Functional characterization of transcriptional regulatory networks of yeast species

PAULO DIAS, Instituto Superior Técnico, Portugal

Transcriptional regulatory networks are responsible for controlling gene expression. These networks are composed of many interactions between transcription factors and their target genes. Carrying a combinatorial nature that encompasses several regulatory processes, they allow an organism to respond to disturbances that may occur in the surrounding environment. In this thesis, we explore different possibilities for the study of transcriptional regulatory networks. The intention is to reveal which functions and/or processes are encoded in the regulatory patterns that constitute the transcriptional regulatory networks. To accomplish that, we study a set of regulatory networks from closely related yeast species using different methods, dividing the workflow into two phases. The first phase consists of a detection of modules followed by their functional characterization. With this, we showed that the regulatory networks can be divided into functional modules that represent the biologic functions of the respective species. In the second phase, we move towards a cross-species analysis. Here, we compare the functional elements of the different species and we study the similarities among them. The purpose of this analysis is to discover if there are any functional elements conserved across the distinct organisms. Overall, our thesis provides a novel pipeline to analyze how the structure and function of regulatory networks of different species may relate to each other. In addition, we explore how those similarities between species can help to infer some properties in networks.

Additional Key Words and Phrases: Complex Networks, Transcriptional Regulatory Networks, Multilayer Networks, Community Detection, Functional Modules

1 INTRODUCTION

Gene expression is the biological process that allows a cell to respond to its changing environment. Each cell is the product of specific gene expression events involving the transcription of thousands of genes. The transcription factors (TFs) are the core elements in the control of gene expression. These are responsible for the activation or inhibition of the genes under their regulation, the target genes (TGs). Normally, the expression level of a target gene is the result of the combinatorial regulation of multiple transcription factors. The hundreds of interactions between transcription factors and target genes define a transcriptional regulatory network that underlies cellular identity and function.

The morphological differences between species/organisms arise from the differential regulation of genes. The information for this differential regulation is encoded in the genomic regulatory code of each individual and it is the product of evolution [Davidson 2001]. The transcriptional regulatory networks of organisms are assembled during their evolution and combine new regulatory links and preexisting ones. These networks are of great biological importance and their study is sure to bring new developments to the scientific community. The understanding of differential gene expression is facilitated by the analysis of transcriptional regulatory networks [Davidson et al. 2002; Luscombe et al. 2004]. Therefore, insights from the structure and evolution of these networks can be translated into predictions and used for the analysis of the regulatory networks of different organisms.

Despite their central role in biology, the structure and dynamics of transcriptional regulatory networks are largely undefined. In this thesis, we study transcriptional regulatory networks, represented as graphs. Graphs are the simplest way to represent complex systems. These can be used to portray the most varied complex systems, many of these systems are present in our daily life. Such as our social networks formed by family and friends or the transport network we use to move around. Also, our regulatory interactions between thousands of genes and transcription factors can be represented as graphs.

In this thesis, our goal is to broaden our knowledge about transcriptional regulatory networks. To achieve our goal, we analyze the structure of these networks by inspecting their division into modules and their functional characterization. Furthermore, our analysis continues with a cross-species comparison. Here, compare the information gathered in the functional characterization of the species to identify conserved functional elements across species.

The exploration of the community structure is the most common way to study the structure of a network. In Network Science, a community (or module) is defined as a group of nodes that have a higher likelihood of connecting to each other than to nodes of other communities. In biological networks, communities can express biological functions. For example, the identification of communities can be used to discover new structures associated with specific biological functions previously unknown [Lewis et al. 2010; Voevodski et al. 2009]. The field of Network Science has offered several community detection methods based on different approaches targeting different types of problems [Fortunato 2010].

The second phase of our analysis consists of a cross-species comparison. Cross-species studies have proven to be crucial in modern biology. They are important to study the differences and similarities between species, which is fundamental to understanding their evolution. For example, it allows the identification of conserved structures between different organisms [Borneman et al. 2007; Matthews et al. 2001]. Related to cross-species studies, are the studies between different types of data, also important in the biologic field. For example, in medical research, the study of different types of cancer tissues allows the prediction of candidate driving genes in cancer [Cantini et al. 2015]. The use of a multilayer network can be very useful in this cases since it allows the representation and comparison of different types of information, which is not possible with the simple and common graph representation.

1.1 Objective

In this work, we make a characterization of transcriptional regulatory networks of several closely related species with the generic goal of gaining insight into this kind of network. In particular, we consider the readily available dataset from YEASTRACT+ [Monteiro et al. 2019] which provides a set of closely related yeast species with annotated data, both in terms of functional annotation and in terms of mapping between nodes of different species. To accomplish this, we use some Network Science methods. We outline our approach by dividing it into two phases: (1) detection and functional characterization of communities/modules; (2) cross-species comparison. With our approach, we aim to analyze the interplay between structure and function for each species and also between species.

The first step of the characterization involves the examination of the community structure of the networks. In transcriptional regulatory networks, we assume that a community is associated with one or more biologic functions. We begin by performing detection of communities on the networks. Then, we proceed with the functional characterization of the communities found. The motivation is to verify if the networks can be divided into well-defined functional communities that reflect the different functions present in the regulatory code of the species. In this step, we apply several community detection techniques to understand which technique is the most suitable for our problem. Within the set of algorithms to be applied, some are suitable for the study of overlapping and signed communities. Regarding the study of overlapping communities, the transcription factors may be associated with multiple regulatory processes. Therefore, the study of this type of communities can be useful as it allows genes to belong to different functional groups. Transcriptional regulatory networks can be seen as signed networks, the relationship between transcription factors and target genes may be negative (denoting inhibition) or positive (denoting activation). Thus, the study of polarized communities may be useful in the identification of regulatory processes in our species.

Moving to the second stage, we focus on the cross-species comparison to identify the main similarities between species. Comparing the functional modules discovered in the first stage, we intend to find out if there are strongly connected modules between different species, which can reveal the functional similarity between organisms or help us to infer functional elements in other species. To finalize our characterization, we use a multilayer network approach combining networks of different species. Then, we proceed with a final community detection step. The goal is to detect modules that may reunite genes from different species that may encode important regulatory patterns conserved across species.

1.2 Outline and Contributions

This document is organized as follows. In Section 2, we start by presenting some concepts that are fundamental to understanding our approach. Then, Section 3 is focused on the state-of-the-art underlying our work. Moving to Section 4, we begin the study of the transcriptional regulatory networks. The second phase of the analysis is described in Section 5. Finally, in Section 6, we draw the concluding remarks, also commenting on the limitations of our approach and possible future analyses related to our results.

During the development of this thesis, we contributed with two accepted talks at two main conferences in the Network Science Community: Networks 2021 and CompleNet 2021.

2 CONCEPTS

In this section, we present some concepts essential for the reading of the document. First, we introduce some Network Science concepts. Then, we review two biological concepts, the transcriptional regulatory networks (the focus of our work) and the Gene Ontology (GO), a useful resource to functionally characterize our species.

Graph. A *graph*, G, is represented by the tuple (V, E) where V is the set of vertices/nodes and $E \subseteq V \times V$ is a set of edges/links that connect the nodes. The size of V is denoted by $N = |V|$ and is the size of the graph, the size of E is denoted by $L = |E|$. A *node/vertex* represents an entity in a graph. This entity and can be a person in a social network, a company in a financial market, a station in a transport network, or a gene in a transcriptional regulatory network. An *edge*, or *link*, represents a relation between two nodes. This interaction can represent a friendship between two people in a social network, a connection between two companies that do business together in a financial market, a connection between two stations in a transport network, or a connection between a transcription factor and a target gene in a transcriptional regulatory network. Two nodes i and j are *adjacent* or *neighbors* if there is an edge e connecting them, i.e., $e = (i, j) \in E$.

Community. In Network Science a *community* or a *module* is defined as a group of nodes that have a higher likelihood of connecting to each other than to nodes of other communities. Thus, communities are locally dense connected subgraphs in a network. Communities play a particularly important role in some areas. They allow us to obtain important information about the functional components of a system and the impact of local structures on dynamics at a global scale. *Modularity* [Newman and Girvan 2004] is the measure that allows us to quantify the quality of a partition c in a graph G.

Multilayer Network. The basic representation by graphs is the most common and simple way to portray complex systems. However, with the evolution of research in complex systems, it became necessary to study systems that are increasingly complex but closer to reality. It has become essential to go beyond the simple representation by graphs. This lead to the emergence of a new approach, the representation of systems by a multilayer network [Boccaletti et al. 2014; Kivelä et al. 2014]. In this scenario, we consider layers in addition to nodes and edges, Figure 1.

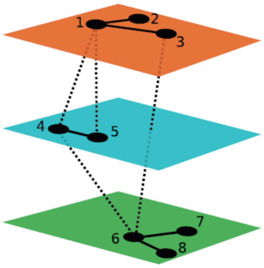


Fig. 1. Multilayer network with three layers. The intra-layer edges are represented by solid lines and the inter-layer edges by the dotted lines. Figure obtained from [Kivelä et al. 2014].

In a multilayer network, each layer is associated with an *aspect*. This aspect can be the type of links or an instant of time. This way, it is possible to build a network in which all edges of different types are embedded in different layers of interaction. There are three types of edges in multilayer networks:

- **Intra-layer edges** - edges connecting two nodes in the same layer
- **Inter-layer edges** - edges connecting two nodes in different layers
- **Couplings** - edges connecting two copies of the same node in different layers

In our approach, we build a multilayer network in which the layers represent different species. The inter-layer edges are established between homologous genes of different species. There are no couplings in this case.

Transcriptional Regulatory Networks. The transcriptional regulatory networks are responsible for the gene regulation that controls genomic expression. This process allows a cell or an organism to respond and adapt to a variety of stimuli from the environment like unexpected and stressful situations. A gene can be:

- **Transcription factor (TF)** - gene that have a regulatory role
- **Target gene (TG)** - gene regulated by the transcription factors

The regulatory association between two genes can represent the activation or inhibition of the expression of the target gene by the transcription factor.

A transcriptional regulatory network can be represented as a directed graph and a variation of a signed graph $G = (V, E, w)$. The nodes of V are transcription factors or target genes (or both), and the edges of E are the directed connections between transcription factors and target genes. The edges denote activation or repression effects on transcription, so, we define edge labels $w \in \{-1, 1, 0\}$ between two nodes (i, j) as follows: $w_{i,j} = 1$ when the transcription factor i is an activator of the target gene j and $w_{i,j} = -1$ when the transcription factor i is a repressor of the target gene j. When we do not know if it is an activator or repressor, $w_{i,j} = 0$. The out-degree of a gene is the number of target genes that it regulates and the in-degree of each gene is the number of transcription factors controlling its transcription. A gene can be a transcription factor and a target gene at the same time. If a gene acts only as a transcription factor, its in-degree is 0, if it acts only as a target gene, its out-degree is also 0.

Gene Ontology (GO). The Gene Ontology[1] [Ashburner et al. 2000] is the most comprehensive resource about the functions of genes and is also the one that is mostly used to support modern biological research. It subdivides the functionality of a gene into three distinct ontologies:

- **Molecular function** - the activity of the gene at the molecular level
- **Cellular component** - the location of the activity of the gene in relation to the biological structures
- **Biological process** - biological process that contains the molecular function of the gene

Each of these three ontologies is a hierarchy of terms, each term has a definition that allows us to define the relationships the terms have with each other. A hierarchy is composed of several levels, apart from the terms that are leaves, all terms can have children, and these children represent a more specific term/process than the parent. The Gene Ontology allows the characterization of a gene in three distinct aspects, each corresponding to one of the ontologies. Having a set of genes, we can use the relationships between the terms to identify the main processes associated with that set. Therefore, this is a useful resource in biology, since it allows the functional characterization of organisms.

3 RELATED WORK

In this section, we present the background of community detection and cross-species analysis. In the field of community detection, we review some proposed methods and concepts. Moreover, we cite some works regarding community detection in biological networks. Closing this section, we mention some studies on cross-species analysis and some involving multilayer network approaches in biological networks.

3.1 Community Detection

Methods. The study of communities has become the most studied property regarding the structure of a network. Here, we review several Network Science methods developed to solve the problem of community detection. Starting with the divisive algorithms, the well-known Girvan-Newman [Girvan and Newman 2002] is the most commonly used algorithm. Modularity-optimization-based methods are the most popular class for community detection. Here, we highlight the Louvain [Blondel et al. 2008], Clauset-Newman-Moore [Clauset et al. 2004] and Leiden [Traag et al. 2019] algorithms. Another class of methods is the class of spectral algorithms, as an example, we have the Donetti-Muñoz algorithm [Donetti and Munoz 2004]. Not belonging to the previous mention categories, we have the Infomap algorithm [Rosvall and Bergstrom 2008], the Label Propagation algorithm*et al.* [Raghavan et al. 2007] and the Markov Cluster algorithm [Van Dongen 2000]. Regarding the detection of overlapping communities, CFinder [Adamcsek et al. 2006] is the software package implementing the Clique Percolation technique developed by Palla *et al.* [Palla et al. 2005]. Lastly, for detection of communities in signed networks we point the spectral algorithm by Cucuringu *et al.* [Cucuringu et al. 2019]. A more specific description of some of these algorithms can be found at [Fortunato 2010].

Communities significance. The significance of a partition is related to its robustness and stability against random perturbations of the graph structure. The idea is that if a partition is significant, it will be recovered if the structure of the graph is changed. On the other hand, if a partition is not significant, it will collapse when the structure of the graph is modified. Karrer *et al.* proposed a method to test the significance of a partition [Karrer et al. 2008]. Also Lancichinetti *et al.* [Lancichinetti et al. 2010] proposed two measures, C-score and B-score, to estimate the significance of single communities and not only of the whole partition.

Communities in Biological Networks. The study of communities is regularly used in the investigation of cellular systems of organisms such as protein-protein interaction (PPI) networks, gene regulatory networks (GRN), and metabolic networks (MN). In PPI networks

[1] http://geneontology.org/

of yeast species, the study of communities allowed the identification of modules corresponding to important protein complexes [Chen and Yuan 2006; Rives and Galitski 2003; Sen et al. 2006; Spirin and Mirny 2003]. In MN, with the investigation of the community structures, it was possible the detection of functional modules [Ahn et al. 2010; Ravasz et al. 2002]. Finally, in GRN, it was achievable the discovery of functionally related groups of genes [Wilkinson and Huberman 2004] and identification of groups of genes associated with functions that drive cancer [de Anda-Jáuregui et al. 2019].

3.2 Cross-species Comparison

A major challenge of biological research is to understand the complex networks of interacting genes and proteins that give rise to biological form and function. Approaches based on cross-species comparisons usually provide a valuable framework to address these challenges, in this section, we cite some of the works related to this topic. In PPI networks, cross-species can be used to predict protein-protein interactions (interologues) conserved across species [Matthews et al. 2001; Sharan et al. 2005; Wiles et al. 2010]. The characterization of interspecies differences in gene regulation is fundamental for understanding the diversity and evolution of species. For example, Borneman *et al.* [Borneman et al. 2007] identified considerable divergence in binding sites of transcription factors across closely related yeasts species. In another study [Stuart et al. 2003], from the comparison of correlated patterns of gene expression from different species, the authors were able to find co-expressed genes in these.

***Multilayer Approach in Biologic Networks*.** Multilayer networks are useful in the study of biological networks since it allows the combination of multiple levels of genomic and molecular interaction data. In PPI networks, this type of approach has already helped to make predictions of protein functions in yeast [Zhao et al. 2016] and in human [Liang et al. 2019]. In the medical field, this approach supported the recognition of candidate driver cancer genes [Cantini et al. 2015; Yu et al. 2019].

4 IDENTIFICATION OF FUNCTIONAL MODULES

In this section, we begin the study of the transcriptional regulatory networks. First, we introduce the networks of the species by presenting their characteristics. Then, we start the study of networks with the detection of modules and following functional characterization.

4.1 Data

The data we use in this work is a series of transcriptional regulatory networks from different yeast species. In particular, we consider the data from the Yeastract+[2] portal which provides the transcriptional regulatory networks of 10 closely-related yeast species [Monteiro et al. 2019]. The characteristics of these networks are presented in Table 1.

From Table 1, it is clear that the species have different levels of documentation, as reflected by the number of nodes and edges. *S. cerevisiae* is the network with more regulatory associations between transcription factors and target genes. These associations may be classified into two major groups: (1) those supported by DNA binding evidence; (2) those supported by expression evidence. Due to the high level of information of *S. cerevisiae*, we add a new network to our set. *S. cerevisiae B* consists of filtering the original network keeping only the regulatory associations supported by binding evidence. This filtering aims to clarify the future interpretation of the results in this species. Comparing the characteristics of the original and filtered networks, we observe that the number of nodes, transcription factors, and target genes remains close to the original. This indicates that the filtering of the original network managed to retain most of the genetic evidence of *S. cerevisiae*. Unlike the species mentioned above, there are species whose networks are small and sparse. Enumerating these species we have: *C. tropicalis*, *K. pastoris*, *K. lactis*, *Z. bailii* and *K. marxianus*. This lack of genetic evidence suggests that the characterization of these species may not reflect their biological nature. Therefore, we decide to discard these networks from the current analysis. *S. cerevisiae* and *C. albicans* are the species with the highest node degree, which is normal since these have more transcription factors than the others and each node is expected to be involved in multiple processes. On the other hand, *Y. lipolytica* is the one with fewer transcription factors, only five, and its clustering coefficient is the highest among the networks, revealing that the nodes are concentrated around those transcription factors. Due to this structural organization, it is likely that in this species the number of modules detected is limited by the number of transcription factors.

Network	**#Nodes**	**#Edges**	**#TFs**	**#TGs**	$\langle k_{in} \rangle$	$\langle k_{out} \rangle$	**CC**	**D**
S. cerevisiae	6 886	195 498	220	6 886	28.40	28.40	0.47	4
S. cerevisiae B	6 478	45 209	176	6 475	6.98	6.98	0.22	5
C. albicans	6 015	35 687	118	6 015	5.93	5.93	0.28	5
Y. lipolytica	5 288	9 238	5	5 288	1.75	1.75	0.36	4
C. parapsilosis	3 381	6 986	11	3 380	2.07	2.07	0.25	4
C. glabrata	2 133	3 508	40	2 116	1.64	1.64	0.09	6*
C. tropicalis	665	698	16	663	1.05	1.05	0.01	5
K. pastoris	561	581	4	559	1.04	1.04	0.01	5
K. lactis	111	126	10	106	1.14	1.14	0.15	2*
Z. bailii	32	31	1	31	0.97	0.97	0.00	2
K. marxianus	4	3	1	3	0.75	0.75	0.00	2

Table 1. Networks Properties. CC stands for Clustering Coefficient and D for Diameter, $\langle k_{in} \rangle$ for average in-degree and $\langle k_{out} \rangle$ for average out-degree. In the Diameter field, a value followed by a * represents the value of the Diameter for the largest component of the graph.

4.2 Comparative Analysis of Modules

The first phase of our approach is the detection of modules. We select a collection of algorithms that exploit the diverse ideas and techniques of Network Science developed over the years. The set is composed of the following algorithms: Girvan-Newman (GN), Louvain, Leiden, Clauset-Newman-Moore (CNM), Label Propagation (LP), Markov Clustering (MC), Infomap, CFinder (CF), and a spectral clustering technique (SC) for modules detection on signed networks. With the application of the spectral technique, we hope to verify if the networks contain polarized modules that may be important for understanding their structure. To execute the introduced algorithms, we used libraries where they are already implemented.

[2] http://yeastract-plus.org

Some of the considered algorithms are stochastic, i.e, the result may change in each run because their procedure depends on random events. The Louvain, the Label Propagation, and the Infomap are the non-deterministic algorithms we use in our approach. To compare the different outputs of the algorithms, we run these 1 000 times. To study the different partitions given, we compare each pair of different partitions having the number of modules equal to the value of the mode. To make this comparison, we use the package *clusim* [Gates and Ahn 2019] that allows us to compare different partitions using similarity measures, in our case we use *Rand Index* [Rand 1971]. The similarity results achieved a high value and with low variance. Therefore, despite the stochasticity of the algorithms, the high similarity and low variance show that the structural differences between the partitions are minimal. Thus, regarding stochastic algorithms, we adopt one of the results having the number of modules equal to the mode.

Due to the temporal complexity of Girvan-Newman and CFinder algorithms, it was not possible to run them on some of the biggest networks. We tried to run these algorithms for a timeout of two weeks, however, the execution of these algorithms did not come to an end. Table 2 displays the number of modules obtained for the networks using the different algorithms of our set.

Network	GN	Louvain	Leiden	CNM	LP	MC	Infomap	CF	SC
S. cerevisiae	-	5	5	3	1	1	54	-	2
S. cerevisiae B	-	12	11	6	1	78	48	34	2
C. albicans	-	12	12	7	1	11	23	19	-
Y. lipolytica	1	4	4	4	1	1	1	3	-
C. parapsilosis	25	8	8	6	1	2	5	4	-
C. glabrata	17	14	13	12	16	24	29	14	-

Table 2. Number of modules obtained for each network using the different algorithms.

The results in Table 2 show that different algorithms give different results regarding the number of modules obtained. Some of the algorithms fail to detect modules, such as the Label Propagation, Girvan-Newman, and Spectral Clustering in signed networks, this lead us not to choose to study these results. There is a great divergence between the number of modules obtained between *S. cerevisiae* and *S. cerevisiae B*. Therefore, the filter applied to create *S. cerevisiae B* network reveals to be essential in the search for modules in this species. Whereas that the division of *S. cerevisiae B* in modules points to a better division of species, we decide on using the filtered network to study *S. cerevisiae*. In *Y. lipolytica* few modules were detected, a consequence of the low number of transcription factors. Thus, in the future functional analysis of these modules, few functions should be identified for this species. Regarding the rest of the species, it was possible to extract some modules. Indicating that the functional characterization of these may be more complete. To better understand the division in modules, we decide to study the distribution of their sizes for the different algorithms. We Figure 2 we present the distributions for *C.albicans* as example.

We expect that a balanced division of the networks (modules of the same magnitude) should be the case that better reflects the division of species according to their biological function. The distribution shows that the modularity-based algorithms (Louvain, Leiden and, Clauset-Newman-Moore) have a more balanced division than the others. Infomap, despite some very small modules, produced others with equivalent size to those mentioned above. In the case of CFinder, although it has modules that include almost the entire network, the smaller ones can help us to understand if the species benefit from an overlapping communities study. Lastly, Markov Clustering gives us a very unbalanced division, so we decide to discard these results.

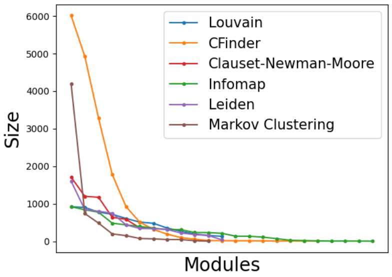


Fig. 2. Modules size distribution for *C. albicans*

To close the first phase of our analysis, we analyze the significance of the modules obtained with the modularity-based algorithms. For this purpose, we calculate their *C*-score and *B*-score, Table 3. Looking at the *C*-score values, in none of the algorithms it was possible to identify significant modules. However, the *B*-score says the opposite, indicating that the *C*-score is a very restrictive measure. According to the *B*-score values, the Louvain algorithm only produced one significant module, which may be a consequence of its stochasticity. Regarding the other two algorithms, both produce significant modules. Combining the significance of some modules and the balanced division, at that point, Leiden seems to be the one that best captures the structure of the species. Nevertheless, in the functional analysis, we take into account the results of Infomap, CFinder, Louvain, and Clauset-Newman-Moore, which also present interesting results.

	Louvain		Leiden		Clauset-Newman-Moore	
C	*C*-score	*B*-score	*C*-score	*B*-score	*C*-score	*B*-score
0	1.00	1.00	0.99	1.02e-27	0.97	6.53e-67
1	1.00	1.00	0.99	0.39	1.00	2.07e-69
2	0.99	0.29	1.00	0.01	0.98	1.17e-16
3	1.00	1.00	1.00	0.99	0.99	0.99
4	0.99	1.00	0.99	0.63	0.99	0.01
5	0.99	1.00	0.99	0.01	0.99	0.99
6	0.99	1.00	0.99	1.00	-	-
7	1,00	1.00	0.99	0.83e-9	-	-
8	0.99	1.00	0.99	1.00	-	-
9	0.99	1.00	0.99	0.01	-	-
10	1.00	1.00	0.99	0.32	-	-
11	0.99	1.33e-70	-	-	-	-

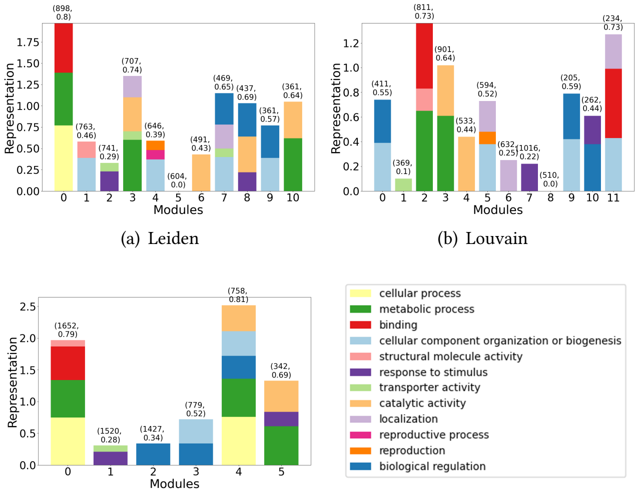
Table 3. Significance of the modules obtained for *S. cerevisiae*.

4.3 Functional Analysis of Modules

This section refers to the label assignment process that consists of assigning one or more labels to the found modules. These labels represent specific functionalities of species. Therefore, it allows the functional characterization of these. The idea of the labeling process is to associate to the modules the most represented Gene Ontology terms among their genes. To obtain all terms associated with a module, we need to obtain all terms directly linked with the genes, then, we go through all the terms in the hierarchy until we reach the root to obtain the higher level terms of the module.

To label the modules, we have to identify the most significant and representative terms of each module. Given the set of terms associated with a module, we perform a three-step filtering of the terms: (1) select only the most global terms (level 2 and 3 terms); (2) keep only the most specific terms of the module using the statistical measure p-value; (3) retain the terms with a good representation in the module (represented in at least 10% of the genes).

Algorithms Performance. Using *S. cerevisiae* as a reference, we compare the performance of algorithms that we consider to have interesting results. Beginning with the modularity-based methods, Figure 3. A first look shows that most modules have more than one label, exposing the functional diversity of these. However, it is observable that not all genes in the modules are linked to functionalities that characterize the modules they belong to. By applying the p-value filtering, we obtain only the most specific terms of each module. Therefore, there are always fractions of genes in the modules that are not associated with any of the terms. These genes correspond to behaviors that end up being captured in other modules.



(a) Leiden

(b) Louvain

(c) Clauset-Newman-Moore

(d) Gene Ontology Terms

Fig. 3. Modules and respective functions for modularity-based methods on *S. cerevisiae*. The bar of each term symbolizes its representation in the module. The pair of values at the top of each bar are respectively the size of the module and the percentage of genes of the module related with at least one term (in the module).

In Figure 3, we observe that some functions appear with high representation in the modules. Such as the metabolic process, cellular process, biological regulation, or response to stimulus. This points to the importance that these functions have in the species. In contrast, others seem to be less represented. Being specific functions, these are associated with a smaller set of genes. Reproduction, reproductive process, and transporter activity are good examples of specific functions detected in the modules. The Clauset-Newman-Moore algorithm is the worst performer algorithm, capturing the least diversity of functions. Comparing the results from Louvain and Leiden we can observe that some modules are very similar in terms of functionality. However, Leiden was able to identify functions that Louvain could not, such as the cellular process (usually heavily represented in modules) or reproductive process. Moreover, in general, the modules from Leiden have more functional diversity and the ratio of genes that contribute to the classification of the modules is greater. This combination of factors leads us to conclude that the Leiden algorithm had a better performance in dividing and capturing the functionalities of the species.

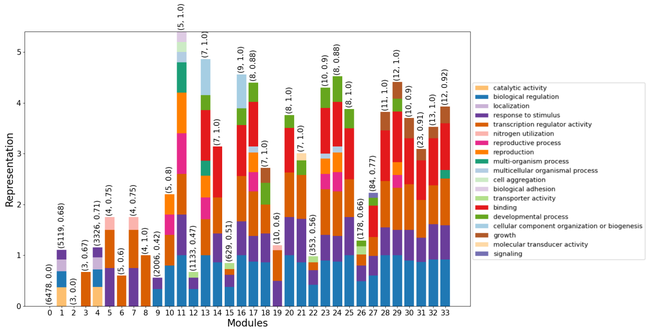


Fig. 4. Modules of *S. cerevisiae* obtained with CFinder and respective functions.

In Figure 4 we present the results of the label assignment process in the modules of *S. cerevisiae* found with CFinder. The majority of the modules are too small and their classification does not help us to characterize the species. However, we can retain some new information about the species, such as the presence of previously not detected functions in modules of acceptable size. For example, in *M27* we notice the presence of the functions: transcription regulator activity, developmental process, and signaling. Finally, we also notice *M1*, which represents almost the entire species and has four associated functions with good representation (all of them previously detected with the Leiden algorithm). This evidence helps us to confirm that these are important functions in this organism.

Lastly, in Figure 5, we witness the poor performance of Infomap. Although it managed to classify some modules of relevant size, it failed to classify the vast majority of modules.

Functional Analysis of Remaining Species. Closing this chapter, we analyze the results of the label assignment process for the remaining species in the study. We use the results for the modules obtained with the Leiden algorithm, Figure 6, since it is the algorithm with the best performance for *S. cerevisiae*.

Starting with *C. albicans*, we notice the absence of terms in *M0*, *M9*, *M10*. In *M0*, since the module encompasses a large part of the species, it is difficult to detect significant terms using the p-value. All the remaining modules are associated with at least one function.

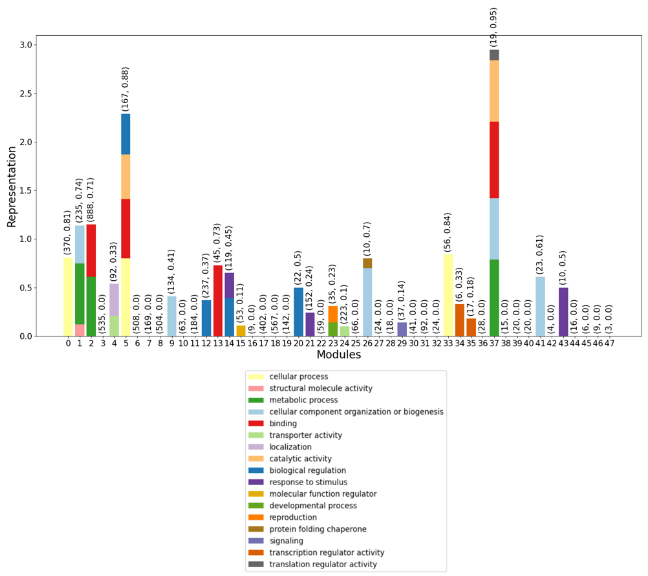


Fig. 5. Modules of *S. cerevisiae* obtained with Infomap and respective functions.

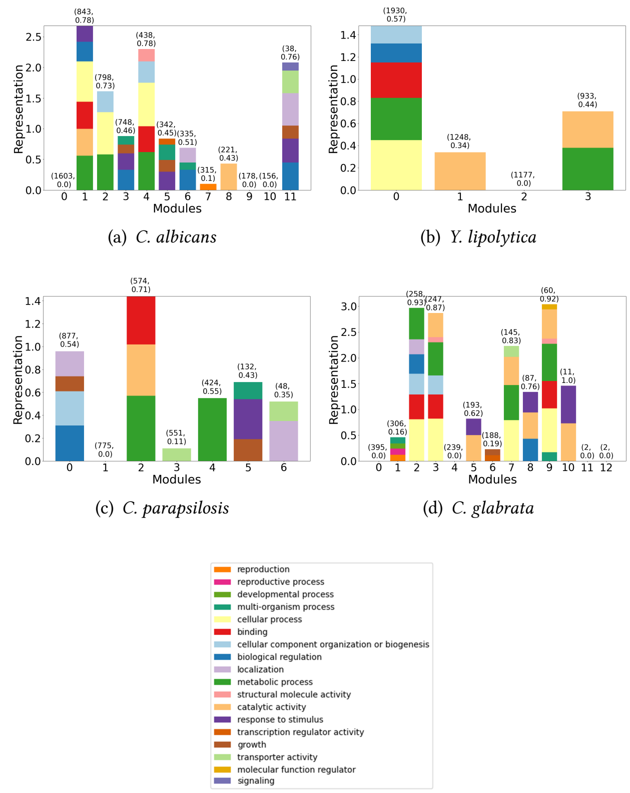
Many of those are associated with three or more terms, capturing many of the functions of the species. An interesting point is the association of some modules to functions such as multi-organism process and growth, which are not sufficiently representative/significant to be associated with a module in *S. cerevisiae*. Also in *C. parapsilosis* and *C. glabrata*, some modules are associated with functions not detected in *S. cerevisiae*. Due to the large sizes of *S. cerevisiae* modules, it is difficult for specific terms to have a good representation in these, since they are associated with few genes. In all of these species, general functions already captured in *S. cerevisiae* were also detected, such as metabolic process, response to stimulus, or biological regulation. Revealing once again the central role these have in the functionality of different organisms. It is noticed that the modules of *C. glabrata* are associated with more functionality than the modules of *C. parapsilosis* and *Y. lipolytica*, although we have more generic evidence on the last two. Whereas that *C. glabrata* has more transcription factors, we assume that the information about this species contains genetic evidence about more biological processes. This results in a more diversified classification of modules in comparison to *C. parapsilosis* and *Y. lipolytica*.

5 CROSS-SPECIES COMPARISON

Here, we begin our cross-species comparison. First, we compare the functional modules discovered in Section 4 and we settle some similarities between species. Then, we move to a multilayer approach where we search for potential functional structures conserved among species.

5.1 Functional Comparison of Modules

We resort to the homology mappings between species to establish the connections between modules. Each link in a homology mapping denotes the connection between two homologous genes. In biology, it is established that the DNA sequence of two homologous genes derives from a common ancestor (may or may not have the same function). For this work, the homology mappings are obtained from YEASTRACT+ [Monteiro et al. 2019].



(a) *C. albicans*

(b) *Y. lipolytica*

(c) *C. parapsilosis*

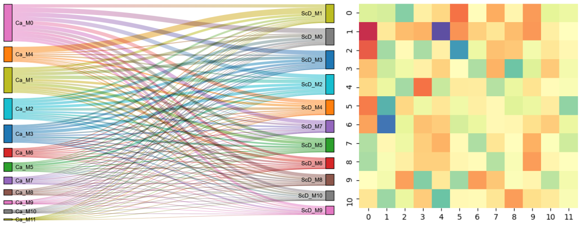
(d) *C. glabrata*

(e) Gene Ontology Terms

Fig. 6. Label Assignment results for the different species using Leiden algorithm.

***S. cerevisiae* vs *C. albicans*.** In this subsection, we portray the comparing process between *S. cerevisiae* and *C. albicans*. For this purpose, we explore the level of connection between the functional modules obtained with the Leiden algorithm. In Figure 7(a) we present a Sankey diagram representing the connections between the modules for both species.

To understand the level of connection between modules, we perform an analysis to assess the quality of the mappings. First, we calculate the number of links shared between every pair of modules of the two species. Then, we compare these distributions with 1 000 realizations of the same process in a null model, which consists of maintaining the community structure of both networks but with randomization of the nodes. Consequently, this procedure results in different mappings between species. In Figure 7(b) we introduce the heat map of the z-scores representing the level of connection between modules. The heat map reveals the existence of some pairs



(a) Sankey Diagram

(b) Heat map

Fig. 7. Figure 7(a) - Sankey diagram representing the connections between the modules of *S. cerevisiae* and *C. albicans*. Figure 7(b) - heat map representing the level of connectivity between the modules of *S. cerevisiae* and *C. albicans*.

	Terms					
Connections	GO:0071840	GO:0005198	GO:0008152	GO:0009987	GO:0005488	GO:0065007
M1-Sc	🟩 0.10	🟩 0.14	🟥 0.17	🟥 0.18	🟥 0.07	
M4-Ca	🟩 0.13	🟩 0.16	🟩 0.21	🟩 0.23	🟩 0.11	
M0-Sc	🟥 0.03		🟩 0.09	🟩 0.10	🟩 0.06	🟥 0.04
M0-Yl	🟩 0.01		🟩 0.04	🟩 0.05	🟩 0.03	🟩 0.02
M0-Ca	🟥 0.03		🟥 0.10	🟥 0.12	🟥 0.07	🟥 0.04
M0-Yl	🟩 0.03		🟩 0.09	🟩 0.10	🟩 0.06	🟩 0.04

Terms	**Function**
GO:0071840	cellular component organization or biogenesis
GO:0005198	structural molecule activity
GO:0008152	metabolic process
GO:0009987	cellular process
GO:0005488	binding
GO:0065007	biological regulation

Table 4. Strongly connected pairs of modules from different species. For each module, we can consult the percentage of genes that have homologous with the same function in the other one that is part of the connection. Looking at the first pair, it is possible to verify that in *M6* of *S. cerevisiae*, 0.09% of the genes participate in the connections related to the metabolic process. A green cell means that the term was found in the module through the label assignment process, a cell in red denotes the opposite (the term was not found in the module).

of modules with strong connections in relation to others (green and blue colors).

By consulting the functional characterization of module pairs with stronger connectivity, we verify the sharing of functions between some of the modules. This circumstance points to homologous genes with the same function as the cause for the strong connectivity in some of the pairs of modules. One good example is the pair of modules *M0* and *M2* of *S. cerevisiae* and *C. albicans* respectively. In both cases, the metabolic and cellular processes are widely represented terms, homologous genes associated with those functions may be the origin for this solid connection. However, in other cases, mutual labels only represent a small part of the genes of the modules. Such as in *M1* of *S. cerevisiae* and *M4* of *C. albicans*, that is by far the strongest connection between the two species. In this case, the mutual functions between modules seem not to be sufficient justification for such a strong connection. Thus, this strong connection may arise from other events, such as the sharing of functions that were only detected in one of the modules (cellular and metabolic process). Closing the analysis, we notice the connection between *M5* of *S. cerevisiae* and *M1* of *C. albicans*. The functions of *M1* of *C. albicans* may serve as predictions for possible functions in *M5* of *S. cerevisiae* since this one is unlabeled.

Detailed Analysis of Connections. Here, we perform a detailed analysis of some strong connections between modules. For this purpose, we examine the terms associated with the links of the connections. A term is associated with a link if the term is common to the homologous genes in it. In Table 4 we present some of the most relevant connections.

The detailed analysis of the connections demonstrates that there are functional elements of considerable size in different species formed by homologous genes with the same functions. Since a homologous gene is a gene inherited in two species by a common ancestor, this evidence reveals the conservation of functional elements across different organisms. Also, using the information of Table 4, we can diagnose functional elements in some modules that were not detected until now. Such as the metabolic and cellular processes in *M1* of *S. cerevisiae*. Finally, we look at the connection between *M0* of *C. albicans* and *M0* of *Y. lipolytica*. With this cross-species analysis, we unveil some functional elements present in *M0* of *C. albicans*. With this new information, it is clear that the absence of labels assigned to this module in the label assignment process results from its large size.

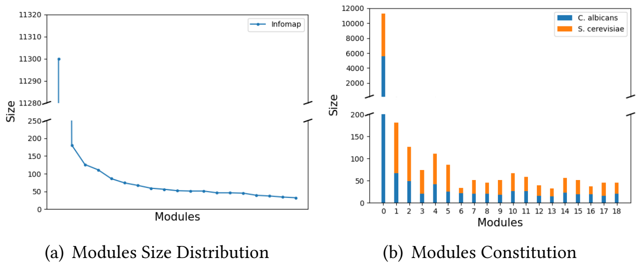
5.2 Multilayer Approach for Cross-Species Comparison

In the previous section, we found functional elements conserved across species. However, we did not check if these elements have other associated functions or even if they overlap, since each gene can be associated with more than one term. Therefore, in this final step, we build a multilayer network between species in which we perform a modules detection using the Infomap algorithm since it is suitable for this type of network. With the detection and functional characterization of the modules, we seek to identify and characterize functional structures conserved across species. In this multilayer network, the inter-layer links are those of the homology mappings between species.

Once again, we use the species *S. cerevisiae* and *C. albicans* to create the multilayer network. From the detection of modules, we could find several modules. The size distribution of those modules is displayed in Figure 8(a). We notice that one of the modules encompasses the vast majority of genes from both species. Therefore, this one should not contain characteristic information about the genetic conjugation of both species. On other hand, the remaining modules are smaller and contain equivalent sizes. In Figure 8(b) it is possible to verify the balanced constitution of the modules.

Going further with our analysis, we study the contribution of the genes of each species for the classification of the modules in the multilayer network. The comparison between the labels of each module and those of the respective gene groups can be seen in Figure 9.

Looking at the classification of the first module, we can see that by dividing the module into two groups it is possible to detect some functionalities (which is not possible in the entire module). However, the labels are unrelated, indicating that this module cannot provide useful information about the similarity between species. Regarding



(a) Modules Size Distribution

(b) Modules Constitution

Fig. 8. Figure 8(a) - size distribution of the modules found with Infomap algorithm in the multilayer network of *S. cerevisiae* and *C. albicans*. Figure 8(b) - constitution of the modules found in the multilayer network of *S. cerevisiae* and *C. albicans*.

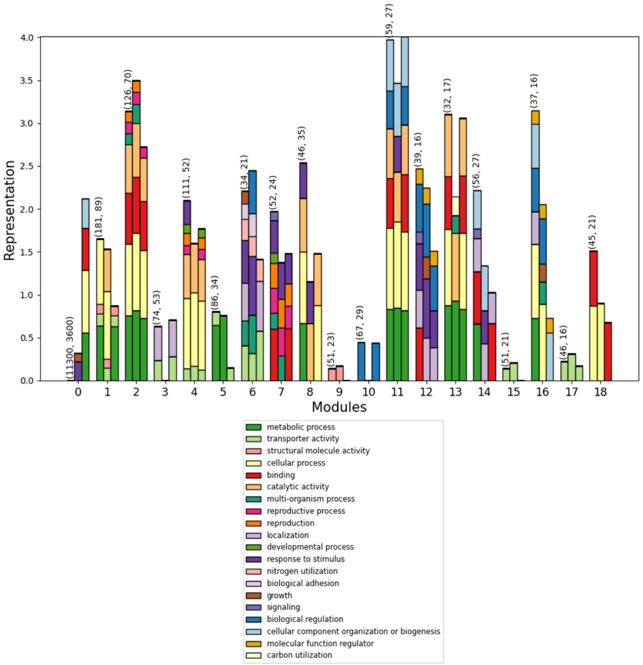


Fig. 9. Comparison of labels between the modules of the multilayer and the respective groups of genes from *S. cerevisiae* and *C. albicans*. The three bars side-by-side respectively describe the labels of the module, of the genes from *S.cerevisiae* and the genes from *C. albicans*. At the top of the first bar of each module is shown the module size and the number of inter-layer links in the module.

the other modules, by comparing the number of inter-layer links with the size of the modules, we deduce that a considerable proportion of these modules are composed of homologous genes of the two species. Thus, the detection of modules was able to detect compact structures composed of genes from both species. Some modules are mostly classified and are the result of the combination of functionally identical homologous genes from the two species. The majority of the functionality of these modules is present in the genes of both species. Thus, we consider these modules as functional structures conserved in the species. In these circumstances, we can include the modules *M2*, *M4*, *M7*, *M11*, *M12* and *M13*. By analyzing the representation of the functions in those modules, we recognize that there are functions equally represented. Such as the metabolic and cellular process in *M2* and *M11* or reproduction and reproductive process in *M7*. This evidence confirms that part of the functional elements identified as conserved in Section 5.1 are actually the same structure.

6 CONCLUSIONS

In this thesis, we studied transcriptional regulatory networks of yeast species. With the results obtained in this work, we managed to contribute with relevant information about the species in study. Let us detail these developments.

Our contributions began in Section 4 with the functional analysis of modules detected in the species. From the algorithms used for the detection of modules, the methods based on optimization of the modularity achieved better performance. Of these, we highlight Leiden, which best managed to combine a balanced division of modules with a good functional classification. The classification of modules revealed that there are biological functions more represented than others among species. Suggesting that these are central processes in the development of the organisms. From these processes, we can enumerate the metabolic process, cellular process, biological regulation ,or response to stimulus. From the comparison of functional modules between species, we identified some functions such as growth or multi-organism process in species with less genetic evidence that were not detected in *S. cerevisiae*. Also, we observed that the quantity of genetic evidence does not translate into a better functional characterization of species. We conclude that the functional diversity detected in species is correlated to the number of transcription factors and the different processes in which they participate.

In Section 5, the cross-species comparison allowed us to draw some conclusions about the genetic similarity that exists between species. First, by evaluating the degree of connection between modules of different species, we verified the existence of some strong connections. It was demonstrated that these strong connections have their origin in the conservation of functional elements in the modules. The structural elements conserved in the modules were identified as being formed by homologous genes associated with important functions such as metabolic or cellular processes. These connections were also fundamental to infer new functional elements in some modules. Finally, with the creation of the multilayer, we confirmed the existence of preserved structures across species. In these preserved structures, we were able to verify the combination of functions previously defined as conserved.

6.1 Limitations and Future Work

Although we have achieved good results with our approach, we have encountered some limitations. The biggest constraints reside in the label assignment process. The first is the difficulty of finding meaningful terms with the p-value approach in large modules (in relation to the others). Therefore, if there is an unbalanced division of the network, it will be difficult to label the large modules. Also, the threshold we used to consider a term as relevant in a module (10%) may be too restrictive. As a consequence, specific terms (only associated with a small set of nodes), may end up not being detected by the method. To overcome this problem, a possible solution would be to adapt the threshold value to the size of the modules. Thus, larger modules would have lower threshold values to facilitate the detection of more specific functions.

Lastly, additional future work is worth exploring. In Section 5.2, it would be possible to explore in more detail the conserved structures found across species. For example, since we only look at global processes, an analysis of their sub-processes could reveal whether or not these modules can encode specific regulatory patterns. Furthermore, we found some genes in those modules that were not associated with any Gene Ontology terms. We could attempt to use the functions of the modules in which these genes are to predict their functionality, always taking into account that we do not have the genetic evidence to confirm the possible predictions for these genes.

REFERENCES

Balázs Adamcsek, Gergely Palla, Illés J Farkas, Imre Derényi, and Tamás Vicsek. 2006. CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22, 8 (2006), 1021–1023.

Yong-Yeol Ahn, James P Bagrow, and Sune Lehmann. 2010. Link communities reveal multiscale complexity in networks. *nature* 466, 7307 (2010), 761–764.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics* 25, 1 (2000), 25–29.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesús Gómez-Gardenes, Miguel Romance, Irene Sendina-Nadal, Zhen Wang, and Massimiliano Zanin. 2014. The structure and dynamics of multilayer networks. *Physics reports* 544, 1 (2014), 1–122.

Anthony R Borneman, Tara A Gianoulis, Zhengdong D Zhang, Haiyuan Yu, Joel Rozowsky, Michael R Seringhaus, Lu Yong Wang, Mark Gerstein, and Michael Snyder. 2007. Divergence of transcription factor binding sites across related yeast species. *Science* 317, 5839 (2007), 815–819.

Laura Cantini, Enzo Medico, Santo Fortunato, and Michele Caselle. 2015. Detection of gene communities in multi-networks reveals cancer drivers. *Scientific reports* 5, 1 (2015), 1–10.

Jingchun Chen and Bo Yuan. 2006. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics* 22, 18 (2006), 2283–2290.

Aaron Clauset, Mark EJ Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical review E* 70, 6 (2004), 066111.

Mihai Cucuringu, Peter Davies, Aldo Glielmo, and Hemant Tyagi. 2019. SPONGE: A generalized eigenproblem for clustering signed networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 1088–1098.

Eric H Davidson. 2001. *Genomic regulatory systems: in development and evolution*. Elsevier.

Eric H Davidson, Jonathan P Rast, Paola Oliveri, Andrew Ransick, Cristina Calestani, Chiou-Hwa Yuh, Takuya Minokawa, Gabriele Amore, Veronica Hinman, Cesar Arenas-Mena, et al. 2002. A genomic regulatory network for development. *science* 295, 5560 (2002), 1669–1678.

Guillermo de Anda-Jáuregui, Sergio Antonio Alcalá-Corona, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. 2019. Functional and transcriptional connectivity of communities in breast cancer co-expression networks. *Applied Network Science* 4, 1 (2019), 1–13.

Luca Donetti and Miguel A Munoz. 2004. Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment* 2004, 10 (2004), P10012.

Santo Fortunato. 2010. Community detection in graphs. *Physics reports* 486, 3-5 (2010), 75–174.

Alexander J Gates and Yong-Yeol Ahn. 2019. CluSim: A Python package for calculating clustering similarity. *Journal of Open Source Software* 4, 35 (2019), 1264.

Michelle Girvan and Mark EJ Newman. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99, 12 (2002), 7821–7826.

Brian Karrer, Elizaveta Levina, and Mark EJ Newman. 2008. Robustness of community structure in networks. *Physical review E* 77, 4 (2008), 046119.

Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. 2014. Multilayer networks. *Journal of complex networks* 2, 3 (2014), 203–271.

Andrea Lancichinetti, Filippo Radicchi, and José J Ramasco. 2010. Statistical significance of communities in networks. *Physical Review E* 81, 4 (2010), 046110.

Anna CF Lewis, Nick S Jones, Mason A Porter, and Charlotte M Deane. 2010. The function of communities in protein interaction networks at multiple scales. *BMC systems biology* 4, 1 (2010), 1–14.

Lifan Liang, Vicky Chen, Kunju Zhu, Xiaonan Fan, Xinghua Lu, and Songjian Lu. 2019. Integrating data and knowledge to identify functional modules of genes: a multilayer approach. *BMC bioinformatics* 20, 1 (2019), 1–15.

Nicholas M Luscombe, M Madan Babu, Haiyuan Yu, Michael Snyder, Sarah A Teichmann, and Mark Gerstein. 2004. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431, 7006 (2004), 308–312.

Lisa R Matthews, Philippe Vaglio, Jérôme Reboul, Hui Ge, Brian P Davis, James Garrels, Sylvie Vincent, and Marc Vidal. 2001. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome research* 11, 12 (2001), 2120–2126.

Pedro T Monteiro, Jorge Oliveira, Pedro Pais, Miguel Antunes, Margarida Palma, Mafalda Cavalheiro, Mónica Galocha, Cláudia P Godinho, Luís C Martins, Nuno Bourbon, Marta N Mota, Ricardo A Ribeiro, Romeu Viana, Isabel Sá-Correia, and Miguel C Teixeira. 2019. YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic Acids Research* 48, D1 (Oct. 2019), D642–D649. https://doi.org/10.1093/nar/gkz859

Mark EJ Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69, 2 (2004), 026113.

Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *nature* 435, 7043 (2005), 814–818.

Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. 2007. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E* 76, 3 (2007), 036106.

William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66, 336 (1971), 846–850.

Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. 2002. Hierarchical organization of modularity in metabolic networks. *science* 297, 5586 (2002), 1551–1555.

Alexander W Rives and Timothy Galitski. 2003. Modular organization of cellular networks. *Proceedings of the national Academy of sciences* 100, 3 (2003), 1128–1133.

Martin Rosvall and Carl T Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* 105, 4 (2008), 1118–1123.

Taner Z Sen, Andrzej Kloczkowski, and Robert L Jernigan. 2006. Functional clustering of yeast proteins from the protein-protein interaction network. *BMC bioinformatics* 7, 1 (2006), 1–13.

Roded Sharan, Silpa Suthram, Ryan M Kelley, Tanja Kuhn, Scott McCuine, Peter Uetz, Taylor Sittler, Richard M Karp, and Trey Ideker. 2005. Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences* 102, 6 (2005), 1974–1979.

Victor Spirin and Leonid A Mirny. 2003. Protein complexes and functional modules in molecular networks. *Proceedings of the national Academy of sciences* 100, 21 (2003), 12123–12128.

Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *science* 302, 5643 (2003), 249–255.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports* 9, 1 (2019), 1–12.

Stijn Marinus Van Dongen. 2000. *Graph clustering by flow simulation*. Ph. D. Dissertation. Faculteit Wiskunde en Informatica, Universiteit Utrecht.

Konstantin Voevodski, Shang-Hua Teng, and Yu Xia. 2009. Finding local communities in protein networks. *BMC bioinformatics* 10, 1 (2009), 1–14.

Amy M Wiles, Mark Doderer, Jianhua Ruan, Ting-Ting Gu, Dashnamoorthy Ravi, Barron Blackman, and Alexander JR Bishop. 2010. Building and analyzing protein interactome networks by cross-species comparisons. *BMC systems biology* 4, 1 (2010), 1–16.

Dennis M Wilkinson and Bernardo A Huberman. 2004. A method for finding communities of related genes. *proceedings of the national Academy of sciences* 101, suppl 1 (2004), 5241–5248.

Liang Yu, Yayong Shi, Quan Zou, and Lin Gao. 2019. Studying the drug treatment pattern based on the action of drug and multi-layer network model. *bioRxiv* (2019), 780858.

Bihai Zhao, Sai Hu, Xueyong Li, Fan Zhang, Qinglong Tian, and Wenyin Ni. 2016. An efficient method for protein function annotation based on multilayer protein networks. *Human genomics* 10, 1 (2016), 1–15.