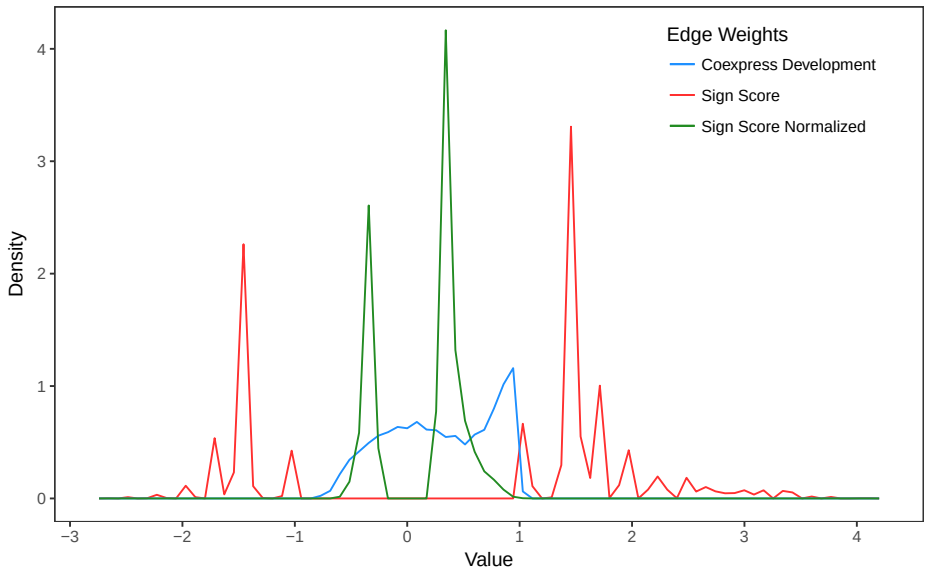


Figure 3.1: Density of signed weights and gene expression correlation. The original signs we further normalized by dividing all values with the maximum absolute value in order for the distribution to lie in the $[-1, 1]$.

31 duplicated interactions which is due to the inclusion of signs from literature. From these interactions we kept the ones from literature for our analysis.

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 is not considered a good practice (Ou-Yang, Dai, and Zhang 2015). Nevertheless the authors (Vinayagam et al. 2014) compared gene expression time-course data with the signs predicted from their methods and found big differences, 2158 interactions have the opposite sign score compared with the gene expression correlation (Table 3.2).

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 3.2) (Barabási and Albert 1999). For the rest of this article when we refer to the network we will mean its giant component.

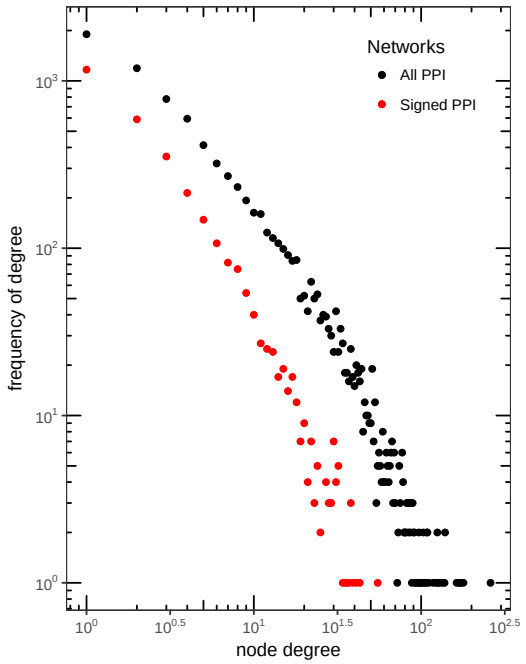


Figure 3.2: The degree distribution of the network is scale-free.

3.1.2 Protein essentiality

To annotate the proteins of the signed network 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 3.3 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. 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 3.3). 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.

3.1.3 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

Table 3.3: Gene essentiality consensus from OGEE database

Consensus	All *D.melanogaster's* proteins	Giant component of *D.melanogaster*	Complete signed network	Giant component
Nonessential	13373	7224	3009	2737
Essential	267	215	154	146
Conditional	141	73	33	29
NA	0	494	156	146
Total	13781	8006	3352	3058

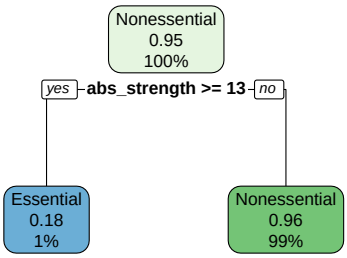
COMPLEAT database (Vinayagam et al. 2013). COMPLEAT database has freely available data and also provides a platform for analyses for various types of data. We downloaded the protein complexes of *D.melanogaster* and their proteins. There are 2 types of complexes in COMPLEAT, those collected from individual experiments referred to as *literature* and those inferred from 2 algorithms, *CFinder* and *NetworkBlast*. When we plotted the distribution of complex's size in terms of number of containing proteins we saw a pattern (Figures A.2c and A.2a). NetworkBlast, which predicted $\approx 50\%(2893)$ of the complexes (Table A.1), has a upper limit of 16 proteins in complex size (Figures A.2c and A.2d). 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. More information about the COMPLEAT database and the bias we observed is discussed in Appendix A.

3.2 Evaluation of essentiality prediction methods

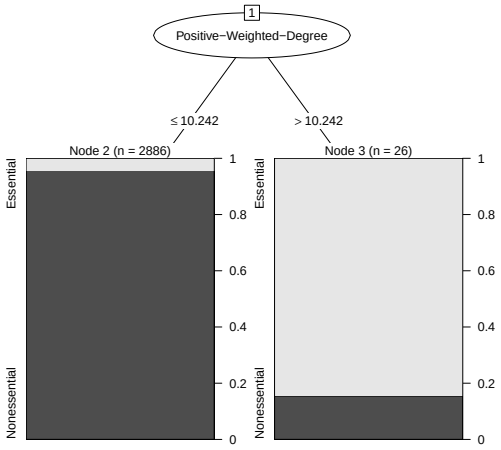
After the calculation of centrality indices for all network proteins we created decision trees for essentiality consensus prediction. We chose the centralities as variables for decision rules from which we constructed three trees using the algorithms from rpart package, C4.5 and C5.0. The C4.5 algorithm created a tree with higher complexity, more branches, than the rpart and C5.0 algorithms (Figure 3.3c). The C4.5 algorithm had also better precision, because it had less false positives but lower recall than the other algorithms (Table 3.4, Equations 9 and 10). 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 3), positive degree (Equation 2) and betweenness (Equation 4) 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 3.4: Confusion matrix for the 3 different algorithms of decision trees.

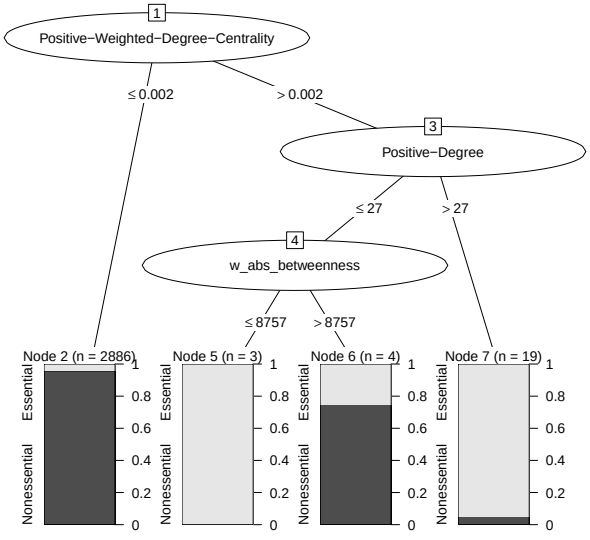
Type	C5.0	rpart	C4.5
True Positives	22	23	21
False Negatives	124	123	125
True Negatives	2762	2761	2765
False Positives	4	5	1
Precision	0.846	0.821	0.955
Recall	0.151	0.158	0.144



(a) rpart package algorithm



(b) C5.0 algorithm



(c) C4.5 algorithm

Figure 3.3: Trees from different algorithms. **(a)** and **(b)** generated oversimplified trees but **(c)** generated a little more complex and more precise tree.

We used ROC curve, Precision Recall curve and Jackknife curve to compare the predictability power of centralities and decision trees (Figure 3.4).

In the original PPI network of *D.melanogaster* degree scores very low although in the signed network is a good essentiality predictor. In Figure 3.4 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 algorithm ($AUC = 0.874$). Degree centrality is the best performed centrality with $AUC = 0.804$. Worth mentioning is the low performance of betweenness centrality ($AUC = 0.609$) and closeness centrality ($AUC = 0.673$). In the Jackknife curve (Figure 3.4b) 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. Even though decision trees had faster essential protein accumulation (i.e higher precision) they reached a plateau in 23 essential proteins. Furthermore, closeness centrality eventually and gradually reached the top methods in correct essential protein prediction.

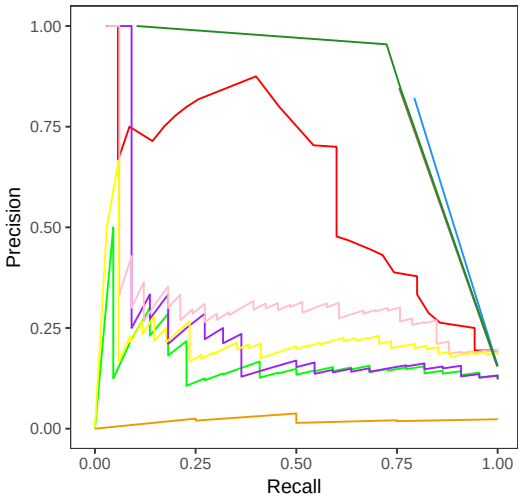
3.3 Essential subgraph

We investigated the subgraph of essential proteins of the signed network which contains only essential proteins and their interactions (Figure 3.5). 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 and KEGG pathways annotation.

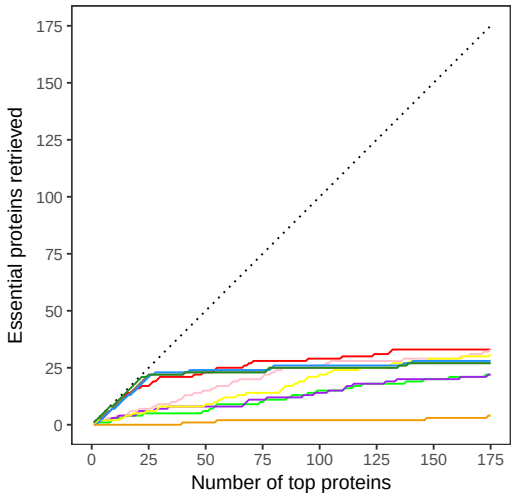
3.3.1 Decomposition of essential cluster

In order to investigate the inner structure of the essential cluster it is important to know the direction of the links. The signed network has only the signs of the interactions (Vinayagam et al. 2013), hence we have to add the network. For the inference of the direction of protein interactions we used the manually curated database SignaLink 2 (Fazekas et al. 2013). SignaLink 2 has 366 directed protein interactions between 180 protein of *D.melanogaster* (Fazekas et al. 2013). Because this is a small set we used the *H.sapiens* directed protein interaction that was created by (Vinayagam et al. 2011). In order to use this network we downloaded the protein interlogs between *D.melanogaster* and *H.sapiens* from DroID database (Murali et al. 2011). Interlogs are the conserved interactions between organisms (Walhout et al. 2000). We assumed that interlogs keep also the same interaction orientation although this is not always the case (Korcsmaros et al. 2011). The combination of the *H.sapiens* directed protein interactions and interlogs resulted in 34,270 directed interactions between 2756 proteins. In the signed protein network the 3302 of these interactions were present. Finally, we combined the data from SignaLink with the directed interlogs and we got 3474 directed interaction in the signed network of *D.melanogaster*. These were our training set for the naive Bayes model.

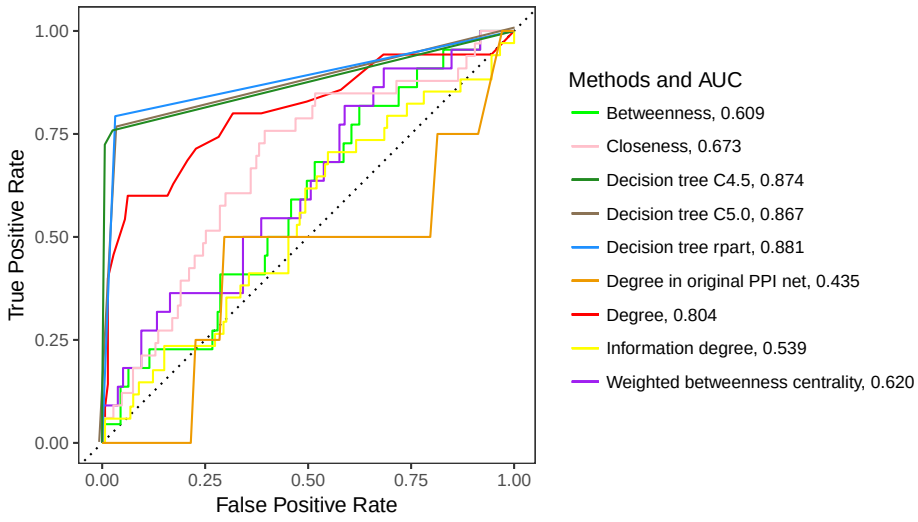
We duplicated the edgelist of the giant component, to the total of 11860, so we had all possible directions



(a) Precision Recall curve.



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



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

Figure 3.4: Evaluation methods for the different prediction methods of protein essentiality. **(a)** rpart and C5.0 methods have the above curves because they generated trees with one decision rule (Figure 3.3).

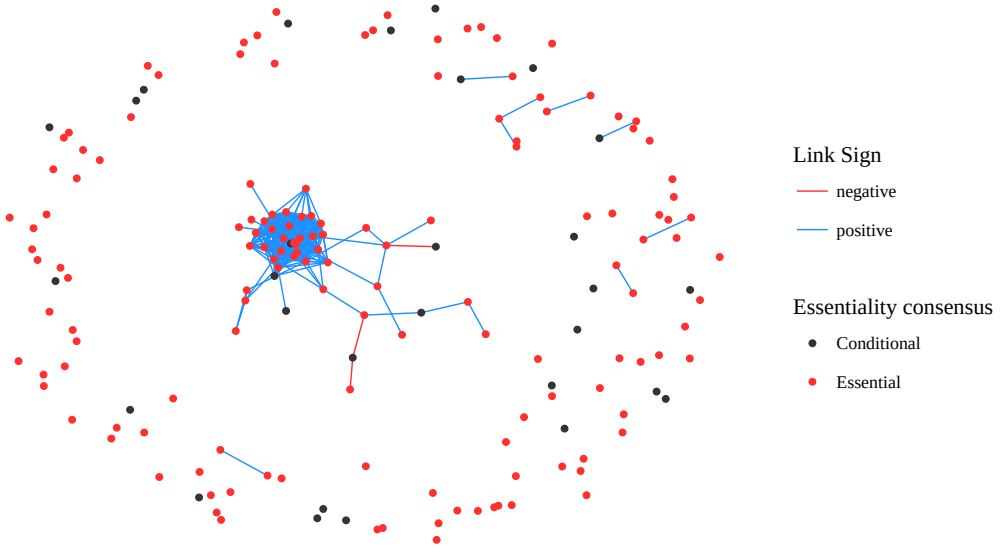


Figure 3.5: Interactions between essential proteins are positive. The only negative interactions are from conditionally essential proteins.

of the interactions. Using the aforementioned training data set in the naive Bayes algorithm we were able to predict the direction of the rest of the interactions using the sign score, co expression coefficient and edge betweenness as classification variables. To conclude, we predicted the edges directions of the signed network which resulted in 198 directed interactions and 27 proteins in the essential cluster. Our implementation had 0.5 precision and 0.3 recall which means that improvements are due in the future.

Next we decomposed the essential cluster using the Frobenius Decomposition Theory. The goal of this graph decomposition decipher how the information flows in the directed network. Using Tarjan's algorithm (Tarjan 1971) we found that there isn't a strongly connected component (or equivalently the irreducible component, Theorem 1) in the essential cluster (Table 3.5). All the 27 proteins are part of a reducible component (Figure 3.6). Information in the strongly connected component can reach all proteins from any protein in the component (Definition 1).

Table 3.5: Essential cluster information

Type	Values
Essential proteins in network	146 + 29 conditional
Connected essential cluster	36 proteins, 243 positive interactions
Directed essential cluster	27 proteins, 197 directed positive interactions
Irreducible component	Only singular components

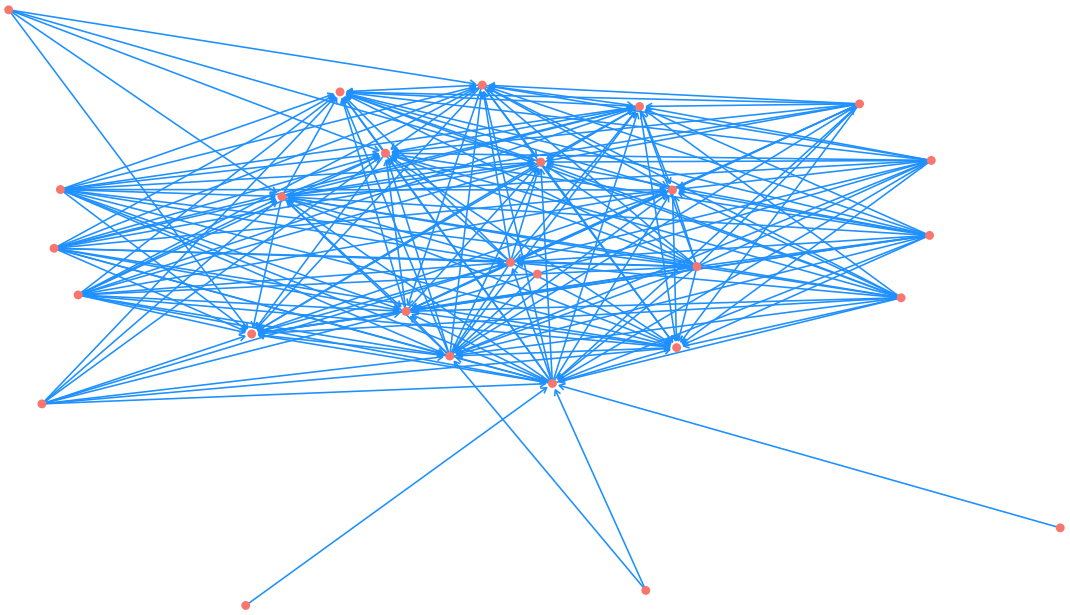


Figure 3.6: Essential proteins strongly connected component.

The essential cluster is weakly connected so it is reducible which by Definition 2 means that its adjacency matrix can take an upper triangular form. Ultimately this means that some proteins have only outgoing interactions and some only incoming interactions. So information in the essential cluster has direction. In Figure 3.7 we reconstructed the network to present the direction of interactions in 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 (Figure 3.7).

Because there is not a irreducible component (Figure 3.6) the analysis couldn't proceed further for the inner structure of the essential cluster (Table 3.5).

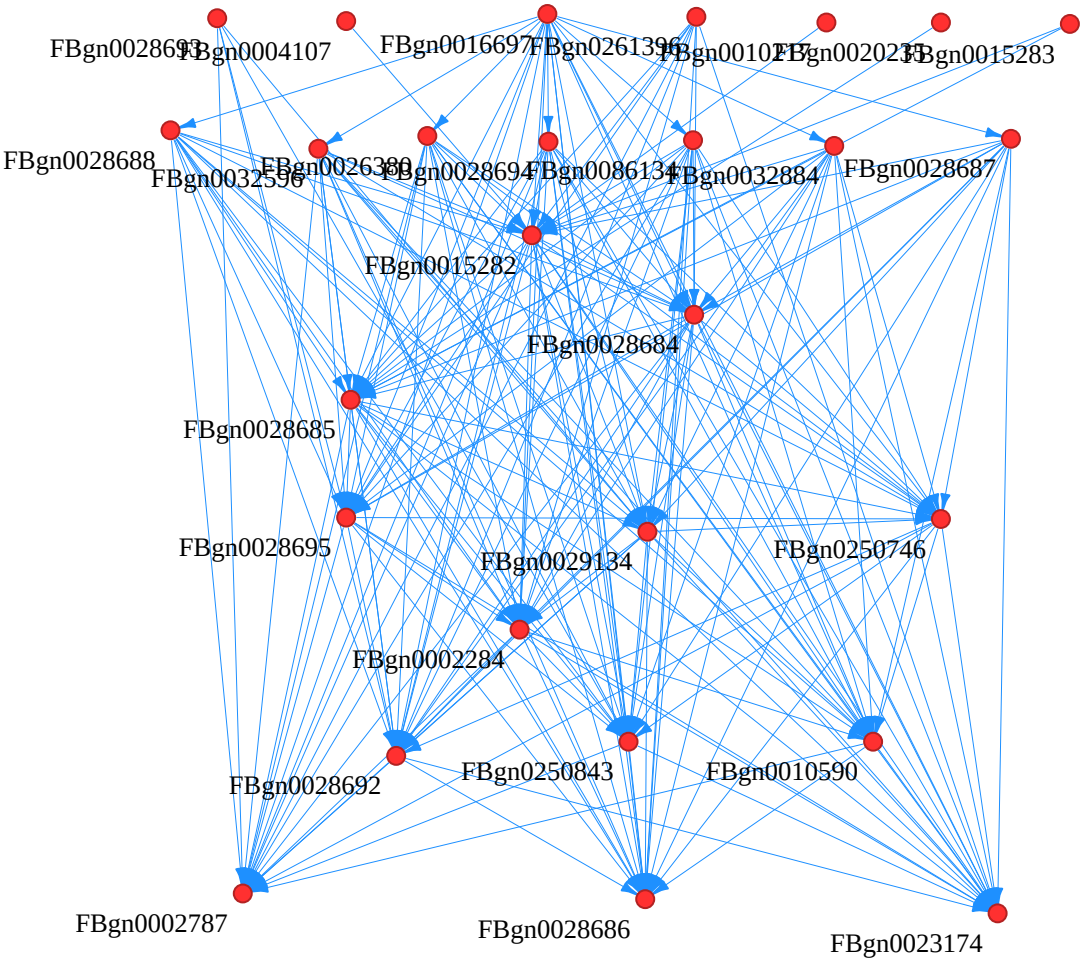


Figure 3.7: Graph decomposition using Frobenius decomposition theory. Top proteins act as input of signal flow, middle proteins process and traverse the signal to the final proteins, the output.

3.3.2 Gene ontology annotation

To examine the functions of proteins in the essential cluster we annotated them to Gene Ontology. First we found that 317 proteins, from the signed network, had no annotation in any ontology and 2378 proteins were annotated in all three ontologies (Table 3.6). 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 3.7).

Table 3.6: Gene ontology annotations of proteins of the signed network in the three ontologies

Ontology	Network Proteins	Number of ontologies	Number of proteins
Biological Process	2858	0	317
Molecular Function	2721	1	214
Cellular Component	2655	2	443
None	317	3	2378

Table 3.7: Biological process ontology Fisher's exact test significant terms

Method	Number of significant terms
Classic p-value	58
FDR	23
Bonferroni's correction	21

The subgraph of Biological Process Ontology with the significant terms is shown in Figure 3.8. After the enrichment of the essential protein cluster we found the following processes:

1. Protein catabolism
   (a) Proteasome subunits
   (b) Ubiquitin action
   (c) Response to stress
   (d) ATPases
2. ATP biosynthesis
3. Hydrogen membrane transport
4. Cell cycle G1/S transition

The essential cluster participates in these processes and operates with activating interactions. We also compared these results with the enrichment analysis of the essential proteins of the network as well as all the essential proteins of *D.melanogaster* (Table 3.8). All the GO enrichments contained the proteasome and protein catabolism. Also the essential cluster isn't enriched with translation and centrosome duplication, two processes that are highly enriched in the other essential protein sets (Figures 3.9 and 3.10).

Furthermore, proteins and their significant gene ontology terms represent a bipartite network. We projected this bipartite network to its terms to create the Functional network of significant terms (Figure 3.11). In

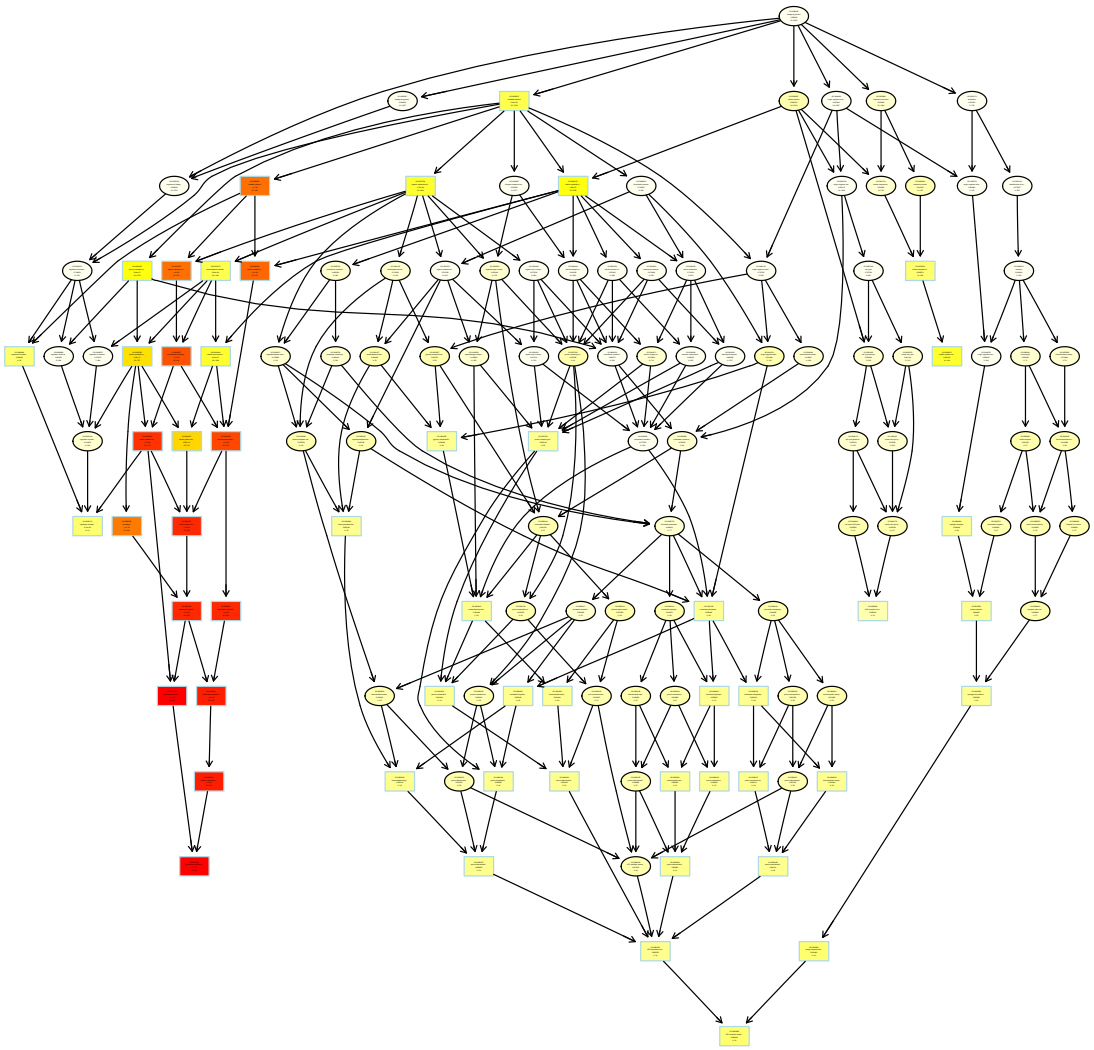


Figure 3.8: Singular enrichment analysis for the essential cluster genes in respect to network genes (universe). The left part of the graph is mostly for the proteasome and catabolism in general. The right part is mainly for response to stress and the middle part for nucleotide synthesis. Boxes indicate the 53 most significant terms that are connected through other terms represented with circles. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). The most significant terms are for the catabolism because red color is for very low p-values ($< 10^{-}4$). The arrows indicate is-a relationships. This graph was created using the *topGO* package from Bioconductor.

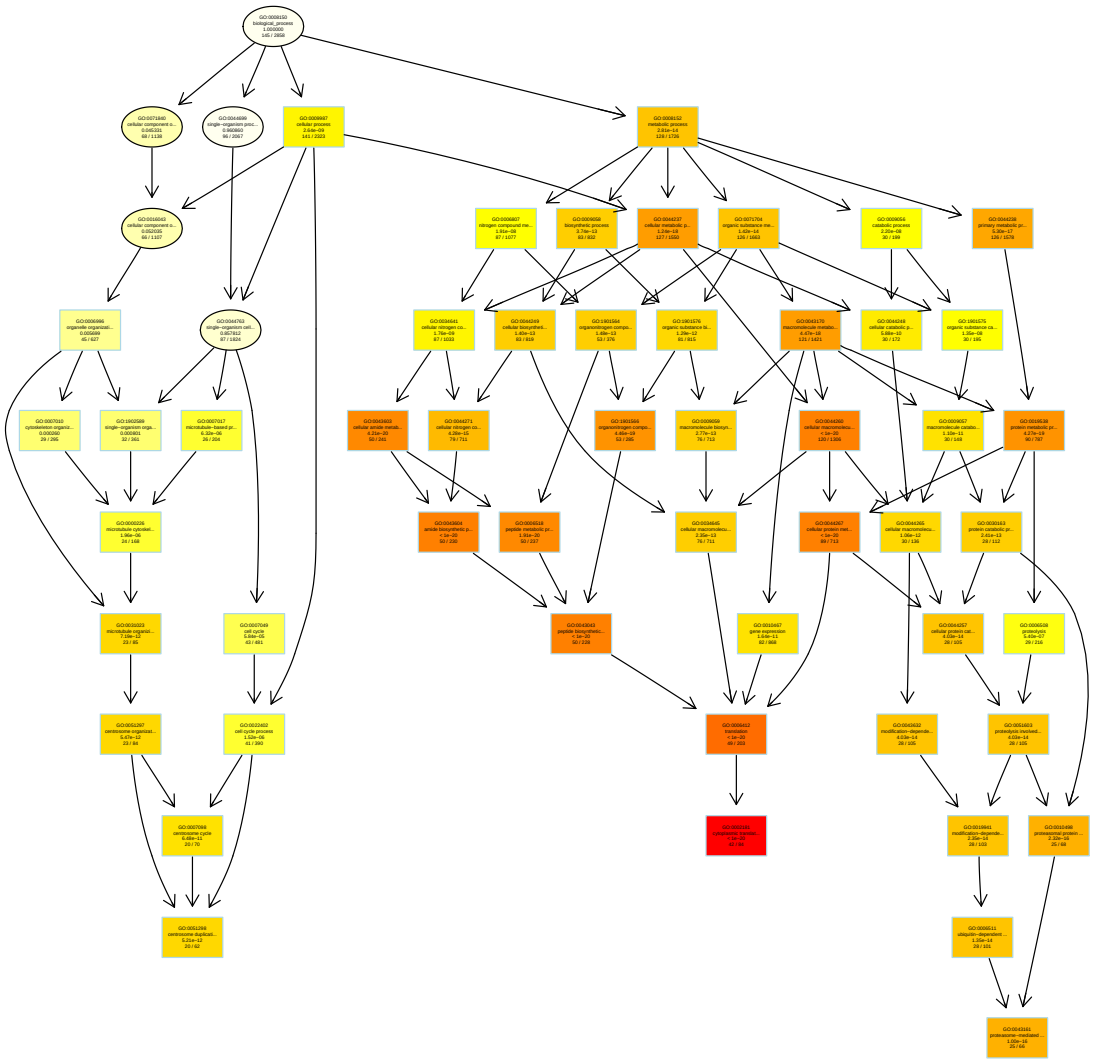


Figure 3.9: Singular enrichment analysis for the essential genes in the network in respect to all network genes (universe). Three main processes appear to be enriched, these are from the left, centrosome duplication as part of cell duplication, translation (peptide synthesis) and finally proteasome formation and protein catabolism. Boxes indicate the 51 most significant terms that are connected through other terms represented with circles. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). The arrows indicate is-a relationships. This graph was created using the *topGO* package from Bioconductor.

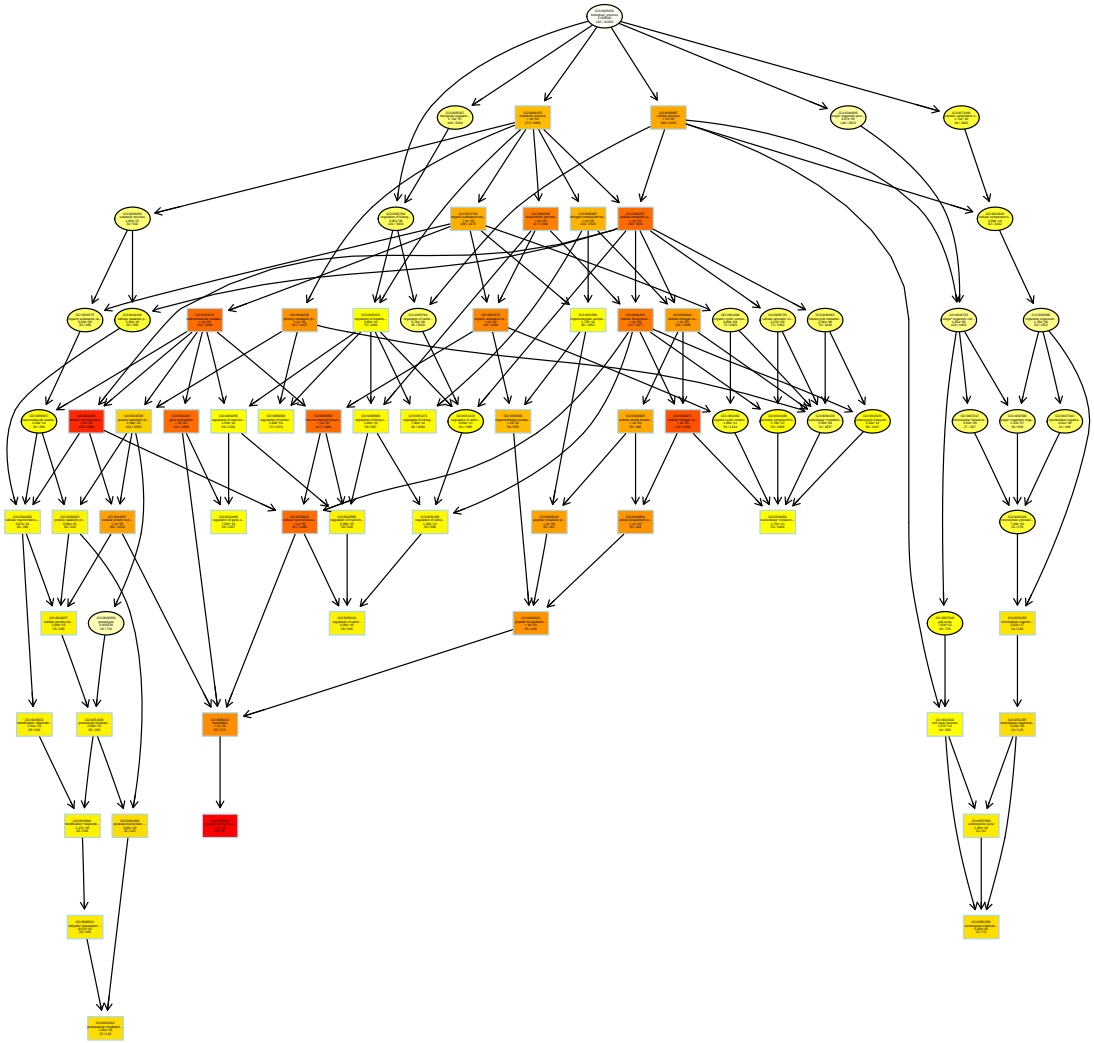


Figure 3.10: Singular enrichment analysis of all the essential genes of *D.melanogaster* from the OGEE database. The gene set we used as null distribution (universe) contained all genes of *D.melanogaster* that have been tested for their essentiality. The enriched part of the Biological Process ontology indicates that essential proteins in *D.melanogaster* are part of 5 processes which are interlinked. From left to right, is protein catabolism and the proteasome, next is the peptide synthesis through translation, followed by transcription regulation and nucleobase biosynthesis. At the far right part of the graph is the centrosome duplication process which is part of cell replication. Boxes indicate the 50 most significant terms (from 297 sig. terms in total) that are connected through other terms represented with circles. Box color represents the relative significance, ranging from dark red (most significant) to light yellow (least significant). The arrows indicate is-a relationships. This graph was created using the *topGO* package from Bioconductor.

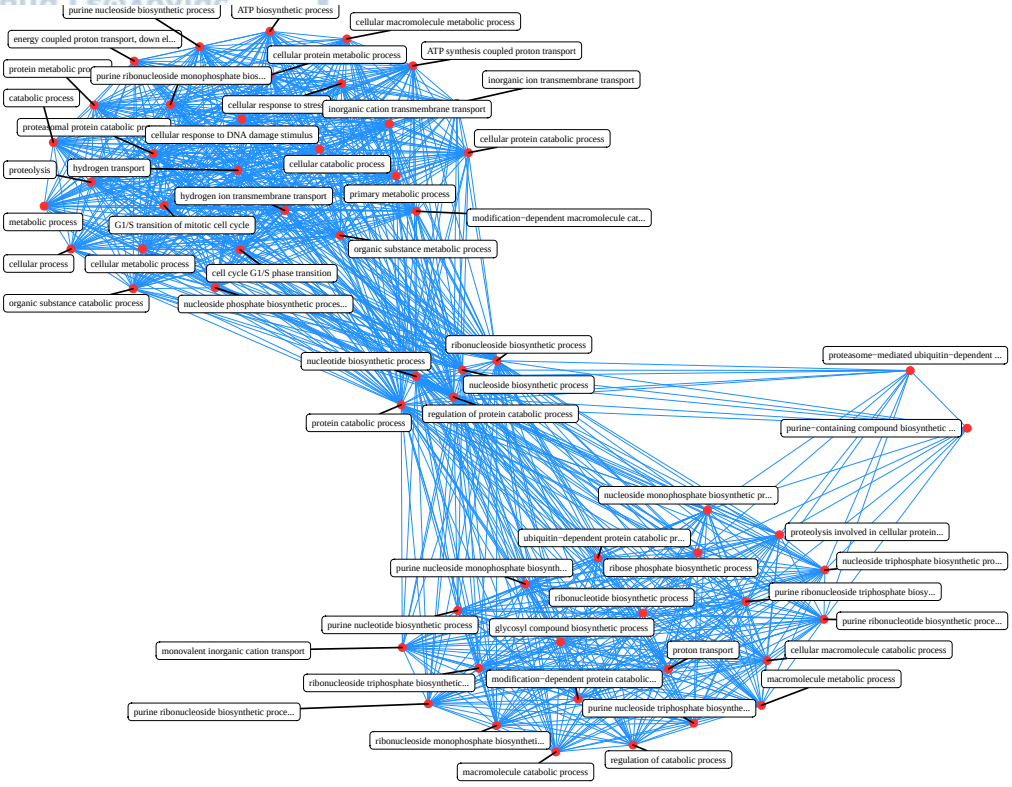


Figure 3.11: Functional enrichment analysis.

this network two terms interact with each other if they share a protein. The Functional network is dense and it shows how interconnected the catabolic process and nucleotide synthesis are in the essential cluster.

3.3.3 KEGG pathway annotation

We also used the KEGG database for pathway annotation of the essential cluster of *D.melanogaster* (Kanehisa and Goto 2000). The annotation had similar results with the GO annotation. In the essential cluster there are 4 pathways present as shown in Table 3.9. Again the proteasome has 24 protein in the

Table 3.8: Comparison of GO enrichment analysis of different essential protein sets.

Essential proteins	Proteins with Entrez IDs	Universe with Entrez IDs	Significant BP terms (<0.01)
Essential cluster	36	Signed network, 3256 proteins	53
Essential proteins in signed network	151	Signed network, 3256 proteins	51
All essential proteins	246	12812 proteins of *D.melanogaster*	297

cluster but there are other pathways involved as well which indicates a activating relationship between them.

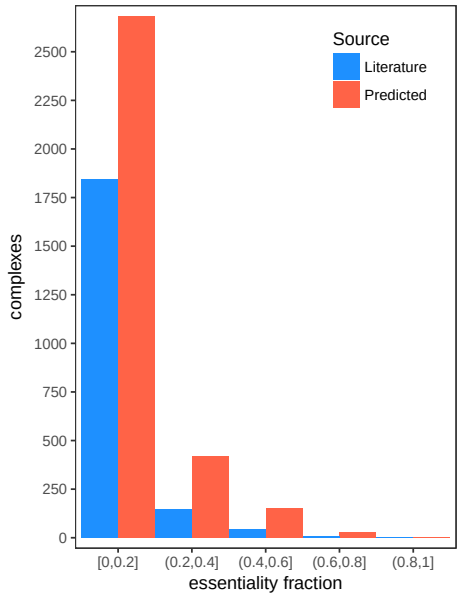
Table 3.9: KEGG pathway annothation of the protein of the essential cluster.

KEGG pathway ID	Number of proteins	KEGG pathway name
dme190	3	Oxidative phosphorylation
dme1100	3	Metabolic pathways
dme3010	3	Ribosome
dme3050	24	Proteasome
NA	6	NA

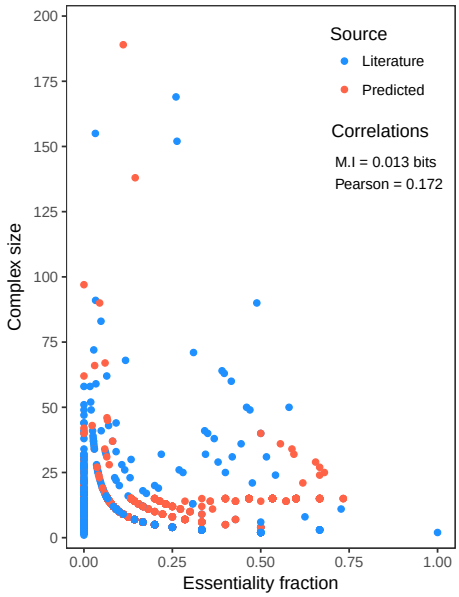
3.4 Modular essentiality

Proteins interact with each other temporally and conditionally to form complexes. These complexes are the functional machines that participate in biological processes. We tested the complexes of *D.melanogaster* to see how the essentiality consensus of their proteins is distributed. Some authors claim that a complex would either contain mostly essential proteins or not at all (Hart, Lee, and Marcotte 2007; Ryan et al. 2013). For all complexes from COMPLEAT database we calculated the essentiality fraction (Equation 14). Half of protein complexes have essentiality fraction in $[0, 0.2]$ bin (Figure 3.12a) which means that they are not essential. In order to avoid any bias due to complex size we plotted it against essentiality fraction and we didn't find any correlation (Figure 3.12b).

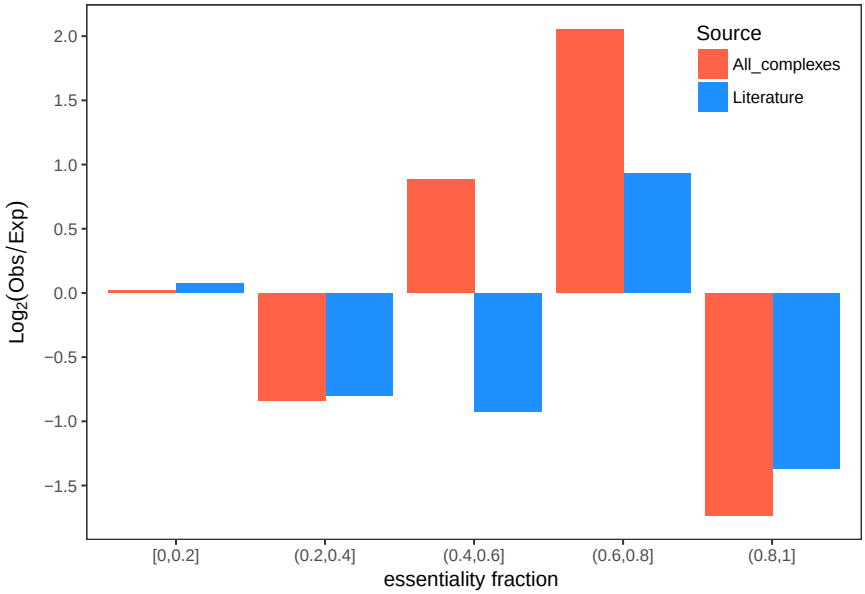
We bootstrapped the essential proteins of complexes to generate a random distribution of complexes for essentiality fraction. These are the expected values. We then binned the data to 5 equally spaced intervals. The log ratio (Equation 15) of observed to expected number complexes was calculated (Figure 3.12c) in each bin. In the $[0, 0.2]$ bin the log ratio is positive, although low, so the observed values are higher then the expected. This is also the case in the $(0.6, 0.8]$ bin. There is also a difference between all complexes and literature complexes in the $(0.4, 0.6]$ bin with all complexes having more abundance than expected bin which might be due to the bias of the predicted complexes from NetworkBlast (see Appendix A). In addition, there is only one complex with essentiality fraction in $(0.8, 1]$ in *D.melanogaster* which is lower than expected and contradicting to the modularity hypothesis (Table 3.10). In Figure 3.13 the bootstrapped distributions for each bin are shown. With vertical lines are the observed number of complexes. All the observed values are statistically substantial in one-tailed tests. These results are not so strong about the modularity of essentiality of complexes of *D.melanogaster* as for unicellular organisms.



(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 of the observed values to the expected.

Figure 3.12: Essentiality of the *D.melanogaster's* protein complexes. Almost half of the complexes have very low essentiality fraction. Also there appears to be modularity in the essentiality of complexes

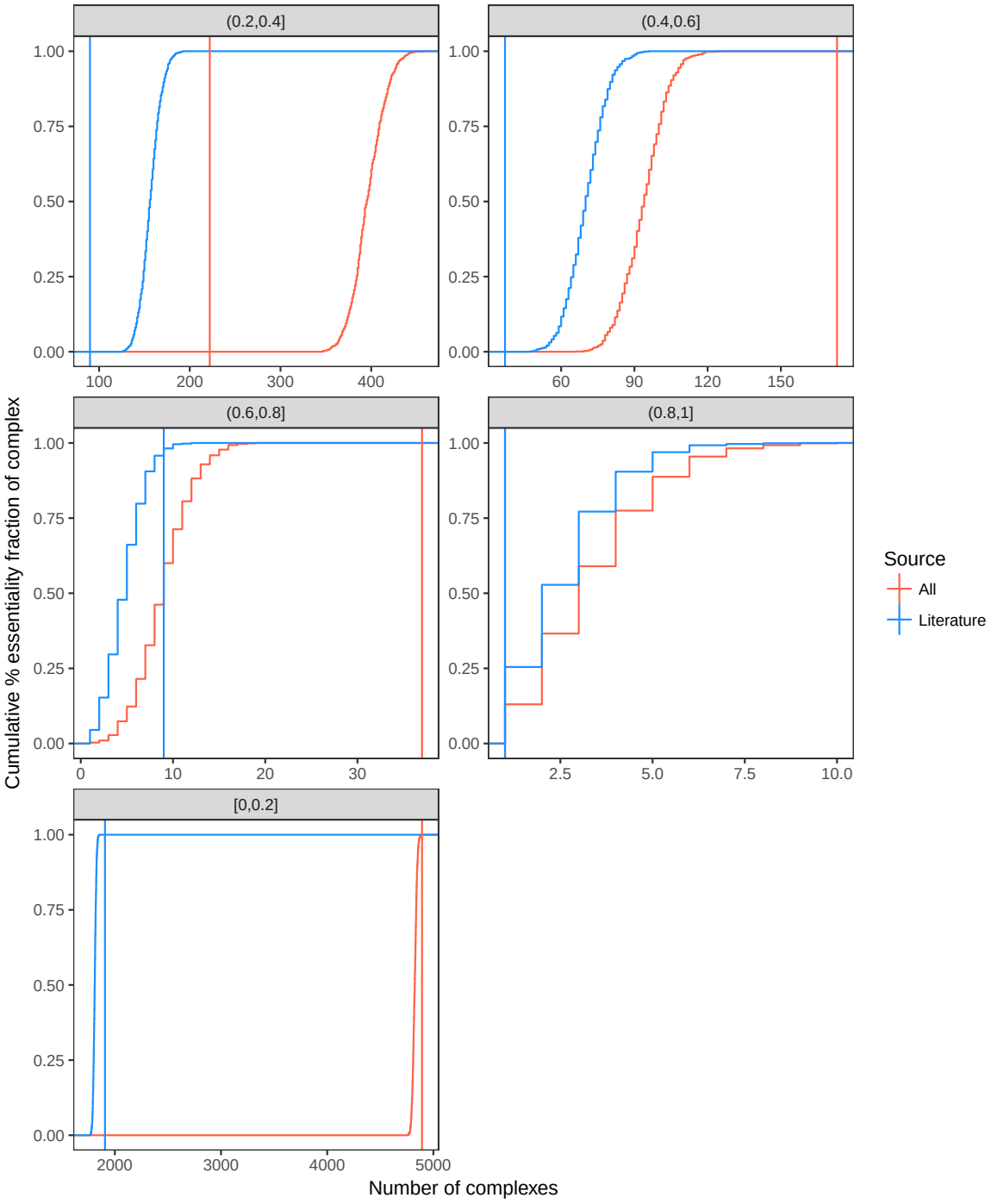


Figure 3.13: Cumulative distributions of bootstrapped essentiality of the complexes. The vertical lines are the observed number of complexes belonging to the respective bin.

Table 3.10: 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	4.729	0.928
	(0.8, 1]	1	2.584	-1.370

4 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 suggests 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*. Also, reexamination of the essentiality of *D.melanogaster's* proteins with screens other than RNAi may be informative.

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, especially degree centrality. 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. On the other hand, the more complex centralities like betweenness had lower prediction power than degree which may be due to missing interactions in the network.

In the literature, local interaction density centrality (LIDC) is considered the best centrality measure for essential protein prediction 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 3.3, 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 like the proteasome. 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. The creators of the signed network (Vinayagam et al. 2014) and (Lin et al. 2013) have also found that positive interactions are mostly found between proteins of the same complexes and negative interactions between proteins of different complexes. After the prediction of the interactions orientation using the naive Bayes method we found that the essential cluster is reducible, and we reconstructed the network with its components. This result indicates that information flow is directed which decomposes the essential cluster in 3 parts, input, processing and output. This general structure has been found in many signaling networks (Hyduke and Palsson 2010). Further analysis must be done to decipher the biological implications of this finding. In order to further study 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 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 out 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; 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 (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 14). Their tests were performed on unicellular organisms. We followed their methodology for *D.melanogaster's* complexes and found similar results (Figure 3.12). 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 3.13). 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[3], remains a huge challenge

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

(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) and predict this dynamic behavior with the use of temporal networks (Holme and Saramaki 2012). The temporal activation of protein interactions is due to spatial effects because proteins function in specific locations in the cell (Aebersold and Mann 2016) and because of the differential nature of interactions. It is a fact that different environmental conditions lead to radically different processes in organisms and consequently in 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 (Hartwell et al. 1999). This finding adds a new scale of interactions, interactions between complexes, 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). To conclude, as new data become avalaible that are more reliable and more diverse and this progress is accompaned with the advancement of mathematical and computational tools it will be more clear to decifer the essential components and interactions of organisms and processes.

List of Figures

List of Tables

A Appendix: COMPLEAT database

The COMPLEAT database (Vinayagam et al. 2013) provides both protein complex data and a platform for annotation and enrichment of RNAi and other data. To our knowledge it's the most complete database for protein complexes of *D.melanogaster*, *S.cerevisiae* and *H.sapiens* yet. While analyzing the complex data we discovered an irregularity in the complex size distribution. The authors didn't mention this irregularity which is apparent in Figure A.2a.There is a gap in the distribution between 15 and 16 number of proteins of complexes size for all organisms (Figure A.2a). This gap disappears in the reverse distribution which is the only distribution that was published by the authors, i.e the number of complexes that each protein participates in (Figure A.2b). This pattern looks like a phase transition which if it was true then it would have huge biological meaning. But with a more thorough look we discovered the source of this irregularity (Figures A.2d and A.2c).

Table A.1: Summary of COMPLEAT database for *D.melanogaster*

Source	Complexes	Proteins
Literature (326 distinct experiments)	2045	4501
NetworkBlast	2893	3525
CFinder	389	1362
Total	5327	5786

The complexes are provided by 3 different approaches, literature from specific experiments both small scale and high-throughput and computationally inferred from CFinder and NetworkBlast algorithms (Kalaev et al. 2008). In Table A.1 we see that half of the complexes are provided from the NetworkBlast algorithm. This tool has a plateau of 16 proteins as maximum complex size (Figure A.1) although the other methods show a heavy-tailed distribution to complexes size. This creates a bias towards medium sized complexes that is reflected to other analysis like the modular essentiality discussed here (Figure 3.12c). Further investigation is needed to determine if this bias of NetworkBlast is due to authors' implementation of NetworkBlast or the algorithm has an inherent bias towards medium sized protein complexes.

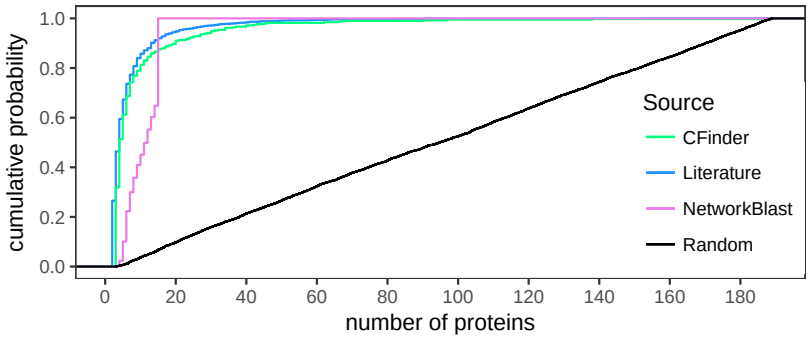
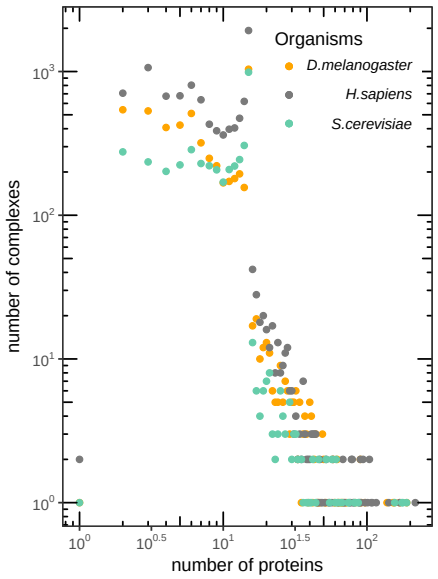
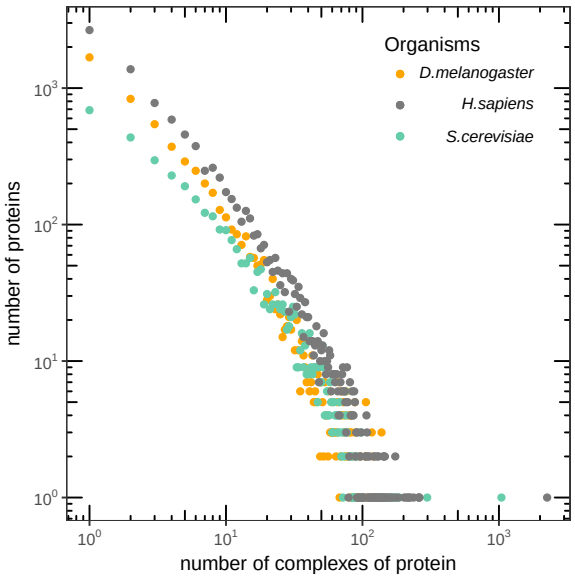


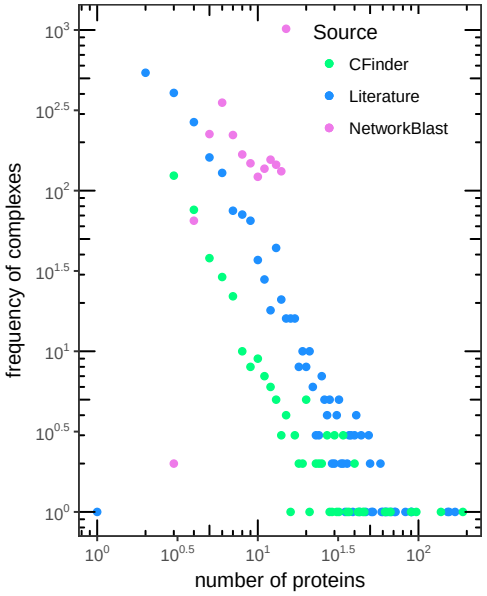
Figure A.1: Complex size cumulative distribution of *D.melanogaster* based on inference methods of COMPLEAT database. NetworkBlast reaches 100% in complex size of 16 proteins.



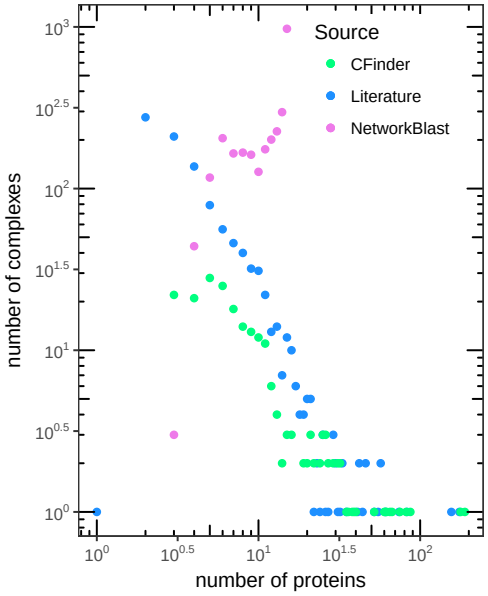
(a) Complexes size distribution.



(b) Proteins participation in complexes distribution.



(c) *D.melanogaster* complexes size distribution with different methods.



(d) *S.cerevisiae* complexes size distribution with different methods.

Figure A.2: COMPLEAT database distributions.

B Appendix: Network contraction with complexes

B.1 Complexes in the signed network

Which of these protein complexes are present in our data set? To answer this question we annotated the signed network proteins with complexes data. Most complexes have missing proteins in the interval [0.10] (Figure B.1) which is expected since most complexes are small (Figure A.2a). We found that 585 complexes were complete (Figure B.2).

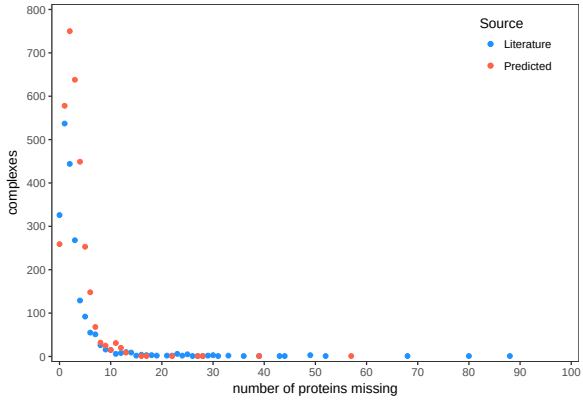


Figure B.1: Histogram of the missing proteins of complexes when comparered to the signed network.

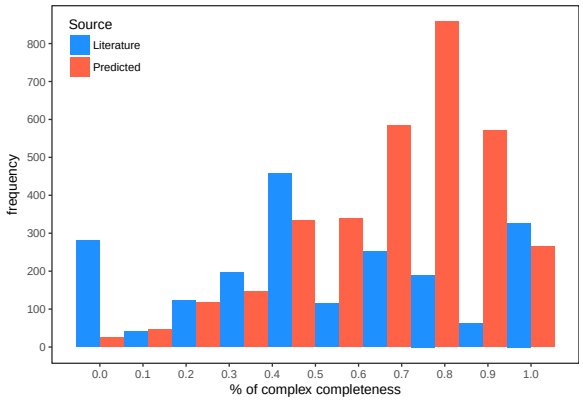


Figure B.2: Histogram of the percentage of proteins that appear in the signed PPI network per complex.

B.2 Network contraction with complexes

Since complexes are the molecules that facilitate most processes of the organisms it is very important to construct networks with complexes interactions. This requires experimental procedures and computational tools that can change the resolution to complexes scale. Scalability is one of the main goals for network science in the following years. To contract network with complexes from the protein - protein interaction

network it is necessary to determine which complexes to use. The rule we applied in this instance is to use only the complexes that all of their proteins are present in the network. This resulted in 585 complexes. Others can use a different threshold, like to use complexes that have >80% of their proteins present. Or take a completely different approach, like using GO annotation in the original network for the selection of complexes or applying clustering methods in protein networks like linked communities (Ahn, Bagrow, and Lehmann 2010; Kalinka and Tomancak 2011).

These 585 complexes contain 1063 proteins which have 2123 interactions in the signed network. So the 1/3 of the signed network is used. After we created the complexes network, two complexes are interacting if their proteins interact in the signed network. We got a network that contained duplicated edges and self loops which we deleted. There were multiple edges between complexes, we kept those that were distinct in the signed network.

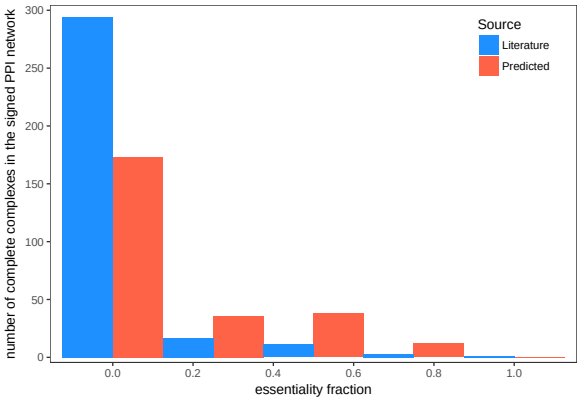


Figure B.3: Histogram of the essentiality fraction of the complexes that have all of their proteins in the signed PPI network. Forty nine protein complexes, from the 585 complexes that are complete in the signed network of drosophila, consist of 50% or more essential proteins.

In order to keep as much information as possible we treated positive and negative edges independently. More specifically, from all the redundant edges with the same direction, we kept 2, one positive and one negative. The weight of the positive edge and negative edge will be the normalized weight from all the positive and negative edges, respectively. Finally we normalized all the weights with the absolute value of the maximum weight, in order to have all the edge weights in the [-1,1]. This method resulted in a very dense network (Table B.1).

Table B.1: This is a summary of the network between complexes based on the signed PPI network.

Type	Total
Positive edges	14269
Negative edges	6081
Total	20350

References

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

Ahn, Yong-Yeol, James P Bagrow, and Sune Lehmann. 2010. "Link communities reveal multiscale complexity in networks." *Nature* 466 (7307). Nature Publishing Group: 761–64. doi:10.1038/nature09182.

Alexa, Adrian, and Jorg Rahnenfuhrer. 2016. *topGO: Enrichment Analysis for Gene Ontology*.

Allaire, J J, Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, Rob Hyndman, and Ruben Arslan. 2017. *rmarkdown: Dynamic Documents for R*. https://cran.r-project.org/package=rmarkdown.

Ashburner, Michael, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, et al. 2000. "Gene Ontology: tool for the unification of biology." *Nat. Genet.* 25 (1): 25–29. doi:10.1038/75556.

Barabási, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science (80-.).* 286 (5439): 509 LP–512. http://science.sciencemag.org/content/286/5439/509.abstract.

Boutros, Michael, Amy A Kiger, Susan Armknecht, Kim Kerr, Marc Hild, Britta Koch, Stefan A Haas, Heidelberg Fly Array Consortium, Renato Paro, and Norbert Perrimon. 2004. "Genome-Wide RNAi Analysis of Growth and Viability in Drosophila Cells." *Science (80-.).* 303 (5659): 832–35. http://science.sciencemag.org/content/303/5659/832.abstract.

Brohée, Sylvain, and Jacques van Helden. 2006. "Evaluation of clustering algorithms for protein-protein interaction networks." *BMC Bioinformatics* 7: 488. doi:10.1186/1471-2105-7-488.

Carey, Vince, Li Long, and R Gentleman. 2016. *RBGL: An interface to the BOOST graph library*. http://www.bioconductor.org.

Carlson, Marc. 2016a. *GO.db: A set of annotation maps describing the entire Gene Ontology*.

———. 2016b. *org.Dm.eg.db Genome wide annotation for Fly*.

Chatr-Aryamontri, Andrew, Bobby Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, et al. 2015. "The BioGRID interaction database: 2015 update." *Nucleic Acids Res.* 43 (D1): D470–D478. doi:10.1093/nar/gku1204.

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.

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.

Croft, David, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, et al. 2014. "The Reactome pathway knowledgebase." *Nucleic Acids Res.* 42 (D1): D472. doi:10.1093/nar/gkt1102.

Csardi, Gabor, and Tamas Nepusz. 2006. "The igraph software package for complex network research." *InterJournal* Complex Sy: 1695. http://igraph.org.

D'Elia, Michael A., Mark P. Pereira, and Eric D. Brown. 2009. "Are essential genes really essential?" *Trends Microbiol.* 17 (10): 433–38. doi:10.1016/j.tim.2009.08.005.

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.

Fazekas, Dávid, Mihály Koltai, Dénes Türei, Dezső Módos, Máté Pálfy, Zoltán Dúl, Lilian Zsákai, et al. 2013. "SignaLink 2 – a signaling pathway resource with multi-layered regulatory networks." *BMC Syst. Biol.* 7 (1): 7. doi:10.1186/1752-0509-7-7.

Fraser, Claire M, Jeannine D Gocayne, Owen White, Mark D Adams, Rebecca A Clayton, Robert D Fleischmann, Carol J Bult, et al. 1995. "The Minimal Gene Complement of Mycoplasma genitalium." *Science (80-.).* 270 (5235): 397 LP–404. http://science.sciencemag.org/content/270/5235/397.abstract.

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

Freeman, Linton C., Stephen P. Borgatti, and Douglas R. White. 1991. "Centrality in valued graphs: A measure of betweenness based on network flow." *Soc. Networks* 13 (2): 141–54. doi:10.1016/0378-8733(91)90017-N.

Gantmacher, F.R. 1987. *The Theory of Matrices vol 2*. 2nd ed. AMS Chelsea Publishing. doi:10.1007/978-3-642-99234-6.

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.

Gitter, Anthony, Judith Klein-Seetharaman, Anupam Gupta, and Ziv Bar-Joseph. 2011. "Discovering pathways by orienting edges in protein interaction networks." *Nucleic Acids Res.* 39 (4). doi:10.1093/nar/gkq1207.

Glass, Kimberly, and Michelle Girvan. 2014. "Annotation enrichment analysis: an alternative method for evaluating the functional properties of gene sets." *Sci. Rep.* 4: 4191. doi:10.1038/srep04191.

Gluecksohn-Waelsch, Salome. 1963. "Lethal Genes and Analysis of Differentiation." *Science (80-.).* 142: 1269–76. doi:10.1126/science.142.3597.1269.

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.

Holman, Alexander G, Paul J Davis, Jeremy M Foster, Clotilde KS Carlow, and Sanjay Kumar. 2009. "Computational prediction of essential genes in an unculturable endosymbiotic bacterium, Wolbachia of Brugia malayi." *BMC Microbiol.* 9 (1): 243. doi:10.1186/1471-2180-9-243.

Holme, Petter, and Jari Saramaki. 2012. "Temporal networks." *Phys. Rep.* 519 (3). Elsevier B.V.: 97–125. doi:10.1016/j.physrep.2012.03.001.

Hornik, Kurt, Christian Buchta, and Achim Zeileis. 2009. "Open-Source Machine Learning: {R} Meets {Weka}." *Comput. Stat.* 24 (2): 225–32. doi:10.1007/s00180-008-0119-7.

Hutchison, C. A., R.-Y. Chuang, V. N. Noskov, N. Assad-Garcia, T. J. Deerinck, M. H. Ellisman, J. Gill, et al. 2016. "Design and synthesis of a minimal bacterial genome." *Science (80-.).* 351 (6280): 6253–3. doi:10.1126/science.aad6253.

Hyduke, Daniel R., and Bernhard Ø. Palsson. 2010. "Towards genome-scale signalling-network reconstructions." *Nat. Rev. Genet.* 11 (4). Nature Publishing Group: 297–307. doi:10.1038/nrg2750.

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.

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.

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.

Jordan, I King, Igor B Rogozin, Yuri I Wolf, and Eugene V Koonin. 2002. "Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria." *Genome Res.* 12: 962–68. doi:10.1101/gr.87702.

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.

Kabacoff, Robert I. 2011. *R in Action : Data analysis and graphics with R.*

Kalaev, Maxim, Mike Smoot, Trey Ideker, and Roded Sharan. 2008. "NetworkBLAST: Comparative analysis of protein networks." *Bioinformatics* 24 (4): 594–96. doi:10.1093/bioinformatics/btm630.

Kalinka, Alex T., and Pavel Tomancak. 2011. "linkcomm: An R package for the generation, visualization,

and analysis of link communities in networks of arbitrary size and type." *Bioinformatics* 27 (14): 2011–2. doi:10.1093/bioinformatics/btr311.

Kanehisa, Minoru, and Susumu Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Res.* 28 (1). Oxford, UK: Oxford University Press: 27–30. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/.

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

Koonin, E V. 2000. "HOW MANY GENES CAN MAKE A CELL: The Minimal-Gene-Set Concept." *Annu. Rev. Genomics Hum. Genet.* 01: 99–116.

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.

Korcsmaros, Tamas, Mate S. Szalay, Petra Rovo, Robin Palotai, David Fazekas, Katalin Lenti, Ill??s J. Farkas, Peter Csermely, and Tibor Vellai. 2011. "Signalogs: Orthology-based identification of novel signaling pathway components in three metazoans." *PLoS One* 6 (5). doi:10.1371/journal.pone.0019240.

Kotsiantis, S. B. 2013. "Decision trees: A recent overview." *Artif. Intell. Rev.* 39 (4): 261–83. doi:10.1007/s10462-011-9272-4.

Krogan, Nevan J, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, et al. 2006. "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae." *Nature* 440 (7084): 637–43. doi:10.1038/nature04670.

Kuhn, Max, Steve Weston, Nathan Coulter, and Mark Culp. C code for C5.0 by R. Quinlan. 2015. *C50: C5.0 Decision Trees and Rule-Based Models*. https://cran.r-project.org/package=C50.

Lieben, Liesbet. 2015. "Redefining gene essentiality." *Nat. Publ. Gr.*, no. December. Nature Publishing Group: 2015. doi:10.1038/nrg.2015.23.

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.

Liu, Gaowen, Mei Yun Jacy Yong, Marina Yurieva, Kandhadayar Gopalan Srinivasan, Jaron Liu, John Soon Yew Lim, Michael Poidinger, et al. 2015. "Gene Essentiality Is a Quantitative Property Linked to Cellular Evolvability." *Cell* 163 (6). Elsevier Inc.: 1388–99. doi:10.1016/j.cell.2015.10.069.

Lu, Long Jason (editor). 2015. *Gene Essentiality Methods and Protocols*. Edited by Long Jason Lu. New York: Springer Science+Business Media. doi:10.1007/978-1-4939-2398-4.

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.

Luo, Weijun, Brouwer, and Cory. 2013. "Pathview: an R/Bioconductor package for pathway-based data

integration and visualization." *Bioinformatics* 29 (14): 1830–1. doi:10.1093/bioinformatics/btt285.

Manfred Eigen. 1971. "Self organization of matter and the evolution of biological macromolecules." *Naturwissenschaften* 58: 465–523. doi:10.1007/BF00623322.

Manning, Christopher D., Raghavan Prabhakar, and Hinrich Schutze. 2008. *Introduction to Information Retrieval*. 1st ed. Vol. 1. Cambridge: Cambridge University Press 2008. doi:10.1017/CBO9781107415324.004.

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.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2017. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. https://cran.r-project.org/package=e1071.

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.

Murali, Thilakam, Svetlana Pacifico, Jingkai Yu, Stephen Guest, George G. Roberts, and Russell L. Finley. 2011. "DroID 2011: A comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila." *Nucleic Acids Res.* 39 (SUPPL. 1): 736–43. doi:10.1093/nar/gkq1092.

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.

Newman, M E J. 2006. "Modularity and community structure in networks." *Proc. Natl. Acad. Sci. U. S. A.* 103 (23): 8577–82. doi:10.1073/pnas.0601602103.

Newman, M. E J. 2005. "A measure of betweenness centrality based on random walks." *Soc. Networks* 27 (1): 39–54. doi:10.1016/j.socnet.2004.11.009.

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.

Pagès, Hervé, Marc Carlson, Seth Falcon, and Nianhua Li. 2017. *AnnotationDbi: Annotation Database Interface*.

Pedersen, Thomas Lin. 2017. *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. https://cran.r-project.org/package=ggraph.

Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science (80-.).* 334: 1226–7. doi:10.1126/science.1213847.

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.

Piccolo, Stephen R., and Michael B. Frampton. 2016. "Tools and techniques for computational reproducibility." *Gigascience* 5 (1). GigaScience: 30. doi:10.1186/s13742-016-0135-4.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Mach. Learn.* 1 (1): 81–106. doi:10.1023/A:1022643204877.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Rhee, Seung Yon, Valerie Wood, Kara Dolinski, and Sorin Draghici. 2008. "Use and misuse of the gene ontology annotations." *Nat. Rev. Genet.* 9 (7): 509–15. doi:10.1038/nrg2363.

Rivals, Isabelle, Léon Personnaz, Lieng Taing, and Marie Claude Potier. 2007. "Enrichment or depletion of a GO category within a class of genes: Which test?" *Bioinformatics* 23 (4): 401–7. doi:10.1093/bioinformatics/btl633.

RStudio Team. 2016. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc. http://www.rstudio.com/.

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.

Shaw, Marvin E. 1954. "Group Structure and the Behavior of Individuals in Small Groups." *J. Psychol.* 38 (1): 139–49. doi:10.1080/00223980.1954.9712925.

Siek, JG, LQ Lee, and Andrew Lumsdaine. 2001. *The Boost Graph Library: User Guide and Reference Manual*. Boston, MA: Pearson Education.

Sing, T, O Sander, N Beerenwinkel, and T Lengauer. 2005. "ROCR: visualizing classifier performance in R." *Bioinformatics* 21 (20): 7881. http://rocr.bioinf.mpi-sb.mpg.de.

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.

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.

Tatum, E. L., and J. Lederberg. 1947. "Gene Recombination in the Bacterium Escherichia coli." *J. Bacteriol.* 53 (6): 673–84. doi:10.1038/158558a0.

Tenenbaum, Dan. 2017. *KEGGREST: Client-side REST access to KEGG*.

Therneau, Terry, Beth Atkinson, and Brian Ripley. 2017. *rpart: Recursive Partitioning and Regression Trees*. https://cran.r-project.org/package=rpart.

Varga, Richard S. 2000. *Matrix Iterative Analysis*. Edited by H. Yserentant, R. Bank, R.L. Graham, J.

Stoer, and R. Varga. 2nd editio. Heidelberg: Springer-Verlag. doi:10.1007/978-3-642-05156-2.

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, Ulrich Stelzl, Raphaele Foulle, Stephanie Plassmann, Martina Zenkner, Jan Timm, Heike E Assmus, Miguel A Andrade-Navarro, and Erich E Wanker. 2011. "A directed protein interaction network for investigating intracellular signal transduction." *Sci. Signal.* 4 (189): rs8. doi:10.1126/scisignal.2001699.

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.

Walhout, Albertha J. M., Raffaella Sordella, Xiaowei Lu, James L. Hartle, Gary F. Temple, Michael A. Brasch, Nicolas Thierry-Mieg, and Marc Vidal. 2000. "Protein Interaction Mapping in C.elegans Using Proteins Involved in Vulval Development." *Science (80-.).* 287 (5450): 116–22. doi:10.1126/science.287.5450.116.

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

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

———. 2017. "tidyr Easily tidy data with spread() and gather() functions." http://tidyr.tidyverse.org.

Wickham, Hadley, and Romain Francois. 2016. *dplyr: A Grammar of Data Manipulation*. https://cran.r-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Romain Francois. 2017. *readr: Read Rectangular Text Data*. https://cran.r-project.org/package=readr.

Winzeler, E. A. 1999. "Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis." *Science (80-.).* 285 (5429): 901–6. doi:10.1126/science.285.5429.901.

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.

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.

Zhan, Tianzuo, and Michael Boutros. 2016. "Towards a compendium of essential genes – From model organisms to synthetic lethality in cancer cells." *Crit. Rev. Biochem. Mol. Biol.* 51 (2): 74–85. doi:10.3109/10409238.2015.1117053.

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.

Zhang, Zhaojie, and Qun Ren. 2015. "Why are essential genes essential? - The essentiality of Saccharomyces genes." *Microb. Cell* 2 (8): 280–87. doi:10.15698/mic2015.08.218.

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.