TÉCNICO LISBOA

Functional characterization of transcriptional regulatory networks of yeast species

Paulo Alexandre Cardoso Dias

Thesis to obtain the Master of Science Degree in

Engenharia Informática e de Computadores

Supervisors: Prof. Pedro Tiago Gonçalves Monteiro
Prof. Andreia Sofia Monteiro Teixeira

Examination Committee

Chairperson: Prof. Luís Manuel Antunes Veiga
Supervisor: Prof. Pedro Tiago Gonçalves Monteiro
Member of the Committee: Prof. Fernando Pedro Pascoal dos Santos

October 2021

Acknowledgments

I could only begin by thanking my supervisors, Prof. Pedro Monteiro and Prof. Andreia Sofia Teixeira, for all their availability, guidance, and commitment during the last year. They were always present and willing to help, without them this work would not exist.

I would also like to thank my family. Especially to my parents, not only for the support they gave me throughout my academic journey but also for their spirit of sacrifice. They are a great inspiration for me. I also make a special mention to my sisters and my cousins who are always good company and good advisors.

To my friends, I am grateful for the company that was essential throughout the journey, especially in the late pandemic times. Moreover, I clearly have to acknowledge my pets, they were also very important during the course of this year.

Abstract

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.

Keywords

Complex Networks; Transcriptional Regulatory Networks; Multilayer Networks; Community Detection; Functional Modules.

Resumo

As redes regulatórias de transcrição são responsáveis pelo controlo da expressão genética. Estas são compostas por diversas interações entre fatores de transcrição e os seus genes-alvo. Sendo portadoras de uma natureza combinatória que engloba diversos processos regulatórios, estas são responsáveis pelas respostas dos organismos a perturbações que podem ocorrer no ambiente circundante. Nesta tese, exploramos várias possibilidades para o estudo de redes regulatórias de transcrição. A intenção é revelar quais as funções e/ou processos codificados nos padrões regulatórios que constituem as redes de transcrição de genes. Para isso, estudamos um conjunto de redes regulatórias de transcrição de espécies de levedura fortemente relacionadas entre si usando diferentes métodos, dividindo o fluxo de trabalho em duas fases. A primeira fase consiste na deteção de módulos seguida da sua caracterização funcional. Com isso, mostramos que as redes podem ser divididas em módulos funcionais representativos das funções biológicas das respetivas espécies. Na segunda fase, avançamos para uma análise entre espécies. Aqui, comparamos os elementos funcionais das diferentes espécies e estudamos as semelhanças entre eles. O objetivo desta análise é descobrir se existem elementos funcionais conservados entre os distintos organismos. No geral, esta dissertação fornece uma nova forma de analisar como a estrutura e a funcionalidade das redes regulatórias de diferentes espécies se podem relacionar. Além disso, exploramos como as semelhanças entre espécies podem ajudar a inferir algumas propriedades nas diferentes redes.

Palavras Chave

Redes Complexas; Redes regulatórias de transcrição; Redes Multicamada; Deteção de Comunidades; Módulos Funcionais.

Contents

List of Figures

List of Tables

1

Introduction

Contents

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. These transcription events that culminate in gene expression are constantly changing when cells progress through the cell cycle. In response to environmental changes or during the development/growth of the organism [5–9]. The transcription factors (TFs) are the core elements in the control of the 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 combinatorial regulation of multiple transcription factors [10–14]. The hundreds of interactions between transcription factors and target genes defines a transcriptional regulatory network that underlies cellular identity and function. Moreover, the combinatorial nature of the gene transcription provides a flexible gene expression to the organisms.

The development processes in complex animals are governed by the genomic regulatory code present in the transcriptional regulatory networks, whose function is to control the transcription of the genes in space and time [11, 15]. The morphological differences between species arise from the differential regulation of genes. Therefore, the information for this differential regulation is encoded in the genomic regulatory code of each individual and it is the product of evolution [16]. The transcriptional regulatory networks of organisms are assembled during their evolution. They are a puzzle of old and new features, combining 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 [17–19]. 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 represent a wide variety of entities and their interactions. Entities are represented by nodes and interactions by edges. The representation by graphs 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, the transport network we use to move around, the supply network that takes water to our houses or the power grid that provides us with electricity. Even our nervous system, in which innumerable neurons communicate through synapses, and our regulatory interactions between thousands of genes and transcription factors, can be represented as graphs.

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

In this comparison, we use the information gathered in the functional characterization of the species to infer some similarities between 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. The study of communities has gained great importance in different areas. In the case of 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 [20–22]. The field of Network Science has offered several community detection methods based on different approaches targeting different types of problems [23].

The second phase of our analysis consists of a cross-species comparison in which we compare the functional modules identified in the first phase. 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 instance, it allows the identification of conserved structures between different organisms [24–27]. Related to cross-species studies, are the studies between different types of data, which are 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 [28, 29]. In this type of comparison, the use of a multilayer network can be very useful. This approach allows the representation and comparison of different types of information, which is not always possible with the simple and common graph representation.

1.1 Objective

In this dissertation, we provide a characterization of transcriptional regulatory networks of several closely related species with the generic goal of gaining insight about this type of networks. In particular, we consider the readily available dataset from YEASTRACT+ [30] 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 this 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. A community within a network can have different connotations depending on the type of network. In the case of transcriptional regulatory networks, we assume that a community may be associated with one or more biologic functions. We begin by performing a 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. In order to understand which technique is most appropriate for our problem, we compare the results obtained using the different algorithms. 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. Also, those similarities may help us to infer functional elements in other species. To finalize our characterization, we use a multilayer approach combining networks of different species. Then, we proceed with a final multilayer 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 thesis is organized as follows. In Chapter 2, we start by presenting the necessary background. This chapter is divided into two major sections. In the first one, we list the basic concepts of Network Science, which include the different types of graphs, their properties, and different methods to generate random graphs. We also introduce the concept of multilayer network that is fundamental for our approach. Moving to the second section, it includes the biological concepts that are the focus of our work, the transcriptional regulatory networks.

Chapter 3 focuses on the state-of-the-art underlying our work. The first section concerns community detection. Here, we describe the different community-finding algorithms developed in the Network Science field. Then, we refer to the importance/meaning the communities may have within a network. Finally, we provide some studies regarding community detection in biological networks. Moving to the second half, we specify several cross-species studies in biological networks. We also cite some works on multilayer network approaches.

In Chapter 4, we begin the study of the transcriptional regulatory networks. Firstly, we present our

networks and perform an initial analysis of their basic properties. Then, we proceed with the detection of modules and present the results obtained. We continue with the functional analysis of these modules, which is done through the label assignment process. Here, we compare and discuss the performance of the algorithms used, illustrate the label assignment process in *S.cerevisiae*, and conclude with a discussion of the results obtained in the different species.

The second phase of our analysis is described in Chapter 5. Initially, we study the degree of connection between the functional modules obtained in the first stage. For this, we use the homology mappings between species. Afterward, we build a multilayer network between species using the homology mappings as inter-layer links. We finish with the detection of modules in the multilayer network and subsequent functional characterization of these.

Finally, in Chapter 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

Background

Contents

2.1 Network Science Concepts

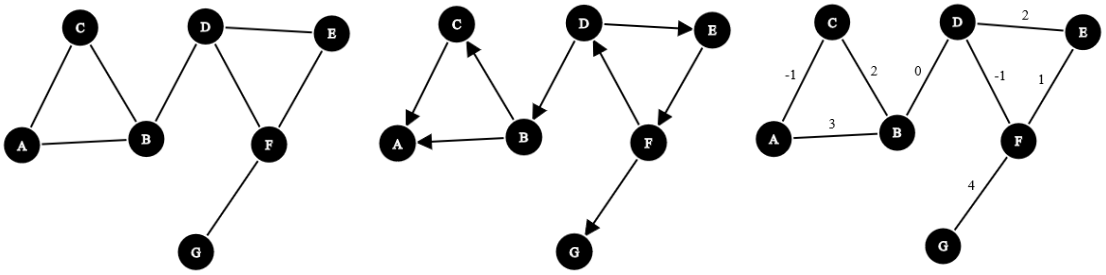
Complex systems can be modeled as graphs that represent the system entities and the interactions among them. The representation of complex systems by graphs is quite simple. However, graphs express very complex behaviors and dynamics that are characteristics of the systems they represent. The characterization of the structure and dynamics of these networks is fundamental to understanding the complex systems and their phenomena. In the following sections, we present some basic concepts of Network Science that are useful for the development of this thesis.

2.1.1 Graph Theory Basic Concepts

Networks are represented by graphs and graph theory is the branch of mathematics that studies the properties of a graph. In this section, we present some crucial concepts of graph theory.

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 $L = |E|$. We say that a graph is sparse if $|E| << N/2$.

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 co-worker in a company 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 two people that work in the same department in a company network. Two nodes i and j are *adjacent* or *neighbors* if there is an edge e connecting them, i.e., $e = (i, j) \in E$.



(a) Undirected Graph **(b)** Directed Graph **(c)** Weighted Graph

Figure 2.1: Different types of graphs. Circles represent the nodes and the lines represent the edges/links. In 2.1(a) the graph is undirected, so, the link connecting the nodes A and B means that A is connected to B and B is connected with A. The Figure 2.1(b) represents a directed graph in which the links are ordered, so, B is connected to A but the opposite is not true. In Figure 2.1(c) is presented a weighted graph, in which edges have weights that represent some measure between the nodes.

A graph $G = (V, E)$ is a *directed graph* (Figure 2.1(b)) if E is a set of ordered pairs (the connection

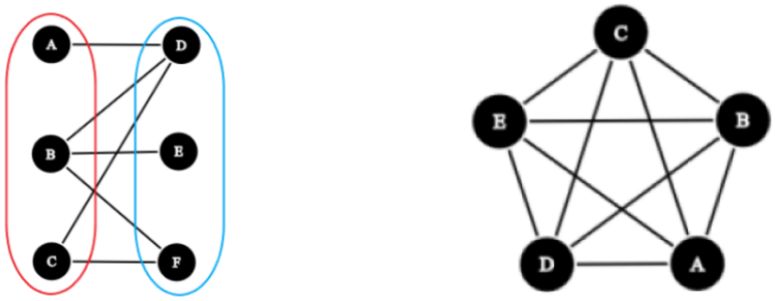
between nodes have a direction), in this case, (i, j) is different from (j, i). A graph $G = (V, E)$ is an *undirected graph* (Figure 2.1(a)) when E is a set of unordered pairs.

A *subgraph* $G' = (V', E')$ of a graph $G = (V, E)$ is a graph whose set of nodes V' is a subset of V and whose set of edges E' is a subset of E, i.e, $V' \subseteq V$ and $E' \subseteq E$, with $i, j \in V' \; \forall (i, j) \in E'$.

A *path* between two nodes i and j is the sequence of links such that each link connects two nodes, and all of these links form the path that connects the nodes i and j. The distance between two nodes i and j is the length of the path between them and is denoted by $d(i, j)$. The *shortest path* between the nodes i and j is the path with the shortest length. If there is no path between two nodes i and j, we set $d(i, j) = \infty$. The *diameter* of a graph G is the longest shortest path between any two nodes of G.

A graph G is *connected* if there is a path in G between any pair of vertices, otherwise it is *disconnected*. In a disconnected graph G, a *connected component* C of G, $C \subseteq V$, is the maximal set of nodes, such that, exists a path between any pair of nodes of C.

A *weighted graph* $G = (V, E, w)$ (Figure 2.1(c)), is a graph where each edge (i, j) has an associated weight $w(i, j)$, $w : E \to \mathbf{R}$. As an example, the weight can represent the cost of traveling between two stations i and j on a transport network. A *signed network* is a graph G where the elements of E have a binary weight $w(i, j) \in \{-1, 1\}$ that expresses a positive or negative relation between nodes.



(a) Bipartite Graph **(b)** 5-Clique Graph

Figure 2.2: Figure 2.2(a) is a bipartite graph, which we can divide into two disjoint sets, one in red (A,B,C) and the other in blue (D,E,F). The second figure, 2.2(b), is a complete graph with five nodes, we denote it by 5-clique graph.

A *bipartite graph* $G = (V, E)$, Figure 2.2(a), is a graph where the set of nodes V can be divided into two disjoint sets U and S such that each edge of E connects a node from U and a node from S.

A *clique* or a *complete graph* $G = (V, E)$, is a graph in which all nodes are connected to each other. If G is complete, we denote G by K_N or N-clique, see Figure 2.2(b).

Depending on the operations, the size of the graph, or other factors, we may need to represent the graphs in different ways. These different ways to represent a graph have their advantages and disadvantages. Next, we present the most usual representations.

Adjacency Matrix

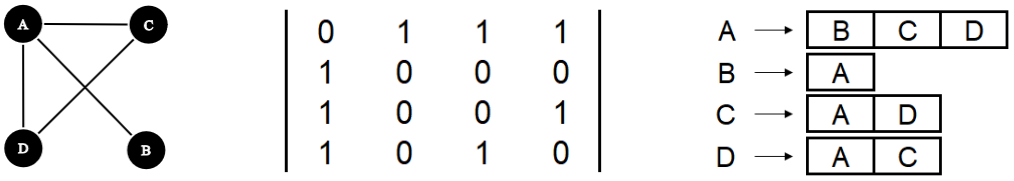
Considering a graph G, the *adjacency matrix* A is a $N \times N$ square matrix in which the entries $A_{i,j} \in \{0, 1\}$ follows the rule:

$$A_{i,j} = \begin{cases} 0, & \text{if } (i,j) \notin E \\ 1, & \text{if } (i,j) \in E. \end{cases} \tag{2.1}$$

As an example, in Figure 2.3(b), we present the adjacency matrix for the graph of Figure 2.3(a). The main advantage of this representation is the access time, it takes $O(1)$ time to check if two nodes are connected or not since we only need to look at the respective entry of the matrix. On the other hand, the main disadvantage is the space needed to represent the matrix. This representation requires $O(N^2)$ space complexity and is inefficient for large and sparse graphs.

Adjacency List

Representing a graph G with a *adjacency list* requires a linked list of size N. Each entry of this list is a list containing the neighbors for each node. Figure 2.3(c) illustrates the adjacency list for the graph of Figure 2.3(a). The main advantage of this one is that requires less space than the adjacency matrix. This representation has a total space complexity of $O(N + L)$, $O(N)$ from storing all nodes, and $O(L)$ for all the neighbors of the nodes. Unlike the previous representation, the principal disadvantage is the access time, because we have to go through all the neighbors of the node to see if there is a connection with another node. This representation is better for sparse graphs.



(a) Graph **(b)** Adjacency Matrix **(c)** Adjacency List

Figure 2.3: Graph Representation. Figure 2.3(a) is an undirected graph and Figures 2.3(b) and 2.3(b) are the representations of the graph in the form of an adjacency matrix and adjacency list respectively.

2.1.2 Network Measures

Complex networks can be characterized by *centrality measures*. These centrality measures show the importance of a node inside the network and can be based on node degree, shortest paths, or how close a node is to others, among many other network characteristics.

The *degree* of a node represents the number of links incident on the node and is denoted by k. In a social network, the degree can represent the number of friends that a person has. In a directed graph, the degree is divided in two components: *indegree* k_{in} (number of ongoing links) and *outdegree* k_{out} (number of outgoing links). The *average degree* of the network, represented by $\langle k \rangle$, is the average between the degrees of the network nodes, and the formula is given by:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{2.2}$$

The *degree distribution* is the probability that a randomly chosen node has degree k and is given by:

$$p_k = \frac{N_k}{N} \tag{2.3}$$

where N_k is the number of nodes with degree k.

The *closeness centrality* was proposed by Bavelas [31] and measures how close a node is from all the other nodes of the network and is given by the formula:

$$C(i) = \frac{N-1}{\sum_{j(j \neq i)} d_{ij}} \tag{2.4}$$

where d_{ij} is the size of the shortest path between the nodes i and j.

The *betweenness centrality* [32] is used to characterize how important a node is in the communication with others nodes. It is defined as the number of paths between pairs of nodes that go through a given node:

$$B(i) = \frac{1}{N^2} \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{2.5}$$

where $\sigma_{st}(i)$ represents the number of paths between s and t that runs through node i, and σ_{st} represents the total number of paths between s and t. Girvan and Newman generalized Freeman's betweenness centrality to edges, the edge betweenness centrality is defined as the number of the shortest paths that go through an edge in a graph [33].

The *clustering coefficient* is a measure that indicates how the nodes of a graph tend to group together [2]. For a node i of degree k the *local clustering coefficient* is defined as:

$$CC_i = \frac{2L_i}{k(k-1)} \tag{2.6}$$

where L_i represents the number of links between the neighbors of node i. The *global clustering coefficient* can be also defined as the number of closed triangles in a network:

$$CC = \frac{3 \times number\_of\_triangles}{number\_of\_triplets} \tag{2.7}$$

where a *triplet* is a group of three nodes i, j and k, which can form a triangle.

2.1.3 Modularity

In Network Science a *community* or a *module* is a group of nodes that have a higher likelihood of connecting to each other than to nodes from 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 is the measure that allow us to quantify the quality of a partition c in a graph G, i.e., allow us to compare particular modules/communities. The modularity of a set of nodes is given by the formula:

$$M_c = \frac{L_c}{L} - (\frac{k_c}{2L})^2 \tag{2.8}$$

where L_c is the number of links within the module c, L is the total number of links of the network, and k_c is the total degree of the nodes in the module c. The total modularity of the network is the sum of the values of modularity of each module [34]:

$$M = \sum_{c=1}^{n_c} [\frac{L_c}{L} - (\frac{k_c}{2L})^2] \tag{2.9}$$

where n_c is the total number of modules.

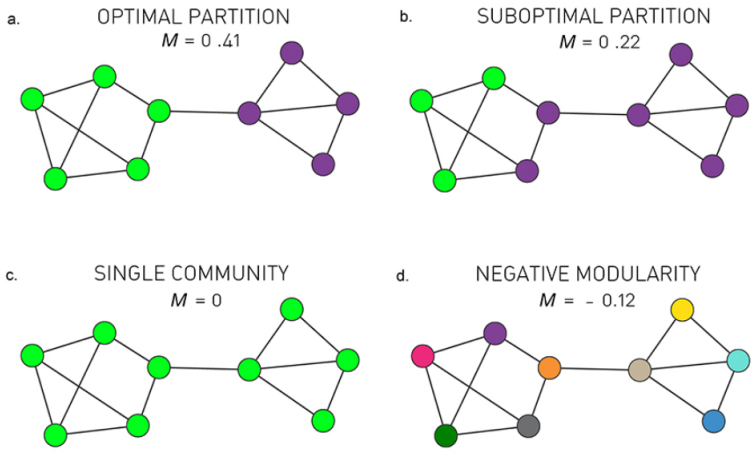


Figure 2.4: Examples of different partitions for the same network. Image from [1].

A network can have different values of modularity. This value depends on how the network is partitioned. Figure 2.4 illustrates examples of different partitions for the same network. The best partition for a network is the one with the greatest value for modularity. In Figure 2.4a., the partition with maximum modularity can accurately capture both communities, in Figure 2.4b. (lower modularity) we deviate from

these two communities. If we take the whole network as a single module, Figure 2.4c., the value of the modularity is equal to zero because the first and second terms of the equation (2.9) are equal. Finally, if each node represents a module, the value of the modularity is negative because the first term of the equation (2.9) becomes zero.

2.1.4 Random Networks

In the last decades, many real-world networks have been submitted to several studies. In these studies, scientists try to propose hypotheses that seek to explain the behaviors of these networks. To test these hypotheses is essential to have realistic network models that can reproduce the properties of these real-world networks such as degree distribution and clustering coefficient. These models act like terms of comparison for studying real-world networks. By applying the hypothesis in these models is possible to draw some conclusions about the networks in study. In this subsection, we discuss some random graph models proposed in Network Science.

The challenge of creating a random network is to decide where to place the borders between nodes to produce a network with the properties of a real system. Two main models were proposed:

- *G(N, L) Model* - proposed by Erdős and Rényi [35], N nodes are labeled and connected with L links chosen at random.
- *G(N, p) Model* - introduced by Gilbert [36], each pair of N labeled nodes is connected with probability p.

These models can produce random networks with a degree distribution that follows a Poisson distribution. However, as large real networks become available, it was found that their degree distribution is not a Poisson distribution [1].

Small-World Networks

The *small world property*, also known as *six degrees of separation* [37, 38], states that any person in the world could meet anyone, anywhere in the world, with a maximum of six or fewer acquaintances between them. This means that the distance between two randomly chosen nodes in a network is typically very short. Given that and the fact that real networks do not follow a Poisson Distribution, a new model to generate random networks was proposed by Duncan Watts and Steven Strogatz in [2], the model is represented in Figure 2.5:

- We start with a ring of nodes, each one connected to the set of k immediate neighbors.

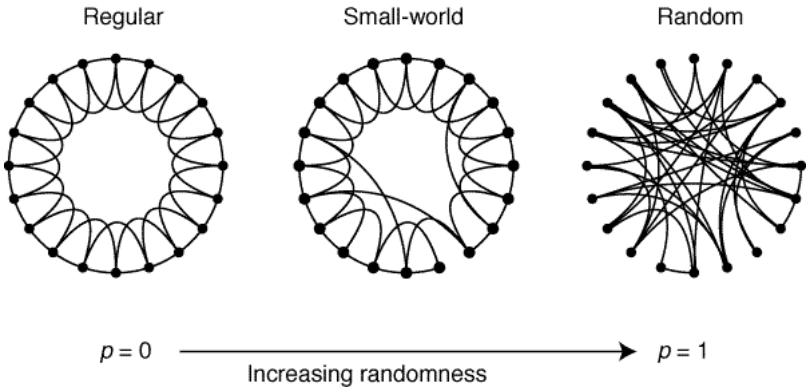


Figure 2.5: Watts-Strogatz model, whereas p increases, the randomness of the network increases. Image from [2].

- With probability p each link is rewired to a randomly chosen node. For a small p the network maintains a high clustering coefficient but the random long-range links can drastically decrease the distances between the nodes.
- For p = 1 all links are rewired, so the network becomes completely random.

With this model, it is possible to achieve a better approximation in what concerns the clustering coefficient observed in real networks but fails to explain the degree distribution.

Scale-Free Networks

With the growth of computational power, many large real networks started to be analyzed, being the World Wide Web a good example [39]. The degree distribution of some real networks can be approximated by the power-law distribution:

$$p_k = k^{-\gamma} \tag{2.10}$$

These networks are called scale-free networks. A power-law degree distribution indicates that most of the nodes of the network have a small degree and only a few have a high degree, these are denoted by *hubs*. This was observed in some complex networks as science collaboration or protein interaction networks [1]. Hubs play a key role in the dynamics of systems. Then, scale-free networks started to play a fundamental role in the study of complex systems, although power-law distributions being rare [1, 40] and may only be observed in large networks [41, 42].

A.L. Barabasi and R. Albert created a model to generate scale-free networks, denoted by B-A model [43]. This model is composed of two steps:

- *Growth* - at each timestep, we add a new node with m links that connect with the nodes already in the network.

- *Preferential Attachment* - the connections between the nodes are probabilistic, depending on the degree of the nodes, making older nodes having a high degree, creating hubs.

Statistical Analysis of Network Properties

The study of networks often involves the analysis of structural properties. Network motifs [44] is a good example of a structural property often studied in Network Science. To verify the occurrence/presence of this type of properties in networks, it is common to use statistical analysis. In this statistical analysis, null models are used as a comparison term. Thus, it is possible to confirm the presence or absence of properties in the original networks. One commonly used statistical measure is the *z-score*, defined as:

$$z - score = \frac{f_{original} - \langle f_{random} \rangle}{\sigma(f_{random})} \tag{2.11}$$

The $f_{original}$ is the frequency or number of occurrences of a given property in the original network, $\langle f_{random} \rangle$ is the average of the frequency of the property in the random model (it can be a set of random networks) and σ is the standard deviation.

2.1.5 Multilayer Networks

The basic representation by graphs is the most common and simple way to portray complex systems. Despite its simplicity, it has been extremely successful. However, with the evolution of research in complex systems, it became necessary to study systems that are increasingly complex but closer to reality. Therefore, it has become essential to go beyond the simple representation by graphs. For example, edges may have heterogeneous characteristics such as connection type, values/strengths, or be active only at certain times. These restrictions lead to the emergence of a new approach, the representation of systems by a multilayer network [3, 45, 46]. In this scenario, we consider layers in addition to nodes and edges, Figure 2.6.

In a multilayer network, each layer can be associated with an *aspect*. This aspect can be the type of links or an instant of time. In this way, we can build a network in which all edges of different types are embedded in different layers of interactions. 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 a multilayer network with links from M different types, to represent the connections at a layer α, with $\alpha = 1, ..., M$, we use an adjacency matrix $A^{[\alpha]} = \{a_{ij}^{[\alpha]}\}$. The entry $a_{ij}^{[\alpha]}$ for $i, j = 1, ..., N$ is 1 in case

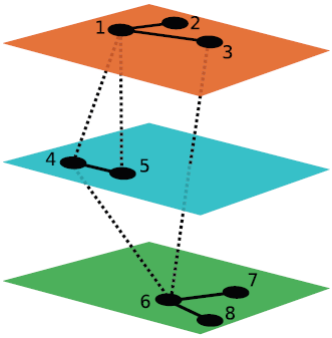


Figure 2.6: Multilayer network with three layers. The intra-layer edges are represented by solid lines and the inter-layer edges by the dotted lines. Figure obtained from [3].

of existence of a link between i and j at layer α and 0 otherwise. Therefore, all intra-layer connection of a multilayer are a set of M adjacency matrices $\mathcal{A} = \{A^{[1]}, A^{[2]}, ..., A^{[M]}\}$, with $\mathcal{A} \in \mathbb{R}_{\geq 0}^{M \times N \times N}$ [47]. This representation can be extended for weighted multilayer networks $\mathcal{W} = \{W^{[1]}, W^{[2]}, ..., W^{[M]}\}$ [47, 48].

For each node i, the connections between different layers α and β, is given by a $M \times M$ matrix $C_i = \{c_i^{[\alpha\beta]}\}$. The entry $c_i^{[\alpha\beta]}$ for $\alpha, \beta = 1, 2, ..., M$ is 1 or 0, depending on whether or not it is possible to go from layer α to layer β through node i. So, the inter-layer connections are represented using a set of N matrices, $\mathcal{C} = \{C_1, C_2, ..., C_N\}$, with $\mathcal{C} \in \mathbb{R}_{\geq 0}^{N \times M \times M}$. Thus, the multilayer network $\mathcal{M}$ can be represented by a tuple with the set of intra-layer connections $\mathcal{A}$ and the set of inter-layer connections $\mathcal{C}$:

$$\mathcal{M} \equiv (\mathcal{A}, \mathcal{C}) \tag{2.12}$$

The connections between two different layers α and β can be different, depending on the type of multilayer network we want to represent. Therefore, a multilayer network can have a different representation from 2.12. We also can represent a multilayer network with a tensor $\mathcal{T} = \{\tau_{ij}^{[\alpha\beta]}\}$ [49]. The entry $\tau_{ij}^{[\alpha\beta]}$ is 1 if there is a connection between node i of layer α and node j of layer β and 0 otherwise. If the multilayer network is a weighted network, the value of the entry $\tau_{ij}^{[\alpha\beta]}$ is the weight of the edge. Therefore, we have a second representation for multilayer networks:

$$\mathcal{M} \equiv \mathcal{T} \tag{2.13}$$

Properties

The study of nodes is the main focus of network investigations, where scientists are always interested in the study of the entities' properties. In this section, we introduce and review a set of local and global properties for the nodes in a multilayer network.

In a multilayer network, not all nodes have connections at all layers. As a consequence, a node i is defined as active on a layer α if has at least one connection with another node at the same layer. The

node-activity vector is defined as:

$$b_i = \{b_i^{[1]}, ..., b_i^{[M]}\}, \tag{2.14}$$

$b_i^{[\alpha]}$ = 1 if node i is active on layer α and $b_i^{[\alpha]}$ = 0 otherwise. The total activity of a node i represents the number of layers in which the node is active, and is defined as $B_i = \sum_{\alpha=1}^{M} b_i^{\alpha}$, $0 \leq B_i \leq M$ [50]. The number of active nodes at layer α is represented by $N^{[\alpha]}$.

The degree of a node in a multilayer network is described by the vector

$$k_i = \{k_i^{[1]}, ..., k_i^{[M]}\}, \tag{2.15}$$

where $k_i^{[\alpha]} = \sum_{i \neq j} a_{ij}^{\alpha}$ is the number of edges incident in node i at layer α. The *overlapping degree* [47] of a node i is the total number of connections of the node in the whole multilayer network and is characterized by

$$o_i = \sum_{\alpha=1}^{M} k_i^{\alpha}. \tag{2.16}$$

The *participant coefficient* [47] measures the heterogeneity of the number of neighbors of a node i across the layers

$$P_i = \frac{M}{M-1}[1 - \sum_{\alpha=1}^{M}(\frac{k_i^{\alpha}}{o_i})^2], \tag{2.17}$$

P_i = 1 when the number of neighbors of node i is equal across the layers and P_i = 0 when a node is active in just one layer.

In a multilayer network, we can study the clustering coefficient separately for each layer, however, this tells us very little about the interplay between the several layers in terms of clustering. A *2-triangle* is a triangle that is formed by an edge belonging to one layer and the two other edges belonging to a second layer, a *3-triangle* is a triangle which three edges are in three different layers. A *1-triad* centered at node i ($j - i - k$), is a triad in which both edges $j - i$ and $i - k$ are on the same layer. Also, a *2-triad*, is a triad in which both edges are in two different layers. Given the previous definitions, we can give two definitions of clustering coefficient for multilayer networks from [51]. For each node i, the first coefficient $C_{i,1}$ is defined as the ratio between the number of 2-triangles with a vertex in i and the number of 1-triads centered in i, we can express this clustering coefficient as:

$$C_{i,1} = \frac{\sum_{\alpha} \sum_{\alpha' \neq \alpha} \sum_{j \neq i, m \neq i} (a_{ij}^{[\alpha]} a_{jm}^{[\alpha']} a_{mi}^{[\alpha]})}{(M-1) \sum_{\alpha} k_i^{[\alpha]} (k_i^{[\alpha]} - 1)}. \tag{2.18}$$

The second clustering coefficient is described as the ratio between the number of 3-triangles having

node i as vertex and the number of 2-triads centered in i. Therefore, we have the equation:

$$C_{i,2} = \frac{\sum_{\alpha} \sum_{\alpha' \neq \alpha} \sum_{\alpha'' \neq \alpha,\alpha'} \sum_{j \neq i, m \neq i} (a_{ij}^{[\alpha]} a_{jm}^{[\alpha'']} a_{mi}^{[\alpha']})}{(M-2) \sum_{\alpha} \sum_{\alpha' \neq \alpha} \sum_{j \neq i, m \neq i} (a_{ij}^{[\alpha]} a_{mi}^{[\alpha']})}. \tag{2.19}$$

Both definitions of clustering above are expressed in terms of adjacency matrix formalism (2.12). For the tonsorial formalism, in 2.13, similar definition can be provided, as in [52].

The *reachability* is an important aspect in the study of graphs. In a multilayer network, the node *interdependence* introduced in [53], is defined as:

$$\lambda_i = \frac{1}{N-1} \sum_{j \neq i} \frac{\psi_{ij}}{\sigma_{ij}}, \tag{2.20}$$

here σ_{ij} is the total number of shortest paths between i and j in the whole multilayer and ψ_{ij} is the number of shortest paths between i and j which make use of links in two or more layers. The network interdependence is given by $\lambda = (\frac{1}{N}) \sum_i \lambda_i$.

A measure of centrality, like the eigenvector centrality, can be calculated by layer and we can build a vector for the different values of centrality:

$$E_i = \{E_i^{[1]}, ..., E_i^{[M]}\}, \tag{2.21}$$

where E_i^{α} is the centrality of node i at layer α.

Several approaches were proposed to define the centrality of a node in multilayer networks, a detailed description of these approaches can be found at [47, 54, 55].

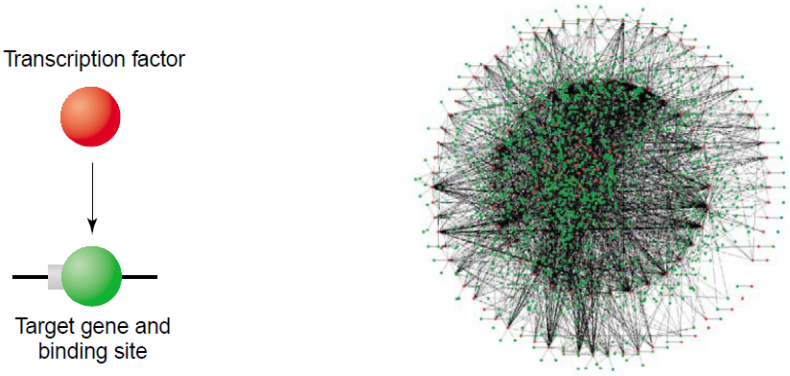
2.2 Biology Concepts

2.2.1 Transcriptional regulatory networks

The transcriptional regulatory networks, Figure 2.7(b), are responsible for the biological process of gene regulation that controls the 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 in these networks can be:

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

The regulatory association between two genes, Figure 2.7(a), can represent the activation or inhibition of the expression of the target gene by the transcription factor.



(a) Regulatory Association

(b) Transcriptional Regulatory Network

Figure 2.7: Figure 2.7(a) represents a simple regulatory association between a transcription factor and a target gene. Figure 2.7(b) illustrates an example of a transcriptional regulatory network. Images obtained from [4].

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, and the edges of E are the directed connections between transcription factors and target genes. Nodes can be transcription factors and target genes at the same time. 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. There is the case when we do not know if it is an activator or repressor, thus in this case $w_{i,j} = 0$. 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.

2.2.2 Gene Ontology

The Gene Ontology (GO)[1] [56] is the most comprehensive resource about the functions of genes and is also the one that is mostly used to support modern biological research. All the functional knowledge present in GO is structured and presented in a way which allows it to be easily used for computational analysis. Gene Ontology is subdivided 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

[1] http://geneontology.org/

- **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

Contents

In this section, we present the background of community detection and cross-species analysis. In the field of community detection, we review the proposed methods to find communities. Moreover, we discuss some methods to evaluate the significance of a community in a network and we close by citing some works regarding community detection in biological networks. Moving to cross-species analysis, we review some studies on the subject in order to understand how the comparison between species can help us to infer some information about previously non-characterized organisms. Finally, we cite some works involving multilayer network approaches in biological networks.

3.1 Community Detection

Since its origin, graph theory has been extremely useful to represent and study a wide variety of systems from different areas. Biological, social or information networks can be represented by graphs, among others. The analysis of these graphs is fundamental for understanding the systems they portray. Structural analysis of a graph often involves the study of the community structure property introduced by Girvan and Newman [33]. Also called clusters or modules, these are groups of nodes that probably share common properties or play similar roles within the network. The identification of modules and their boundaries helps us to classify the nodes concerning their structural importance within communities. Moreover, the detection and subsequent classification of communities also allow us to inspect the connections and similarities between them. In biological networks, communities can represent specific functions of an organism. This way, community detection in these networks is essential to study the organization of the organisms in different functional components.

The study of communities is an important area of research and has received considerable attention from the scientific field. It has become the most studied property with regard to the structure of a network. Thus, there are several methods and techniques developed in the field of Network Science whose objective is to solve the problem of community detection. Next, we consider some of these methods, taking into account their complexity and also the type of graphs they are developed for. These methods are displayed in Table 3.1. Furthermore, we discuss the significance/importance of a community within a network and the similarity that may exist between communities. Lastly, we introduce some works about community detection in biological networks.

3.1.1 Divisive Algorithms

The strategy of divisive algorithms is to detect inter-community edges. Then, by removing these, it is possible to disconnect communities from each other. The crucial point of this type of algorithm is to find a property/measure that allows the identification of those edges.

The algorithm proposed by Girvan and Newman [33, 34] known as the Girvan-Newman algorithm

Algorithm	Directed	Undirected	Weighted	Unweighted	Signed	Overlapping	Multilayer	Code Availability
Girvan-Newman [33, 34]	x	x		x				NetworkX
Radicchi [57]		x		x				Not found
Fortunato-Latora-Marchiori [58]		x		x				Not found
Newman [59]		x	x	x				code
Clauset-Newman-Moore [60]	x	x	x	x				NetworkX
Louvain [61]		x	x	x			x	NetworkX
Leiden [62]		x	x	x				cdlib
Guimerà-Amaral [63]	x	x	x	x			x	Not found
Donetti-Muñoz [64]		x		x				Not found
Capocci-Servedio-Colaiori-Caldarelli [65]	x	x	x	x				Not found
ICS [66]		x		x				Not found
Infomap [67]	x	x	x	x			x	iGraph
Label Propagation [68]		x		x			x	NetworkX
Markov Clustering Algorithm [69]		x		x				cdlib
CFinder [70]						x		NetworkX
Baumes [71]						x		Not found
Link Clustering [72]						x		cdlib
CONGA [73]						x		cdlib, Gregory
Peacock [73]						x		Gregory
Anchuri [74]					x			Not found
Bonchi [75]					x			Not found
Cucuringu [76]					x			SigNet
Esmailian-Jalili [77]					x			Not found

Table 3.1: Community detection algorithms.

(GN), is one of the first proposed algorithms and also one of the most popular. The authors focused on the betweenness centrality as the metric to identify boundaries between communities. The algorithm consists of two steps: (1) Computing the edge betweenness for all edges; (2) Removal of the edges with the largest values. These steps are repeated until all edges are removed. Being a hierarchical clustering technique [78], the resulting partitions are represented in a dendrogram in which each leaf is a node. Each cut in the dendrogram represents a partition of the network in modules, where usually, the partition with the highest modularity value is selected. The main disadvantage of this algorithm is that it can be considerably slower. In sparse graphs, the algorithm has a complexity of $O(N^3)$.

Radicchi *et al.* [57] created a new method, the Radicchi algorithm, proposing a new measure, the *edge clustering coefficient*. This measure generalizes to edges the idea of clustering coefficient introduced by Watts and Strogatz [2]. After computing the values of the clustering coefficient for the edges, the edges removed are the ones with the lowest values for the measure, likely to correspond to inter-community edges. The time complexity for this algorithm is $O(L^4/N^2)$, or $O(N^2)$ on sparse graphs. The method implemented by Radicchi was extended for weighted networks by Castellano *et al.* [79] and for bipartite graphs, by Zhang *et al.* [80].

Another algorithm based on a different centrality measure was proposed by Fortunato *et al.* [58]. The Fortunato-Latora-Marchiori method uses the *information centrality* measure, which is based on the concept of efficiency introduced by Latora and Marchiori [81]. The efficiency of a graph estimates how easy it is for information to travel within the network and it is defined as the average of the inverse of all shortest paths in a network. The information centrality of an edge is the relative variance of the efficiency of a graph when the edge is removed. Therefore, in the algorithm, the authors remove the edges with larger centrality values. Computing the information centrality for an edge requires the calculation of the

distances between every pair of vertices. This can be done using a faster method to perform a breadth-first search (BFS) that takes time $O(LN)$ [82], $O(L^2N)$ for all edges. Since the process is repeated until all edges are removed, the total complexity of the procedure is $O(L^3N)$, $O(N^4)$ on sparse graphs. Since it is also a hierarchical algorithm, the authors select the partition with the largest value of modularity as the final result.

3.1.2 Modularity Optimization Algorithms

The modularity measure has become an essential element for a wide variety of clustering techniques. Nowadays, modularity maximization is the most popular class of methods for community detection. A high value for the modularity indicates a good partition, so, the partition corresponding to the maximum modularity in a network should be the best or at least a very good one.

The first modularity-based algorithm is the greedy method by Newman [59]. The algorithm starts with N clusters, each one containing a node. The edges are added one by one and the addition of a new edge may or not form a new community. In the beginning, the first edge will merge two nodes forming the first module. The edges are selected in a way that the new partition gives the maximum increase or minimum decrease of modularity concerning the previous configuration. The insertion of an edge that is already inside a community does not change the partition and the value of the modularity stays the same. The final result for the algorithm is the partition that maximizes the value of modularity. The running time of this algorithm is $O((L+N)N)$, or (N^2) on sparse graphs. Clauset *et al.* [60] proposed a fast implementation of the technique proposed by Newman, the Clauset-Newman-Moore algorithm. Using more efficient data structures to perform more efficiently the operations of the technique, this approach reaches a complexity of $O(Nlog^2N)$ on sparse graphs.

Introduced by Blondel *et al.* [61], the Louvain algorithm is considered one of the best-known modularity-based algorithms. Initially, all nodes of the graph are a different community. In the first phase, for each vertex i, is computed the gain in the modularity from putting i in the community of its neighbor j, then, i is moved to the community of the neighbour that yields to the largest increase of modularity. In the second phase, the nodes of each community are aggregated in a unique new node representing the community. The two previous steps are repeated until it is not possible to optimize the modularity value. This algorithm achieves a computational complexity of $O(NlogN)$. The Leiden algorithm [62], is a variation of the Louvain algorithm. It tries to solve the problem of the Louvain algorithm which tends to discover communities that are internally disconnected. Therefore, the Leiden algorithm has a second phase in which the communities from the first phase are refined, where these may split into new communities. This algorithm guarantees that the resulting communities are well connected.

Simulated annealing introduced by Kirkpatrick *et al.* [83] is a probabilistic method for global optimization. It consists in performing an exploration of the possible states to optimize a certain function F.

Guimerà-Amaral algorithm [63] is based on this technique and is composed of two types of moves: local moves, where a vertex is moved from one community to another; global moves, which consists of the merge or split of communities. This method can get very close to the true value of maximum modularity, but it has the disadvantage of being very slow. The complexity cannot be estimated as it depends on the parameters chosen for the modularity optimization.

3.1.3 Spectral Algorithms

The spectral properties of graph matrices can be used to find partitions. Spectral clustering uses the eigenvectors of the Laplacian matrices to transform the nodes of a graph in a set of points in some metric space, where the coordinates of the points are the eigenvalues of the vectors. Then, the set of points are clustered using a standard technique such a k-means clustering [84]. The Donetti-Muñoz algorithm [64] is an example of a spectral clustering algorithm. Since the values of the eigenvector components are close for nodes in the same community, the authors used them as coordinates to turn nodes into points in a metric space. Therefore, the nodes can be displayed in a M-dimensional space using M eigenvectors. To group the nodes, the authors use a hierarchical clustering method [78] in which only communities with at least one edge connecting them can be merged. For the similarity measure between vertices, it is used Euclidean and angle distance. The final partition for the graph is the partition of the dendrogram resulting from the hierarchical clustering that has a larger modularity value.

Capocci *et al.* [65] proposed another spectral algorithm, the Capocci-Servedio-Colaiori-Caldarelli method. Here, the authors use the eigenvector components of the right stochastic matrix R. This matrix is obtained from the adjacency matrix by dividing each row by the sum of its elements. If a graph has a number n of connected components, the largest n eigenvalues are equal to 1, with the eigenvectors having equal values for vertices in the same component. Therefore, the communities can be found by inspecting the components of the eigenvectors with an eigenvalue of 1.

We close this section by describing the ICS algorithm by Yang and Liu [66]. In the first step of the algorithm, the adjacency matrix of the network is put in the block-diagonal form, for this, it is computed the clustering centrality for the nodes. This measure is similar to the eigenvector centrality introduced by Bonacich [85], which is given by the eigenvector corresponding to the largest eigenvalue in the adjacency matrix. The value of the centrality is similar to nodes in the same cluster. Therefore, it is possible to see the blocks by listing the nodes in non-decreasing order of their centrality. The cluster found at some step is divided in two if the resulting components are communities, otherwise, the algorithm ends. For the authors, a community is a subgraph such that the external degree of each vertex is bigger than the internal degree.

3.1.4 Alternative Algorithms

In this section, we introduce some alternative algorithms that do not belong in the previous categories.

Infomap is an algorithm developed by Rosvall and Bergstrom [67]. In this approach, the goal is to optimally compress the information needed to describe the dynamic process of information diffusion across the graph. The process of information diffusion in this case is a random walk. The optimal compression is achieved by optimizing the Minimum Description Length quality function [86, 87]. For the optimization, it is used a greedy search combined with the technique of simulate annealing [83].

Another alternative method is the Label Propagation algorithm introduced by Raghavan *et al.* [68]. Each node is initialized with a unique community label. At each iteration, is performed a sweep over all nodes in which each vertex takes the label shared by the majority of its neighbors. In the case of a tie, one of the majority labels is picked at random. Some labels will disappear and others will be propagated through the graph. The process reaches convergence when each node has the majority label of its neighbors.

Lastly, the Markov Cluster algorithm by Van Dongen [69]. The method consists of a simulation of the process of flow diffusion in a network. It starts with the transfer matrix $\mathcal{T}$ of the graph. Then, the algorithm is divided into two steps. In the first step, called expansion, the transfer matrix is raised to some integer power p, generating matrix $\mathcal{M}$. The second step, called inflation, lies in the raising of each entry of the matrix $\mathcal{M}$ to some power α. Next, the elements of each column are divided by their sum, generating the new transfer matrix. The two steps are repeated until the transfer matrix remains equal. The graph described by the final matrix is disconnected and each one of the components represents a module of the network.

3.1.5 Algorithms for Overlapping Communities

The algorithms discussed previously are designed for community detection in partitions where each vertex is assigned to a single community. However, in some networks, each vertex may be shared between communities. The issue of detecting overlapping communities has become quite popular in the last two decades. The most famous technique in respect to this problem is the Clique Percolation method by Palla *et al.* [88]. It is based on the idea that the edges within a community are likely to form cliques. The authors introduced some concepts to implement the idea. Two k-cliques are adjacent if they share $k - 1$ nodes. Therefore, a community is the maximal union of k-cliques that can be reached from each other through a set of adjacent k-cliques. CFinder is the software package implementing the technique developed by Palla *et al.* [70].

Baumes *et al.* [71] proposed two algorithms to find overlapping communities, the Rank Removal (RaRe) and Iterative Scan (IS). To achieve the best performance, they use the Rank Removal algorithm

improved by the Iterative Scan algorithm. Here, a community is seen as a subgraph that locally optimizes a given function W, which is related to the link density of the cluster. Different overlapping sets of nodes may be locally optimal, so, nodes can be shared between them. The IS, is a greedy optimization of function W, starting from a random node/edge, nodes are removed or added one by one until is not possible to increase W. The procedure is repeated using another seed that is randomly picked. The algorithm stops when it finds a previously identified community. RaRe consists of removing important vertices until the graph is fragmented into components of a given size that represents the cores of the clusters. Then, the removed vertices are added to the graph and are associated with those clusters for which the addition increases the value of W. The complexity of the whole process of combining the two algorithms is $O(N^2)$ in sparse graphs.

To facilitate the identification of overlapping communities, some authors proposed to look into communities as sets of edges rather than nodes. One example of this case is the Link Clustering algorithm developed by Ahn *et al.* [72]. It was proposed to group links with a hierarchical clustering technique. The authors used a similarity measure for a pair of adjacent links that is defined by the size of the overlap between the neighbours of the non-coincident end-vertices divided by the total of different neighbors of those end-vertices. The sets of edges are merged pairwise in descending order until all edges are in the same cluster. To select the best partition for the network in the resulting dendrogram, Ahn *et al.* introduced a quality function called *partition density* that measures the edge density within the communities.

Gregory proposed two algorithms in which the original network is transformed into another that can be fed to a clustering algorithm. Then, the disjoint communities found are transformed into potentially overlapping communities of the original network. The first one is called CONGA [89] and is based on the Girvan-Newman algorithm. The algorithms start by calculating the edge betweenness of edges and *split betweenness* of vertices, which is the number of shortest paths that would run between two parts of a vertex if it was split. Then, remove the edge with maximum edge betweenness or split vertex with maximum split betweenness. These two steps are repeated until no edges remain. The complexity of the algorithm is $O(N^3)$ on sparse graphs. The second method is the Peacock algorithm [73]. At first, it is calculated the vertices with highest *split betweenness*. Then, the vertex with the highest value is divided in two with an edge connecting the resulting nodes. This process is repeated until the maximum split betweenness is sufficiently small. In both algorithms, it is performed a community detection algorithm in the transformed network. The overlapping communities are obtained replacing the original names of the vertices that were split.

3.1.6 Algorithms for Signed Networks

Many complex systems can be modeled as signed networks, that contain both positive and negative relations. For example, in a transcriptional regulatory network, each node denotes a gene. A positive link

denotes a positive relationship (activation of gene expression) and a negative link denotes a negative relationship (inhibition of gene expression). The community detection in signed networks has been under-explored. In this section, we cite some proposed community-finding algorithms for this kind of network. In most of the following approaches, the goal is to find polarized communities, in which the edges within them are positive and the edges that connect different communities are negative.

The spectral algorithms have been used to solve the problem of community detection in several types of networks. Thus, some extensions of these have also emerged for signed networks. Some examples are the works done by Anchuri *et al.* [74], Bonchi *et al.* [75] and Cucuringu *et al.* [76]. An alternative method to the spectral clustering techniques is the Esmailian-Jalili algorithm [77]. The authors introduced a Map Equation for signed networks that is based on the assumption that negative edges increase the probability of staying inside a community. Thus, it tries to minimize the positive links between communities and the negative links within communities.

3.1.7 Communities Evaluation

In this section, we discuss how the significance of community structure can be measured. We also review how we can compare the similarity between two different partitions.

Communities Significance

Given a network, we can use a clustering algorithm to find communities in it. We already saw that the modularity measure evaluates the partition of a network. However, high values of modularity do not necessarily indicate that a graph has an established cluster structure. Random graphs may have partitions with a high value of modularity and this is the result of the randomness of structural properties of those graphs. Therefore, the concept of the significance of a partition is related to its robustness and stability it 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.* [90] introduced a method to test the significance of a partition. It performs a sweep over all edges in which each edge is removed with probability α and replaced by another edge between a pair of vertices (i, j) that is chosen at random with probability $p_{ij} = k_i k_j / 2E$. This perturbation affects only the organization of the vertices, the basic structural properties of the graph are conserved. For a given α, some iterations of the perturbation procedure described previously are performed. After obtaining the final modified graph, the communities are identified in this one. Finally, to obtain the stability of the cluster structure, the partition obtained from the modified graph is compared with the original one.

Also Lancichinetti *et al.* [91] used a random graph with similar properties to the original one to access the statistical significance of the cluster structure. In this case, the authors estimate the significance of

single communities and not of the whole partition. The main idea is to verify how likely a community C is also a community of a random graph with the same degree sequence as the original. This likelihood is called C-score and is calculated using the node w of C, which is the node with the lowest internal degree k_w^{in}, it is called the worst node. Being k'^{in}_w the internal degree of the worst node of a random community in the null model, the C-score is the probability that k'^{in}_w is larger or equal to k_w^{in}. A value of the C-score lower or equal to 5% indicates that the community is significant and not a product of random properties of the network. However, relying only on the worst node to evaluate the whole group can be a very severe criterion which can lead to high values for the C-score. Therefore, Lancichinetti *et al.* developed a new measure called B-score that includes a longer list of nodes for the computation of the statistical significance of a community. The rank of a node i if the probability of finding a node with an internal degree equal or higher to k_i^{in} in the null model given its degree k_i and C. The worst nodes correspond to the highest-ranked nodes. So, the B-score is the minimum probability that the sum of the ranks for the worst t nodes of a random community is lower than the given for community C. If a community is significant with respect to the C-score, is also significant according to the B-score. Despite this, the opposite is not necessarily true, low values of the B-score do not necessarily correspond to small C-scores.

Communities Similarity

Using a community detection algorithm, a network is divided into a set of communities that forms a clustering. Some community detection algorithms may output different clusterings for the same network, such as the Infomap and the Louvain that are stochastic algorithms. Therefore, one may want to analyze the differences between clusterings. To understand and evaluate different clusterings, we need to compare them, there are some measures for this purpose that are based on different principles. We have measures based on counting pairs, that consist in counting pairs of objects that are classified in the same way in both clusterings. Then, there are the measures based on set overlaps, they try to match clusters that have a maximum absolute or relative overlap. Finally, we have the measures based on the mutual information metric. Here, we present two examples of measures based on counting pairs.

Given a set S of n elements and two partitions X and Y, a is the number of pairs that are in the same subset in X and Y, b is the number of pairs that are in different subsets in X and Y, c is the number of pairs that are in the same subset in X but in different subsets in Y and d is the number of pairs that are in different subsets in X but in the same subset in Y.

The Rand Index [92] counts correctly classified pairs of elements and is given by:

$$R(X,Y) = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}} = \frac{2(a+b)}{n(n-1)} \tag{3.1}$$

The Jaccard Index [93] is similar to Rand Index, however, it disregards the pairs of elements that are in different clusters for both clusterings.

$$J(X, Y) = \frac{a}{a + c + d} \tag{3.2}$$

3.1.8 Communities in Biological Networks

The vast amount of information currently available about biological organisms allows us to explore the interactions between proteins, genes, and metabolic processes of these organisms. The graph representation is regularly used in the investigation to study the cellular systems of some organisms such as protein-protein interaction (PPI) networks, gene regulatory networks (GRN), and metabolic networks (MN). The study of communities in biological networks suggests that these are characterized by a modular organization in which the modules are associated with specific functions. From this point of view, a module is an entity composed of several elements that perform a specific task, separable from the functions of the other modules. Therefore, the identification of biological modules is fundamental to uncover the functional organization of these networks. However, the information about the proteins or genes in these networks and their interactions is often incomplete, which makes it sometimes difficult to infer some information about the behavior of modules. In the following paragraphs, we present some of the community detection works developed in the context of biological networks.

Regarding PPI networks, Rives and Galitski [94] studied networks of the yeast species *S. cerevisiae*. The authors studied proteins involved in processes that lead the microorganism to a filamentous form. They detected some modules with a hierarchical clustering technique and conclude that the nodes that mostly interact with members of their community are important proteins. Also, it was concluded that edges between communities represent important points of communication between modules. In other work, Spirin and Mirny [95] identified functional modules that correspond to protein complexes in yeast. Different techniques were used for the detection of modules: clique detection, superparamagnetic clustering [96] and optimization of cluster edge density. They estimated the significance of the clusters by computing the p-value for each one in random networks with the same degree sequence of the original network. From the functional annotations of the genes, it was possible to verify that the modules aggregate proteins with the same or similar biological functions, in many cases, the modules coincide with known protein complexes. Chen and Yuan [97] also found functional modules that contain protein complexes in yeast species, applying a modified Girvan-Newman algorithm. Furthermore, they were able to make predictions of unknown functions of some genes based on the functional cluster they belong to. Farutin *et al.* [98] derived a hierarchical decomposition of PPI networks. They managed to identify modules from different levels, modules at lower levels are the nodes of higher-level modules. The authors found that some higher-level modules in human PPI networks correspond to general biological

concepts such as regulation of gene expression or intercellular communication. Sen *et al.* [99] identified protein modules in yeast species using the eigenvector of the Laplacian matrix. Lewis *et al.* [21] explored the relationship between communities in PPI networks and the biological behaviors. For the detection of communities, it was used the multiresolution approach by Reichardt and Bornholt [100]. Lewis *et al.* concluded that many communities are biologically homogeneous, i.e, the functional similarity between protein pairs inside the community is larger than the functional similarity between all protein pairs of the network.

Metabolic networks have also been investigated with community detection algorithms. Holme *et al.* [101] used a hierarchical clustering method based on the Girvan-Newman algorithm. With this method, it was possible to detect modules with different levels of density. Which demonstrates the hierarchical structure of these networks. Ravasz *et al.* [102], studied the metabolic network of *Escherichia coli*. The authors investigated the relation between topological modules and functional properties of the metabolites belonging to the modules. It was found that substrates of a given small molecule class correspond to well-delimited regions of the metabolic network. Therefore, it was possible to associate the topological modules with some specific functionalities. Ahn *et al.* [72], found link communities in metabolic networks. Using the Gene Ontology it was possible to associate functional terms with the communities.

Wilkinson and Huberman [103] analyzed a gene regulatory network (GRN) to detect groups of related genes. The authors built the network by connecting pairs of genes that are mentioned to be related in biological articles. It was used a modified version of the Girvan-Newman algorithm in which the betweenness centrality is computed considering only the shortest paths of a small subset of all pairs of nodes. The results revealed that the genes belonging to the same cluster turn out to be functionally related to each other. In another study [104], the authors inspected the community structure of a transcriptional co-expression network obtained from breast cancer tissue and non-cancer adjacent breast tissue as a control. Then, analyzed the functions associated with the communities using enrichment analysis. The biological functions associated with the modules were different in cancer tissue and healthy tissue. Breast cancer modules were associated with functions that drive disease, whereas modules in healthy tissues were linked to functions associated with the maintenance of homeostasis. However, they observed that the connectivity patterns formed by the association of gene modules and biological functions are similar in the disease and normal tissue, suggesting that the compartmentalization of functional regulation through gene expression remains intact. Bar-Joseph *et al.* [105] developed a new algorithm to detect gene modules in GRN. In this case, a gene module is defined as a set of genes to which the same set of transcription factors binds. The authors applied the algorithm in *S. cerevisiae* and discovered biologically relevant groups of genes. It was also found that the function of the genes present in the modules was consistent with the functions of their regulators.

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. The large amount of data available on biological networks presents a lot of opportunities to study the evolution and function of organisms. Some examples are the PPI networks and the GRN that are crucial to discover conserved evolutionary structures and diversity among species. 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.

Wiles *et al.* [24] compared PPI networks of five different species (human, mouse, fly, worm, yeast) to predict interologues across species. Interologues are protein-protein interactions conserved between organisms. The authors were able to identify interologues conserved across the fives species. Using the Gene Ontology (GO)[1] [56] to analyze the conserved interactions, it was found that orthologous proteins (proteins with the same specificity in different organisms) are highly over-represented in known protein-protein interactions. Wiles *et al.* developed three confidence scores to measure the quality of protein interactions. Lisa Matthews *et al.* [25] used the protein interaction map of *S. cerevisiae* to predict interologues in *C. elegans*. By performing a BLAST search between pairs of orthologs of the proteins in the interaction map of *S. cerevisiae*, it was possible to identify interologues in *C. elegans*. Most of the conserved interactions between the two species have been identified as being involved in metabolic processes. Sharan *et al.* [106] also compared PPI networks of different species, *C. elegans*, *D. melanogaster*, and *S. cerevisiae*. The authors developed a multiple networks alignment framework to create a network alignment graph. In this graph, each node consists of a group of similar proteins and an edge between two nodes represents conserved protein interactions between the corresponding protein groups. By performing a search over the alignment graph, they were able to identify short linear paths of interacting proteins. These paths represent signal transduction pathways and dense clusters of protein interactions that model protein complexes. Brian Kelley *et al.* [107], like Sharon *et al*, also used a strategy for aligning two PPI networks. Analyzing the global alignment graph, the authors demonstrated that between two distantly related species, *S. cerevisiae* and *H. pylori*, there is a large complement of evolutionarily conserved pathways. Caufield *et al.* [108] used protein-protein interactions across different bacterial species to create a meta-interactome. From the meta-interactome, the authors found that 429 protein interactions are conserved across two or more species. These interactions were used to predict protein functions between the different species. Another cross-species analysis was done by Wang *et al.* [109], using the PPI networks of 11 organisms, predicted the interactome of the *Stegodyphus mimosarum* species. Revealing once again that comparative analysis between species may provide additional information about evolution among species.

[1] http://geneontology.org/

Differences between related entities are generally attributed to gene modifications. Therefore, the characterization of inter-species differences in gene regulation is fundamental for understanding the diversity and evolution of species. Borneman *et al.* [26] used chromatin immunoprecipitation and microarray analysis to detect the transcription factors binding sites of three closely related yeast species. The authors were able to identify three different classes of transcription factor binding events. Those conserved across the three species, those in only two of the species, and the specific binding events located in only one of the species. Most of the target genes present in all species are present in only one or two of the species, revealing a considerable divergence in binding sites across the yeasts. These results reflect the specialization of the organisms, this divergence in the regulation of the species may be responsible for their evolutionary adaptation.

Zhang *et al.* [27] performed another study between species in which they predicted human and plant target genes using RNAhybrid [110]. Then, a cross-species regulatory network was built with the target genes previously mentioned. From the regulatory network, it was possible to extract some modules, these modules were associated with three categories: ion transport, metabolic process, and stress response. The similar functions found between human and plant target genes indicate the existence of a relation between exogenous plant miRNA targets and digestive/urinary organs, these findings point to the utility of cross-species comparison in the study of human regulatory mechanisms. Two recent studies have combined cross-species expression and sequence comparisons to infer gene functions. In the first study, Stuart *et al.* [111] compared correlated patterns of gene expression from humans, fruit flies, worms, and yeast. They started by constructing a list of metagenes, a metagene is defined as a set of genes from several organisms that are considered similar when performing a BLAST search. The authors identified pairs of metagenes whose expression is present in multiple organisms, suggesting that the co-expression of those has been conserved across evolution. Then, they build a gene co-expression network in which a node represents a metagene and a link represents significant co-expression between two metagenes. In this network, Stuart *et al.* identified 12 blocks where the components were highly interconnected. From the information about the metagenes, it was possible to infer the biological processes associated with the components in the blocks whose function was unknown. In the other study, Bergmann *et al.* [112] used a slightly different strategy to Stuart *et al.* where six different species were analyzed. Starting with a set of co-expressed genes, S_a, associated with a particular function in organism a, using a BLAST search, they identified the set of homologous genes in organism b, S_b. From S_b, only a subset of genes ${S'}_b$ was co-expressed, these genes were identified as functional conserved homologous of S_a. Furthermore, ${S'}_b$ turned into ${S''}_b$ by including genes from organism b that were co-expressed with genes in ${S'}_b$ but do not share sequence similarity with the genes in S_a. Introducing an example, the authors started with a set of heat-shock genes from yeast. Then, identified a set of co-regulated genes in *E. coli* and *C. elegans*. Later, they found that existed more co-regulated genes in

C. elegans that were also linked with heat-shock response. However, their orthologs were not annotated in this way in the yeast. Once again, it was possible to infer functionality in non-characterized genes.

Multilayer Approach in Biological Networks

The biological relationships characterized by different networks are in most cases not independent, like gene co-expression or transcription factor networks. Therefore, in certain cases, studying single networks turned out to be insufficient to unveil functional regulatory patterns from the interactions across multiple layers of biological information. In recent years, multilayer networks have played a special role in network theory, in the case of biological networks, allowing us to combine multiple levels of genomic and molecular interaction data. In this section, we present some works in which it is used a multilayer approach to study biological networks.

Zitnik and Leskovec [113] developed an algorithm called OhmNet. This algorithm aims to study proteins in different tissues to understand their features. The authors applied the method in a multilayer of PPI networks of 107 human tissues. In 48 tissues with known tissue-specific cellular functions, OhmNet was able to accurately predict the associated cellular functions, and also generated hypotheses about protein actions in the tissues. Kapadia *et al.* [114] used the OhmNet algorithm to predict features of a multilayer blood cell PPI network. Shinde and Jalan [115] used a multilayer PPI network to study the life stages in *C. elegans*, the proteins occurring in different life cycles were distributed over six layers representing the different life stages. The study of the multilayer revealed crucial differences in each layer and also the presence of varying complexity among them. Another study in PPI was introduced by Zhao *et al.* [116], in which they constructed a multilayer of PPI networks for protein function prediction. The authors were successful in predicting protein functions in *S. cerevisiae*. Liang *et al.* [117] combined human and yeast PPI networks to form a multilayer network for the identification of functional modules of genes. The authors developed a clustering algorithm to identify modules in the multilayer network. They were able to predict functional modules that covered over 90% of the proteins in both organisms.

Cantini *et al.* [28], proposed a multilayer network approach combining different layers of genomic data for the identification of candidate driving genes in cancer. They combined transcription factor co-targeting, microRNA co-targeting, protein-protein interaction, and gene co-expression networks. Next, were applied some community detection algorithms. Using enrichment analysis in the communities found, they identified a set of candidate driver cancer genes. From those genes, some of them were already known oncogenes while others were new. The combined information from the different layers allowed the extraction of information on regulatory patterns and functional roles of different cancer driving genes. Rai *et al.* [29] introduced another multilayer approach to understanding the behavior of cancer cells. The authors investigate seven different types of cancer: breast, oral, ovarian, cervical, lung, colon, and prostate. They created two multilayer networks, the normal and the disease network, each one was

composed of seven layers, one for each type of cancer. From the networks, Rai *et al.* extracted three different sets of proteins: common in all normal cells, common in all cancer cells, and common in normal and disease cells. Yu *et al.* [118] in similar work, constructed a multilayer network of three PPI networks in three different tissues: breast, prostate, and blood. Then, observed the overlapping between the genes under the action of a drug used in the treatment of cancer - trichostatin A (TSA). The authors detected two drug-target modules, identifying gene patterns associated with the emergence of cancer.

More recently, Zheng *et al.* [119] develop a method for identifying the disease driver nodes in multilayer networks. The authors used three disease-related biological multilayer networks to test the algorithm. They discover nodes in the minimal set of driver nodes that could act as drug targets in biological experiments.

4

Identification of Functional Modules

Contents

In this chapter, we begin the study of the transcriptional regulatory networks. First, we introduce the networks of the species by presenting their characteristics. Then, we initialize the study of the networks with the detection of modules where we discuss the results obtained using the different algorithms. Further ahead, we continue with the label assignment process on the modules found. We evaluate the performance of the different algorithms and we close with a comparison of results between species.

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+[1] portal which provides the transcriptional regulatory networks of 10 closely-related yeast species [30]. The characteristics of these networks are presented in Table 4.1.

Network	#Nodes	LC #Nodes	#Edges	#TFs	#TGs	$\langle k_{in} \rangle$	$\langle k_{out} \rangle$	CC	D
S. cerevisiae	6 886	6 886	195 498	220	6 886	28.40	28.40	0.47	4
S. cerevisiae B	6 478	6 478	45 209	176	6 475	6.98	6.98	0.22	5
C. albicans	6 015	6 015	35 687	118	6 015	5.93	5.93	0.28	5
Y. lipolytica	5 288	5 288	9 238	5	5 288	1.75	1.75	0.36	4
C. parapsilosis	3 381	3 381	6 986	11	3 380	2.07	2.07	0.25	4
C. glabrata	2 133	2 129	3 508	40	2 116	1.64	1.64	0.09	6*
C. tropicalis	665	665	698	16	663	1.05	1.05	0.01	5
K. pastoris	561	561	581	4	559	1.04	1.04	0.01	5
K. lactis	111	70	126	10	106	1.14	1.14	0.15	2*
Z. bailii	32	32	31	1	31	0.97	0.97	0.00	2
K. marxianus	4	4	3	1	3	0.75	0.75	0.00	2

Table 4.1: Networks Properties. CC stands for Clustering Coefficient, D for Diameter, and LC for Largest Component. In the Diameter field, a value followed by a * represents the value of the Diameter for the largest component of the graph.

From the characteristics in Table 4.1, it is possible to obtain the first impressions about the constitution of the different networks. 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 the *S. cerevisiae* species, 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 a drastic decrease in the number of edges. However, 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

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

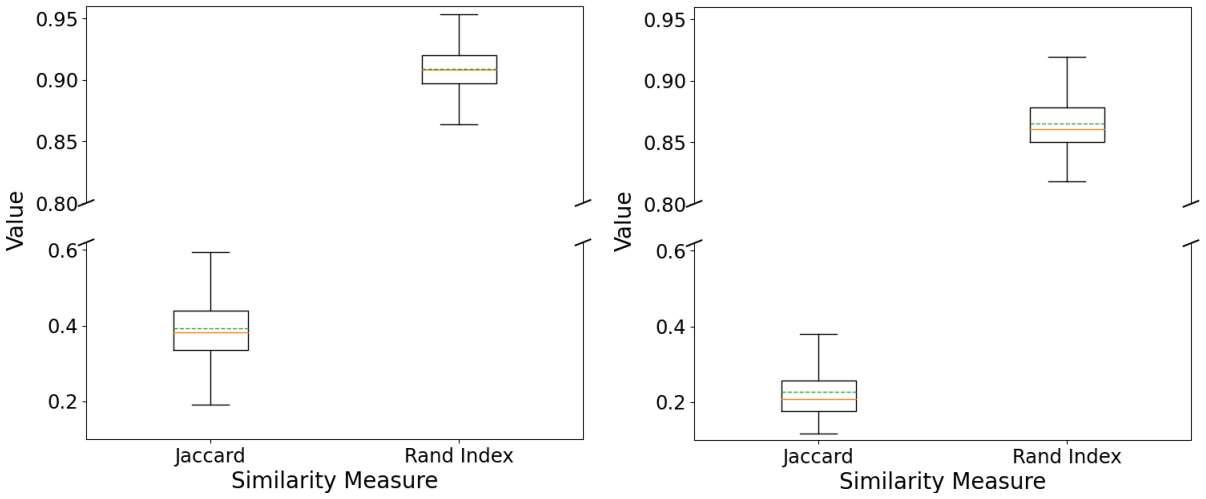
evidence of *S. cerevisiae*. Unlike the species mentioned above, there are species whose networks are small and sparse. Enumerating these species we have: *C. tropicalis*, *K. pastoris*, *K. lactis*, *Z. bailii* and *K. marxianus*. This lack of genetic evidence suggests that the characterization of these species may not reflect their biological nature. Therefore, we decide to discard these networks from the current analysis. The analysis of the rest of the networks shows that the degree of connection differs among the networks. *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. So, 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. The network of the species *C. glabrata* is disconnected, but the largest connected component contains almost all nodes of the entire network. Thus, in the diameter field, we adopt the value of this largest component. In the rest of the document, we will only use the term modules to simplify the reading of the thesis.

4.2 Comparative Analysis of Modules

The first phase of our approach is the detection of modules. In this step, we perform a wide spectrum of modules detection methods in our networks. Once the results are obtained, we evaluate the performance of the different techniques to find out which one has the best performance. In this case, we assume that the best performance is the one that better reflects the series of biological functions of the species. Obviously, we cannot test all modules detection techniques on all networks due to the huge number of necessary analyses. Therefore, we select a collection of algorithms that exploit the diverse ideas and techniques of Network Science developed over the years. The set of chosen algorithms 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 a spectral technique, we hope to verify if the networks contain polarized modules that may be important for understanding their structure. Those methods can be described in Section 3.1. To execute the introduced algorithms, we used libraries where they are already implemented. In Table 3.1 we can consult the library in which each algorithm is available.

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. Therefore, we run these algorithms 1 000 times and compare the results obtained to verify if we can use one of these results in the species analysis. To

study the different partitions, we compare each pair of different partitions having the number of modules equal to the most common result, i.e, equal to the value of the mode. To make this comparison, we use the package *clusim* [120] that allow us to compare different partitions using similarity measures, in our case we use *Jaccard Index* [93] and *Rand Index* [92], both described in Section 3.1.7. To illustrate this process of comparison of partitions, in Figure 4.1 we present a diagram in which we can observe the values of similarity between the different partitions of the two most documented species, *S. cerevisiae B* and *C. albicans*, using the Louvain algorithm. The values obtained for the different similarities are quite different, for the Jaccard Index the values in both networks are considerably low and with high variation. This may be a consequence of comparing two clusters with very different sizes. Comparing a large cluster to a small one decreases the intersection value and increases the union value, this leads to a low value for this measure. The results achieved for the Rand Index show a high value, around 0.9, 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, in our analyzes of the partitions given by stochastic algorithms, we adopt one of the results having the number of modules equal to the mode.



(a) *C. albicans* **(b)** *S. cerevisiae B*

Figure 4.1: Boxplot diagram illustrating the values of similarity for the partitions found with the Louvain algorithm in *C. albicans* 4.1(a) and *S. cerevisiae B* 4.1(b). In the diagram, it is possible to observe the value of the lower and upper limit, first and third quartiles, average (orange), and mean (green).

Due to the temporal complexity of the Girvan-Newman and CFinder method, 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. Some of the algorithms are dependent on a variable k, in CFinder, where k represents the *k*-cliques we pretend to find; for the

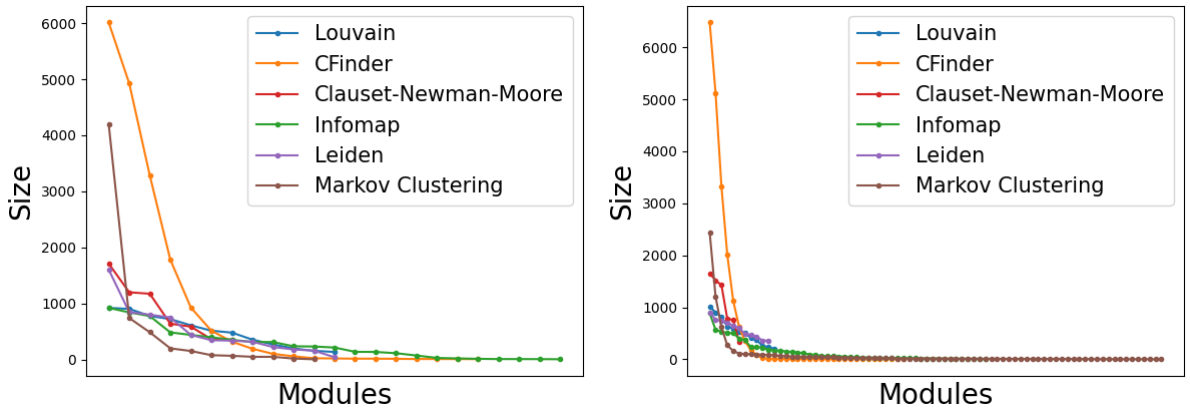
Spectral Clustering, k represents the number of modules we want the algorithm to find. In both cases, we start with k equal to 2. Then, we increment it by 1 until we can not find any more modules. Table 4.2 displays the number of modules obtained for the networks using the different algorithms of our 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 4.2: Number of modules obtained for each network using the different algorithms.

The results in Table 4.2 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 for the biggest networks. Also, the absence of results for the Girvan-Newman and Spectral Clustering in signed networks lead us not to choose to study these results. Louvain and Leiden have similar results for all species. In general, Markov Clustering and Infomap are the algorithms that can find a large number of modules. The CFinder was able to find some modules, which indicates that the study of overlapping modules may help understand these species. The results show that there is a great divergence between the number of modules obtained between *S. cerevisiae* and *S. cerevisiae B*. For this reason, the filter applied to create the *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 into functionalities, we decide on using the filtered network to study the *S. cerevisiae* in the rest of the thesis. In *Y. lipolytica* few modules were detected, as expected since there are a low number of transcription factors in the information we have about this one. This gives us the idea that in the future functional analysis of these modules, few functions should be identified for this species. For the rest of the species under analysis, it was possible to extract some modules. Which indicates that the functional characterization of these may be more complete. To better understand the division in modules, we decided to study the distribution of their sizes for the different algorithms, in Figure 4.2 are these distributions for the species *S. cerevisiae* and for *C. albicans*.

The analysis of the distributions shows different magnitudes between modules of the distinct algorithms. A very large gap between the sizes of the modules can make the classification of modules unbalanced since very large modules may aggregate a lot of functionality and small ones may not be associated with any functionality at all. From there, a balanced division of the networks, in which the modules have sizes of the same magnitude, should be the case that better reflects the division of species according to their biological function. In both cases of Figure 4.2, the modularity-based algorithms (Louvain, Leiden, and Clauset-Newman-Moore) seem to contain a more balanced division, which indicates



(a) *C. albicans* **(b)** *S. cerevisiae*

Figure 4.2: Modules size distribution for *C. albicans* 4.2(a) and *S. cerevisiae* 4.2(b).

that the partitions given by these algorithms may be an important point of study. The Infomap, although it yields some very small modules, others are of a magnitude equivalent to those found with the previously mentioned algorithms. In the field of overlapping modules, CFinder found modules of different orders, from those that include almost the entire network to the smallest. Therefore, in this case, in the analysis of the modules, we considered only the smallest ones, as the larger ones should not be functionally specific and should capture a large amount of behavior. Markov Clustering is the algorithm where the sizes of the modules are most unbalanced, so we decide to discard these results.

To close the first phase of our analysis, we analyze the significance of the modules obtained with the modularity-based algorithms, since these seem to be the best performing algorithms. To test the significance of the modules, we calculate their C-score and B-score. To exhibit an example of the values obtained, in Table 4.3 we list the results for *S. cerevisiae*. According to the C-score values, in none of the algorithms, it was possible to identify significant modules. However, the B-score says the opposite, which indicates that the C-score is a very restrictive measure. In networks with many connections between modules, this one easily reaches a high value. Looking at the B-score values, the Louvain algorithm seems to have only one significant module. This may be a consequence of the stochasticity of the algorithm. Comparing the B-score values of the two other algorithms, both produce significant modules. Despite that, the distribution of the Leiden modules size seen in Figure 4.2 shows that after the first phase of our analysis, this algorithm seems to be the one that best captures the structure of the species. Nevertheless, in the rest of the thesis, we will 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 4.3: c.

4.3 Functional Analysis of Modules

This section refers to the label assignment process. This process consists of assigning one or more labels to each one of the previously found modules. These labels represent specific behaviors/functionalities of the species. Therefore, it allows us to define the modules at a functional level.

Each gene of a species is associated with one or more specific Gene Ontology terms, where each of these terms belongs to one of the three ontologies. For example, a gene can have more than one term, one corresponding to the biological process and another corresponding to the molecular function. The idea behind the labeling process is to calculate the terms of the Gene Ontology present in each module and then associate to these modules the terms that are most represented among their genes. Thus, to perform the labeling process, we must obtain all terms and respective representations for each module. The set of terms associated with a module is composed of the terms directly associated with its genes and by all higher-order terms in the hierarchy of the Gene Ontology that have a relation with the previous ones. To get all terms in a module, we first get the terms directly linked with the genes. Then, we use the relations between terms to go through all the terms in the hierarchy until we reach the root. This way, it is possible to have all terms associated with a module, from the most specific at the bottom of the hierarchy to the most global at the top.

We use the procedure described in the previous paragraph to find all terms associated with each module. Then, to calculate the number of times a term appears in a module, we sum its occurrences in the set of genes that constitute the module. The extraction of terms revealed a large number of terms present in the different modules. To continue with the label assignment process, we obtain the most significant and representative terms of the set of terms of each module. Therefore, we perform a three-step filtering of the terms: (1) select only the most global terms; (2) retain only the most specific terms of the module; (3) retain the terms with a good representation in the module. Each module contains several

terms, from the most specific to the most global. So, in the first step, we select the highest-level terms in the GO hierarchy (levels 2 and 3). To perform the second step of the filtering, we use the statistical measure p-value, which allows us to detect the most specific terms of each module concerning the global network. Additionally, we use different p-value intervals to see the changes in the labeling process as the p-value threshold increases. Lastly, we analyze the degree of representation of each term in the modules. By setting a specific threshold, the terms of the module will be those whose representation is greater than or equal to the threshold. In our work, we decide to consider a threshold of 10%. To be selected, a term must be represented in at least 10% of the genes of the module. This way, we can reveal processes that may be associated with a large number of genes or just a few.

The labeling process was applied to the modules of the different species detected by the algorithms discussed in Section 4.2. In the remainder of this section, first, we compare the performance of the different algorithms in *S. cerevisiae* since it is the species with most genetic evidence. Then, we illustrate the whole label assignment process using the algorithm with the best performance in *S. cerevisiae*. Finally, we close this chapter by evaluating the results of the process on the different species, also using the same algorithm as the previous step.

Algorithms Performance

Using *S. cerevisiae* as a reference, we compare the performance of algorithms that we consider to have interesting results. First, we start by analyzing the modularity-based methods, Figure 4.3. A first look shows that most modules have more than one label, exposing the functional diversity of these. However, not all genes are linked to the functionalities that characterize their modules. 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. By reading the image, it is possible to check the representation of each term in the respective module. This representation is the ratio of genes of the module associated with the term. For some of the modules, the overall representation of the terms exceeds the value 1. This is a consequence of the combination of terms in some genes. Normally, these are related terms, such as the metabolic and cellular processes.

In Figure 4.3, we observe that some functions appear in all classifications and with high representation. Such as the metabolic process, biological regulation, catalytic activity, response to stimulus, among others. This aspect points out the importance that some functions have in the species. In contrast, some terms have a more residual representation. This leads us to believe that these terms represent specific functions and are associated with a smaller set of genes. Reproduction, reproductive process, and transporter activity are good examples of specific functions detected in the modules. In the case of reproduction, this one only emerges in the modules from Leiden and Louvain. A possible explanation for

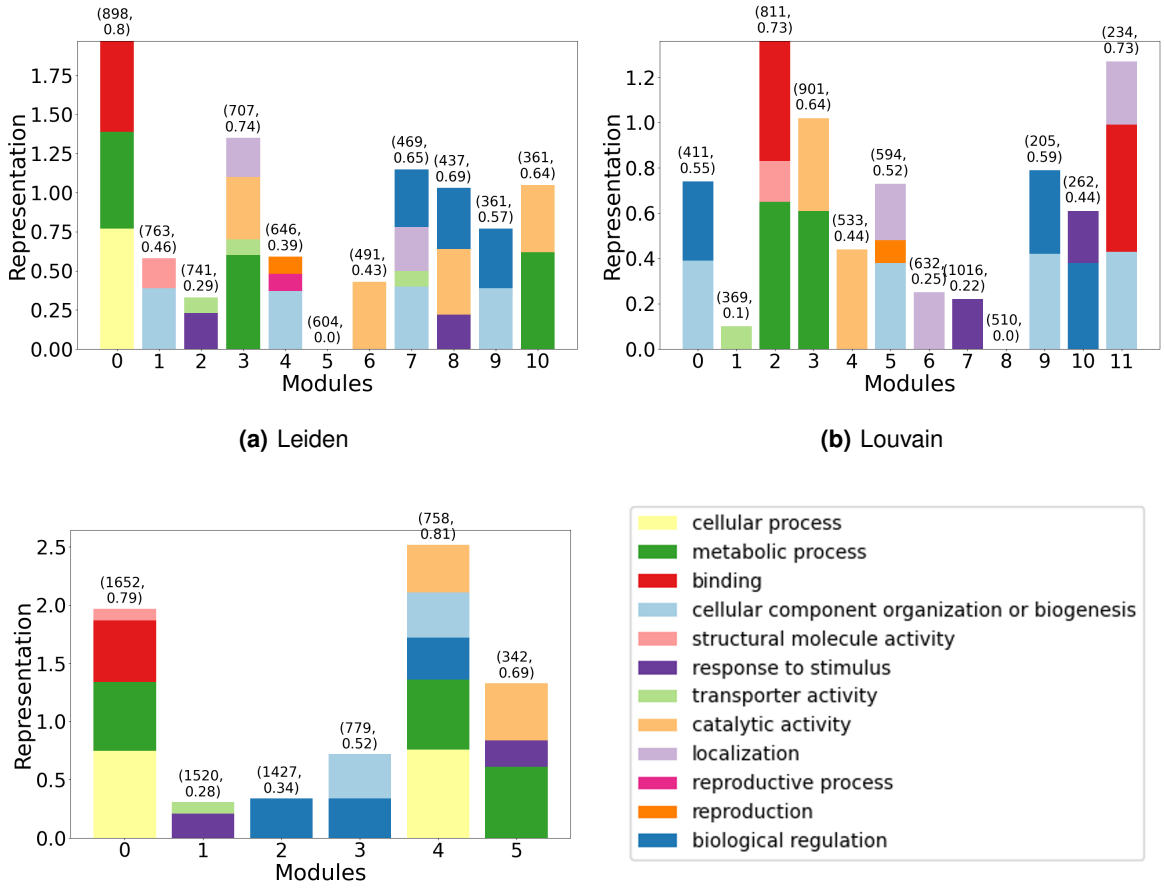


Figure 4.3: Modules and respective functions for modularity-based methods on *S. cerevisiae*. The bar of each term symbolizes its representation in the module. The pair of values at the top of each bar are respectively the size of the module and the percentage of genes of the module related with at least one term (in the module).

this situation is the fact that these terms are associated with a small number of genes, so, their detection is more difficult in modules of larger sizes such as those identified by the Clauset-Newman-Moore algorithm. The Clauset-Newman-Moore is the algorithm with the poorest performance, it fails in the detection of some functionalities such as reproduction, reproductive process, or localization. Furthermore, some modules are associated with terms that only cover a small part of the entire module, such *M1* or *M2*. Comparing the results from Louvain and Leiden we can observe some similarities and differences. Starting with the similarities, it is clear that the classification of some modules is practically identical. Also, in both cases, there is an unclassified module despite its considerable size. We can verify the following similarities between modules from Leiden and Louvain: *M6* with *M4*, *M3* and *M10* with *M3*, *M9* with *M0* and *M9*. Regarding the differences, in the case of Leiden, there are only two modules that have one or no associated terms (*M5* and *M6*) while in Louvain there are five (*M1*, *M4*, *M6*, *M7* and *M8*). The

Leiden algorithm allows us to capture a wide variety of functions in the modules. Furthermore, there are terms detected with Leiden that are not detected in the case of Louvain, such as the cellular process and the reproductive process. This combination of factors leads us to conclude that the classification of the Leiden modules is more diverse and complete. Closing the comparison, we conclude that the Leiden algorithm had a better performance in dividing and capturing the functionalities of the species.

In Section 4.2 we recognized that the results of CFinder and Infomap algorithms seem to be interesting. Here, we analyze the classification of their modules. Beginning with CFinder, in Figure 4.4 we present the outcome of performing the label assignment process in the modules found with this algorithm. We note that some modules are functionally similar such as *M22* and *M26*. This occurs because,

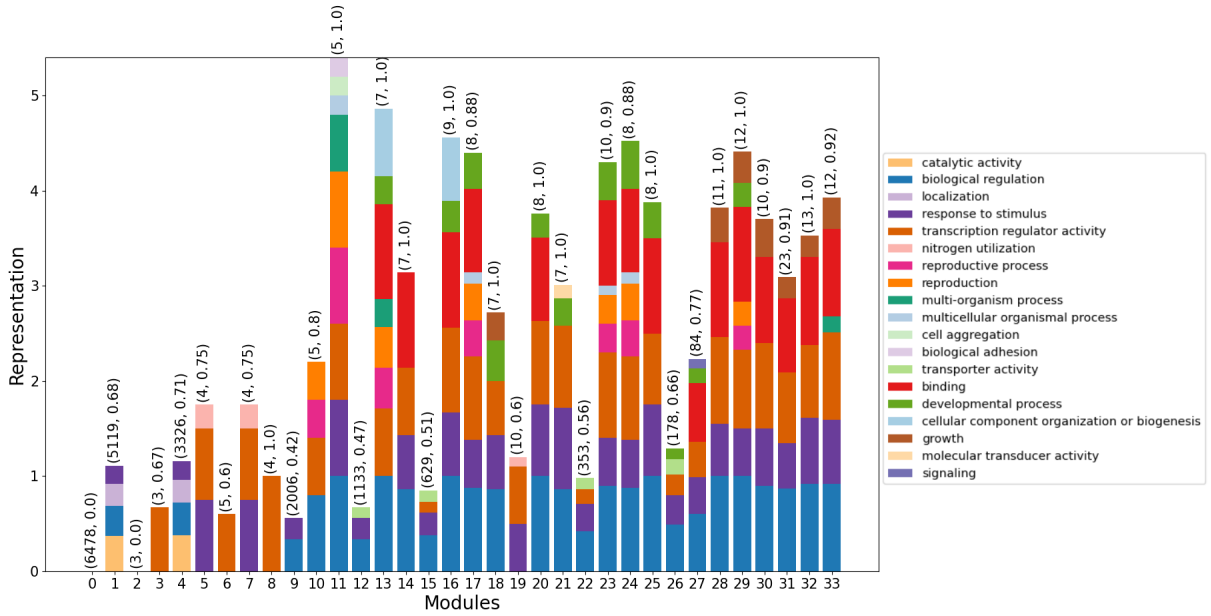


Figure 4.4: Modules of *S. cerevisiae* obtained with CFinder and respective functions.

when k increases in CFinder, it is common to find larger modules that encompass small modules previously found for a smaller k. Therefore, the classification of these modules ends up being similar. We also pay attention to some modules that have a lot of functionality assigned. However, these modules are too small and their classification does not give us a general idea about the functions present in the species. Few modules give us some new information about the species, like *M26* and *M27* which contain functions previously not detected: 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). The results of this module help us to confirm that these are important functions in this organism. About the performance of this algorithm, although it is not as good as the performance of Leiden, it managed to confirm some conclusions previously established.

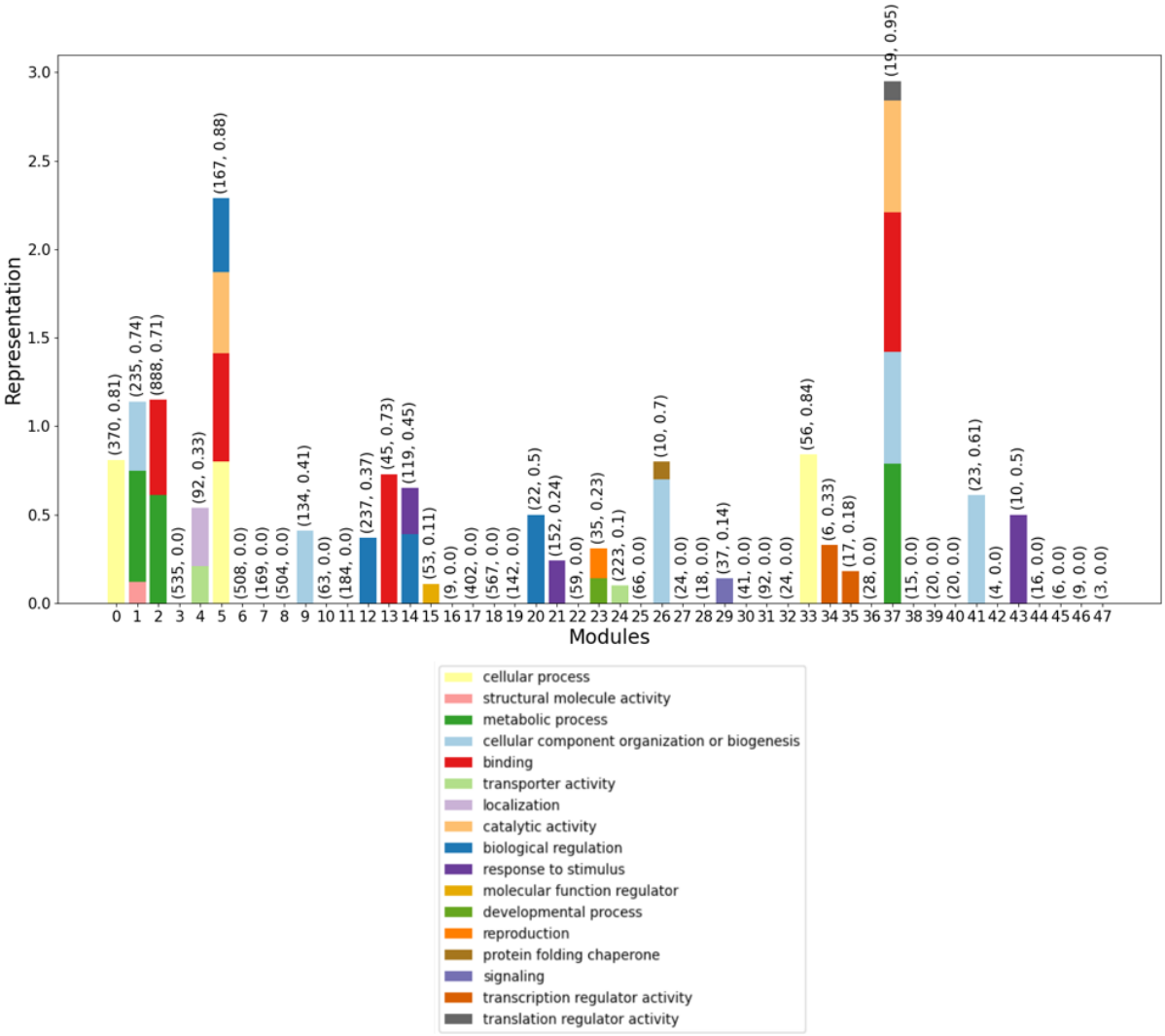


Figure 4.5: Modules of *S. cerevisiae* obtained with Infomap and respective functions.

Finally, we discuss the functional characterization of the modules detected with Infomap. By analyzing Figure 4.5, we find that the performance of Infomap is not the best either. Although it managed to classify some modules of relevant size, it failed to classify the vast majority of modules. Therefore, the algorithm was not able to divide the species into good functional elements.

Functional Analysis on *S. cerevisiae*

To exemplify the label assignment process, we use the modules found by the Leiden algorithm in *S. cerevisiae*. Figure 4.6 displays the number of terms obtained for the modules of *S. cerevisiae*. By examining the figure it is possible to consult the number of terms of levels 2 and 3 for each module, and it is also possible to count the number of terms for each p-value interval.

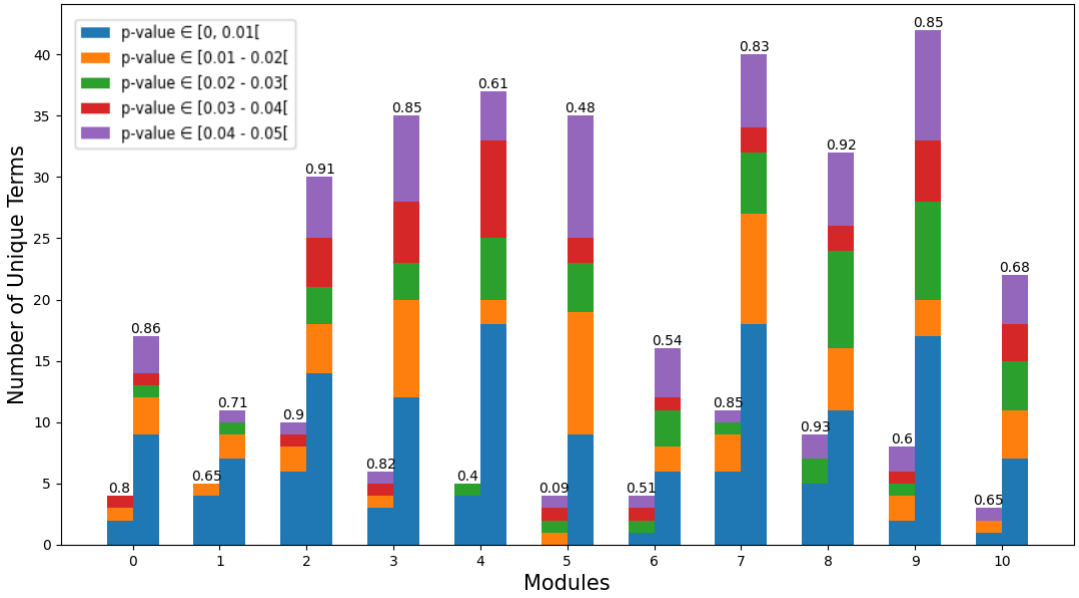


Figure 4.6: Number of Gene Ontology terms associated with the modules of *S. cerevisiae* found by the Leiden algorithm. For each module, the two bars side by side represents the terms of levels 2 and 3, respectively. The value above each column represents the ratio of nodes in the module that are associated with at least one of the terms. Terms with different p-values are identified with different colors as shown in the subtitle.

Analyzing the terms of level 2, we observe that modules have different numbers of associated terms and the ratio of genes of the modules that are associated with the modules is also different. Module 5 is the one in which this ratio is lower and most genes are not associated with the terms. Therefore, it should be difficult to functionally characterize this module. About the remaining modules, almost all of them have a high proportion of nodes associated with the terms, this fact suggests that must be possible to assign functions to them. If we look at the different terms, we see that most are found in the first p-value interval, indicating that these are more representative of the module than the others. However, despite not being so representative, many terms belong to other p-value intervals. This implies that these may complement the functional information about the modules, being essential for the label assignment process. To assign labels to the modules, we decide to look at the level 2 terms associated with the modules. These terms and respective representations are listed in Table 4.4.

As we have already seen, each gene can be associated with terms from different ontologies. However, only the Biological Process and the Molecular Function are essential to classify the modules. The

Term	M0	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Biological Process											
metabolic process (GO:0008152)	0.62			0.60							0.62
cellular process (GO:0009987)	0.77										
cellular component organization or biogenesis (GO:0071840)		0.39			0.37			0.4 0		0.39	
response to stimulus (GO:0050896)			0.23						0.22		
localization (GO:0051179)				0.25				0.28			
reproduction (GO:0000003)					0.11					0.09	
reproductive process (GO:0022414)					0.11					0.09	
biological regulation (GO:0065007)								0.37	0.39	0.38	
detoxification (GO:0098754)			0.01				0.01				
carbon utilization (GO:0015976)			0.01								
developmental process (GO:0032502)					0.09						
multi-organism process (GO:0051704)					0.07						
signaling (GO:0023052)						0.06				0.06	
nitrogen utilization (GO:0019740)						0.01					
cell aggregation (GO:0098743)							0.01				
locomotion (GO:0040011)										0.01	
Molecular Function											
binding (GO:0005488)	0.58										
structural molecule activity (GO:0005198)		0.19									
transporter activity (GO:0005215)			0.10	0.10			0.09	0.10			
catalytic activity (GO:0003824)				0.40			0.43		0.42		0.43
molecular transducer activity (GO:0060089)		0.01									
antioxidant activity (GO:0016209)			0.01								
transcription regulator activity (GO:0140110)				0.06							
protein folding chaperone (GO:0044183)						0.01					
translation regulator activity (GO:0045182)								0.02			
molecular function regulator (GO:0098772)										0.08	
Cellular Component											
organelle part (GO:0044422)		0.53		0.51				0.53	0.51		
protein-containing complex (GO:0032991)		0.39						0.44	0.37		
cell part (GO:0044464)			0.89						0.91		
cell (GO:0005623)			0.89						0.91		
membrane (GO:0016020)			0.35					0.35			
membrane part (GO:0044425)			0.28					0.28			
organelle (GO:0043226)			0.73					0.75	0.79		
membrane-enclosed lumen (GO:0031974)								0.24			
supramolecular complex (GO:0099080)	0.02					0.02			0.02		
extracellular region (GO:0005576)										0.07	0.03

Table 4.4: Gene Ontology Terms of level 2 and respective representation for the modules of *S. cerevisiae* species found with the Leiden algorithm. The color of each cell corresponds to the interval of the p-value of the term.

Cellular Component information only indicates where the genes act, so, it is not necessary to functionally classify the modules. The vast majority of modules contain several associated terms, the high representation of these terms in the modules demonstrates that it is possible to proceed with their classification. This can cause a module to be associated with one or more biological functions. In the third step of the filtering of the terms, we define that terms are significant in a module if their representation in the module is greater than or equal to 10%. Referring to the terms in Table 4.4, we can proceed with the label assignment process. The labeling process for *S. cerevisiae* culminates in the following attributions: *M0* - metabolic process, cellular process and binding; *M1* - cellular component organization or biogenesis and structural molecule activity; *M2* - response to stimulus and transporter activity; *M3* - metabolic process, localization, transporter activity and catalytic activity; *M4* - cellular component organization or biogenesis, reproduction and reproductive process; *M6* - catalytic activity ; *M7* - cellular component organization or biogenesis, localization, biological regulation and transporter activity; *M8* - response to stimulus, biological regulation and catalytic activity; *M9* - cellular component organization or biogenesis

and biological regulation; *M10* - metabolic process and catalytic activity.

Of all the modules of the species, only module 5 could not be classified due to the third step of the filtering. All the others have been associated with functions that characterize their behavior. The terms that were associated with the modules have different levels of representation, such as the term reproduction and the term metabolic process. This contrast reveals that some are global processes, associated with a large number of genes, and others are specific, being affiliated with a small set of elements. By consulting Table 4.4, one can notice that even if the terms appear in different modules, their representation in these modules has a similar value. As the modules are of similar sizes, this indicates that in different modules there are different groups of genes with similar sizes that are associated with the same function. The labeling shows that there are some modules associated with the same functions. To distinguish the behaviors of modules with identical terms, we analyze the terms of level 3 that are sub-processes of level 2 terms, these are listed in Table 4.5.

Term	M0	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Biological Process											
cellular metabolic process (GO:0044237)	0.61										0.59
organic substance metabolic process (GO:0071704)	0.60										0.58
primary metabolic process (GO:0044238)	0.58										
nitrogen compound metabolic process (GO:0006807)	0.56			0.50							
biosynthetic process (GO:0009058)	0.40										
cellular component biogenesis (GO:0044085)		0.21			0.21			0.23			
response to chemical (GO:0042221)			0.13						0.11		
small molecule metabolic process (GO:0044281)				0.15							0.19
cellular component organization (GO:0016043)					0.32			0.33		0.35	
cellular localization (GO:0051641)								0.16			
establishment of localization (GO:0051234)								0.26			
regulation of biological quality (GO:0065008)								0.12			
regulation of biological process (GO:0050789)									0.33	0.32	
cellular response to stimulus (GO:0051716)									0.19		
regulation of molecular function (GO:0065009)										0.11	
oxidation-reduction process (GO:0055114)											0.10
Molecular Function											
hydrolase activity (GO:0016787)				0.18					0.22		
oxidoreductase activity (GO:0016491)							0.11				
transferase activity (GO:0016740)							0.17				0.19
catalytic activity, acting on a protein (GO:0140096)									0.13		

Table 4.5: Gene Ontology terms of level 3 for *S. cerevisiae*.

The introduction of the new results allows us to enrich the attributions captures with the label assignment process started earlier. The previous classification of modules becomes: *M0* - metabolic process (cellular metabolic process, organic substance metabolic process, primary metabolic process, nitrogen compound metabolic process, biosynthetic process), cellular process and binding; *M1* - cellular component organization or biogenesis (cellular component biogenesis) and structural molecule activity; *M2* - response to stimulus (response to chemical) and transporter activity; *M3* - metabolic process (nitrogen compound metabolic process and small molecule metabolic process), localization, transporter activity and catalytic activity (hydrolase activity); *M4* - cellular component organization or biogenesis (cellular component organization and cellular component biogenesis), reproduction and reproductive

process; *M6* - catalytic activity (oxidoreductase activity and transferase activity); *M7* - cellular component organization or biogenesis (cellular component organization and cellular component biogenesis), localization (cellular localization and establishment of localization), biological regulation (regulation of biological quality) and transporter activity; *M8* - response to stimulus (response to chemical and cellular response to stimulus), biological regulation (regulation of biological process) and catalytic activity (hydrolase activity and catalytic activity, acting on a protein); *M9* - cellular component organization or biogenesis (cellular component organization) and biological regulation (regulation of biological process and regulation of molecular function); *M10* - metabolic process (cellular metabolic process, organic substance metabolic process, small molecule metabolic process, oxidation-reduction process) and catalytic activity (transferase activity).

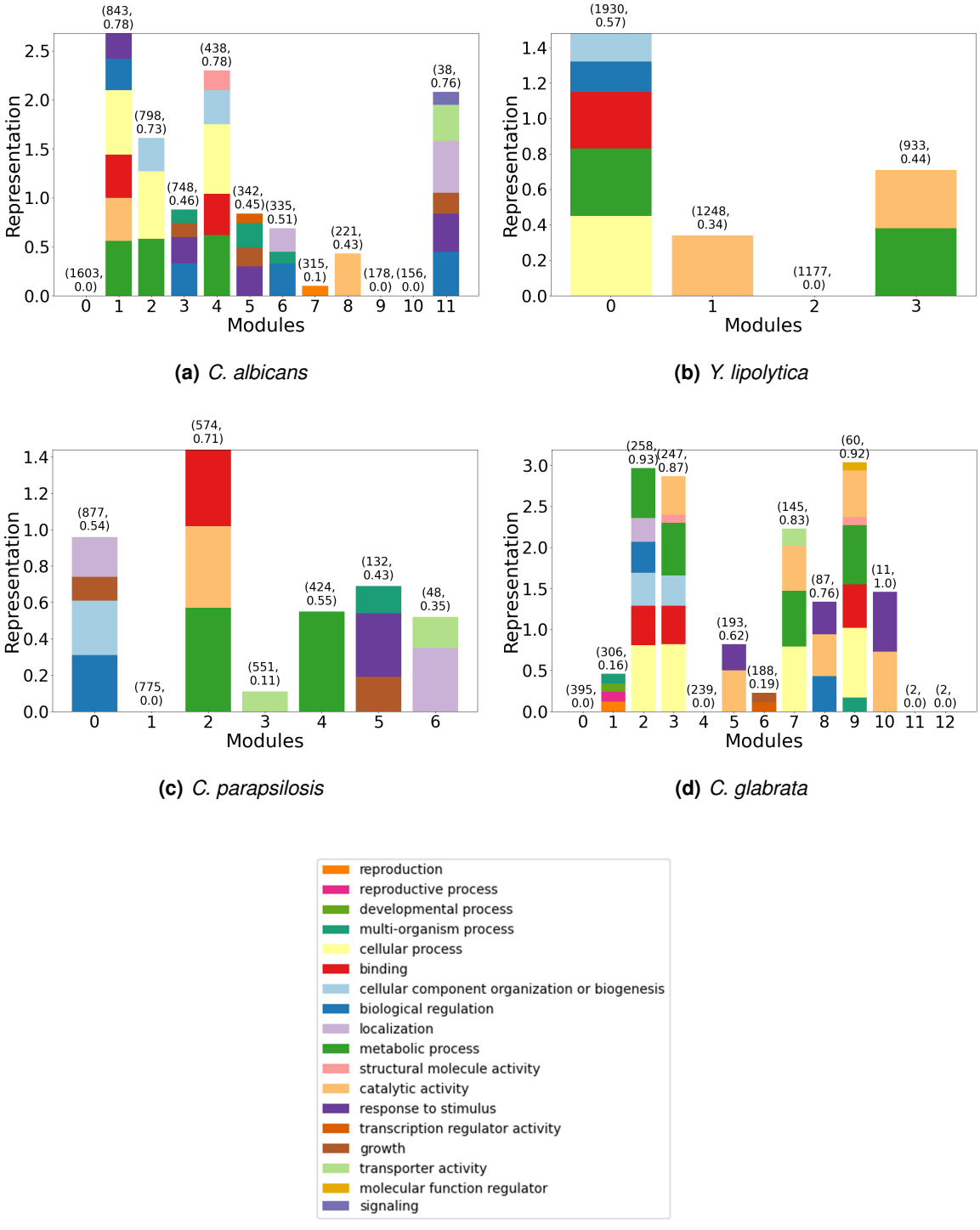
The enrichment of the modules' attributions with more specific labels, allows us to distinguish some of the functionalities between the modules. Regarding the metabolic process, *M0* is the module that has more sub-functionalities and therefore is the most complete in comparison with *M3* and *M10*. The huge number of sub-functionalities of the metabolic process may be the reason for its high representation in the modules. The term cellular component organization and biogenesis are divided into two sub-functionalities, *M4* and *M7* are more complete containing both, while *M1* and *M9* have only one. Modules *M7*, *M8* and *M9* have different behaviors in relation to the biological regulation process. While *M7* is intended for the regulation of biological quality, *M8* and *M9* are related to the regulation of the biological process. However, *M9* is also responsible for regulating the molecular function. Finally, in the catalytic activity, the behaviors of the modules are also distinct from the four sub-functionalities found, each module is associated at most with two, revealing some specificity between them. This deeper study of the level 3 terms allows us to conclude that, although some modules share some general labels, they are in charge of distinct functions.

Functional Analysis of Remaining Species

Closing this chapter, we analyze the results of the label assignment process for the remaining species in the study, these are presented in Figure 4.7. In this analysis, we use the results for the modules obtained with the Leiden algorithm, since it is the algorithm with the best performance for *S. cerevisiae*.

Starting with *C. albicans*, exists an absence of terms in *M0*, *M9*, *M10*. In the first of these modules, a plausible reason for the absence of terms is the fact that this module represents a large fraction of the species, which makes it difficult to detect terms with a good p-value. About *M9* and *M10*, this must be caused by the simple absence of meaningful terms or lack of representation of those terms. All the remaining modules are associated with at least one function. Many of those modules 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*. As with *C. albicans*, also in other species some modules represent a large part of the species and end up not being associated with any biological function. Looking at the functions detected in the different species, we note that also in *C. parapsilosis* and *C. glabrata*, some modules are associated with functions that are not detected in *S. cerevisiae*. These events may have their origin in the size of modules of *S. cerevisiae*. Due to their large sizes, it is difficult for more specific terms to have a good representation in these, as 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. This reveals the central role they play in the functionality of different organisms. By comparing the results between species, it is possible to verify 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 *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*.



(a) *C. albicans*

(b) *Y. lipolytica*

(c) *C. parapsilosis*

(d) *C. glabrata*

(e) Gene Ontology Terms

Figure 4.7: Label Assignment results for the different species using Leiden algorithm.

5

Cross-species Comparison

Contents

In this chapter, we continue the analysis of the species introduced in Section 4.1. After the extraction of the modules and respective functional characterization, our procedure culminates in a comparison between species. Initially, we compare the functional modules discovered in Section 4.3 and we settle some similarities between species. For this, we consider the modules from different species with a high degree of connectivity between them. Finally, we move on to a multilayer network approach where we search for potential functional structures conserved among species. For this, we perform the last modules detection step followed by the functional evaluation of the modules obtained.

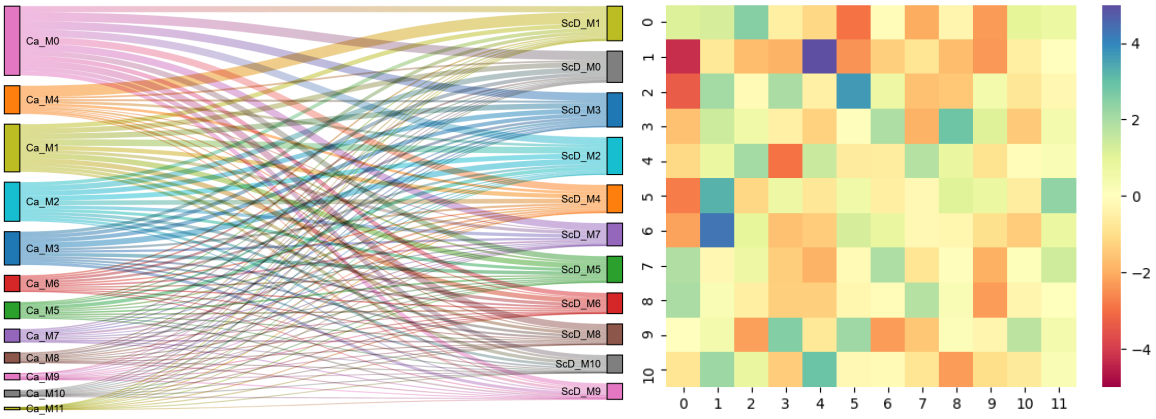
5.1 Functional Comparison of Modules

In this section, we analyze the degree of connection between modules of different species. We resort to the homology mappings between species to establish the connections between modules. Each link in a homology mapping between two species denotes the connection between two genes that have a high degree of similarity. We call these genes, homologous genes. In biology, it is established that the DNA sequence of two homologous genes derives from a common ancestor. They may or may not have the same function. A homology mapping is formed by performing a BLAST search, which allows the comparison between genetic sequences of two different species. For each gene of one species, is picked the gene of the other species with which it is most similar. For this work, the homology mappings are obtained from YEASTRACT+ [30]. We start by illustrating the process of modules comparison between species using the two most documented, *S. cerevisiae* and *C. albicans*. Afterward, we discuss the most relevant connections between species.

S. cerevisiae vs *C. albicans*

Different species are interconnected through pairs of homologous genes. Therefore, also the modules between different species are interconnected. In this subsection, we portray the comparing process between two different species. For this purpose, we explore the level of connection between the functional modules of *S. cerevisiae* and *C. albicans* obtained with the Leiden algorithm. In Figure 5.1(a) we present a Sankey diagram representing the connections between the modules for both species.

The Sankey diagram shows that a module contains connections with multiple modules of the other species. An explanation for this fact is the involvement of a transcription factor in multiple biological functions. We also note that the modules that seem to have a stronger connection are the larger ones. However, *M0* from *C. albicans* was not tagged in the process of label assignment. This suggests that the comparison between species may be crucial to uncover some functionalities in not labeled modules. Although we can draw some conclusions from the Sankey diagram, it is not enough to understand the level of connection between modules. Therefore, we perform an analysis to assess the quality of the



(a) Sankey Diagram **(b)** Heat map

Figure 5.1: Figure 5.1(a) - Sankey diagram representing the connections between the modules of *S. cerevisiae* and *C. albicans*. Figure 5.1(b) - heat map representing the level of connectivity between the modules of *S. cerevisiae* and *C. albicans*.

mapping between modules. 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. This null model 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 5.1(b) we introduce the heat map of the z-scores representing the level of connection between modules, here, we consider the original network against the null model. The heat map reveals the existence of some pairs of modules with strong connections in relation to others (green and blue colors). To a better visualization of the z-score values, we present those in Table 5.1.

	M0	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**	**M11**
M0	1.17	1.25	2.52	-0.54	-1.14	-2.90	-0.10	-2.00	-0.35	-2.27	0.98	0.72
M1	-4.25	-0.71	-1.71	-1.91	20.36	-2.39	-1.31	-0.81	-1.68	-2.31	-0.54	-0.03
M2	-3.34	2.09	-0.17	2.02	-0.47	3.68	0.69	-1.64	-1.47	0.48	-0.76	-0.25
M3	-1.64	1.42	0.60	-0.50	-1.26	-0.06	1.92	-1.91	2.86	1.10	-1.41	-0.44
M4	-1.13	0.77	2.08	-2.91	1.50	-0.66	-0.59	1.81	0.81	-0.89	-0.13	-0.24
M5	-2.77	3.28	-1.14	0.67	-0.75	0.62	-0.54	-0.23	1.08	0.78	-0.62	2.35
M6	-2.18	4.32	-0.90	-1.61	-1.40	1.26	0.82	-0.21	-0.31	-0.97	-1.37	0.77
M7	1.90	-0.25	0.67	-1.36	-1.90	-0.17	1.95	-0.81	0.05	-1.92	0.15	1.38
M8	2.03	0.29	-0.51	-1.32	-1.30	-0.29	-0.15	1.80	0.27	-2.25	-0.32	0.05
M9	-0.01	0.45	-2.19	2.55	-0.72	2.17	-2.23	-1.50	0.18	-0.34	1.70	0.45
M10	-0.81	2.22	0.80	-1.40	2.92	-0.23	-0.16	-0.79	-2.23	-0.98	-0.70	0.25

Table 5.1: Z-score values between the modules of *S. cerevisiae* and *C. albicans*.

In Table 5.1 we highlight some pairs of modules with high connectivity. Given the labels previously assign to the modules, we compare these between the selected pairs of modules. Table 5.2 displays the comparison of labels between modules of the two different species.

S. cerevisiae	C. albicans	Labels	
M0	M2	binding, metabolic process, cellular process	metabolic process, cellular process, cellular component organization or biogenesis
M1	M4	cellular component organization or biogenesis, structural molecule activity	cellular component organization or biogenesis, structural molecule activity, binding, metabolic process, cellular process
M2	M5	response to stimulus, transporter activity	response to stimulus, multi-organism process, growth, transcriptional regulator activity
M3	M8	localization, catalytic activity, transporter activity, metabolic process	catalytic activity
M5	M1		binding, metabolic process, cellular process, catalytic activity, biological regulation, response to stimulus
M6	M1	catalytic activity	binding, metabolic process, cellular process, catalytic activity, biological regulation, response to stimulus
M9	M3	cellular component organization or biogenesis, biological regulation	biological regulation, response to stimulus, multi-organism process, growth
M10	M4	metabolic process, catalytic activity	cellular component organization or biogenesis, structural molecule activity, binding, metabolic process, cellular process

Table 5.2: Comparison of labels between the modules of *S. cerevisiae* and *C. albicans*.

Using the homology mappings between species, we proved that some modules are strongly connected. By analyzing Table 5.2, we try to verify if the high connection translates into the functional similarity between modules. We start by verifying 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. Therefore, homologous genes associated with those functions may be the origin of 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 only detected in one of the modules (cellular and metabolic process). Another hypothesis might be that homologous genes can have different functions, as a result of the evolutionary divergence of species. Hereupon, two modules can be tightly connected and at the same time functionally distinct. By performing an examination of the terms shared across modules, we notice that these are widely represented terms. This evidence confirms that these are very important and common processes in the species. Closing the analysis, we notice the connection between *M5* of *S. cerevisiae* and *M1* of *C. albicans*. The first does not contain any functionality, unlike the other, in which some functions were detected. In this way, the functions of *M1* of *C. albicans* can serve as predictions

for possible functions in *M5* of *S. cerevisiae*. Yet, the current information is not enough to make these predictions. For this, it is necessary to proceed to a more detailed study of the connections between the modules.

Detailed Analysis of Connections

We conclude the comparison between modules of different species by analyzing, in a more detailed way, the most relevant connections. With this more detailed study, we aim to find an explanation for the origin of the stronger connections. In addition, we use this study to infer some functions in modules that were not detected with the label assignment process in Section 4.3. To extract the stronger and most relevant connections, we use the procedure previously illustrated with *S. cerevisiae* and *C. albicans*. Although there are some pairs with strong connections, we set a threshold for the z-score value to select only the most revealing ones.

In Table 5.3 we present the extracted connections. To study these connections, we examine the terms associated with the links that connect two different modules. A term is associated with a link if the term is common to the homologous genes in it. The value of a cell in Table 5.3 represents the ratio of genes in the module that have an association in the other module for the respective term. In other words, denotes the ratio of genes in the module that have homologous with the same function in the other. A green cell implies that the term was detected in the module through the label assignment process carried out in Section 4.3. On other hand, a red cell implies that the term was not detected. Considering the red cells, we can conclude that some modules contain functional elements not detected before.

The connections between modules are assigned to several terms. It is possible to highlight some that are present in almost all of the connections between modules. Among these terms are the metabolic process, cellular process, catalytic activity, biological regulation, and cellular component organization or biogenesis. These terms were previously identified as highly represented among different modules across species. The detailed analysis of the connections demonstrates that there are functional elements 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 the different organisms. The results obtained reinforce the idea that the functions associated with the conserved structures are truly important and essential in the species.

Using the information of Table 5.3, we can diagnose functional elements in some modules that were not detected until now. Some of these elements represent large portions of the respective modules. Good examples are the metabolic process and cellular process in *M1* of *S. cerevisiae*, localization in *M7* of *C. glabrata*, or growth in *M10* of *C. glabrata*. Finally, we look at the strong relation between *M0* of *C. albicans* and *M0* of *Y. lipolytica*. No functionality was identified in *M0* of *C. albicans* in the label

Connections	Terms									
	GO:0071840	GO:0005198	GO:0008152	GO:0009987	GO:0005488	GO:0065007	GO:0051179	GO:0003824	GO:0005215	GO:0050896
M6-Sc *M1-Ca*			0.09 0.05	0.10 0.06	0.06 0.04	0.04 0.02		0.07 0.04		0.04 0.02
M1-Sc *M4-Ca*	0.10 0.13	0.14 0.16	0.17 0.21	0.18 0.23	0.07 0.11					
M0-Sc *M0-Yl*	0.03 0.01		0.09 0.04	0.10 0.05	0.06 0.03	0.04 0.02				
M7-Sc *M0-Yl*	0.05 0.01		0.10 0.03	0.13 0.03	0.09 0.02	0.05 0.01	0.04 0.01		0.02 0.01	
M3-Sc *M7-Cg*			0.05 0.22	0.05 0.23			0.03 0.12	0.04 0.19	0.03 0.12	
M0-Ca *M0-Yl*	0.03 0.03		0.10 0.09	0.12 0.10	0.07 0.06	0.04 0.04				
M1-Ca *M3-Yl*			0.07 0.07	0.06 0.06	0.06 0.05	0.02 0.02			0.08 0.07	0.01 0.01
M2-Ca *M2-Cg*	0.04 0.12		0.05 0.14	0.06 0.19	0.04 0.11	0.02 0.06	0.02 0.05			
M1-Ca *M5-Cg*			0.04 0.17	0.04 0.17	0.03 0.13	0.01 0.06		0.04 0.17		0.02 0.09

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
GO:0051179	localization
GO:0003824	catalytic activity
GO:0005215	transporter activity
GO:0050896	response to stimulus

Table 5.3: 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).

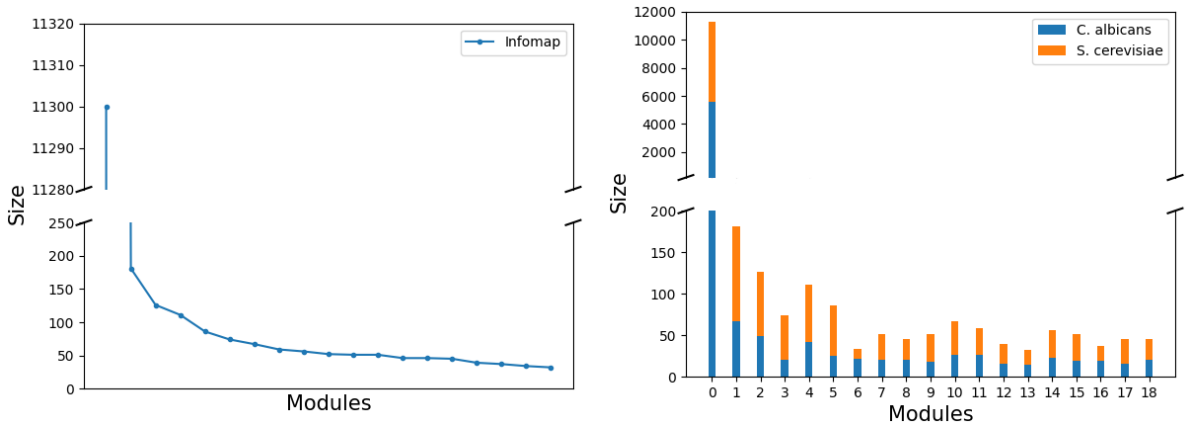
assignment process. However, with this cross-species analysis, we unveil some functional elements present in this module. With this new information, it is clear that the absence of labels assigned to this module results from its large size. Thus, the difficulty of finding particular features in large modules using the p-value is a limitation of our approach.

5.2 Multilayer Approach for Cross-Species Comparison

In the previous section, we used the homology mappings between species to find strong connections between modules. With the analysis of these connections, we end up finding 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. For example, we identified sets of homologous genes linked to the metabolic process and cellular process in the same module. These are two closely related terms since they are often represented in the same modules. Therefore, it is very likely that the sets of homologous genes associated with these terms are the same. In this final step, we build a multilayer network between species in which we perform a modules detection. To carry out the detection of the modules we use 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 in which we apply the detection of the module. From the application of the Infomap algorithm, we could find several modules. The size distribution of those modules is displayed in Figure 5.2(a). By interpreting the size distribution, we notice that there is a module that 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 5.2(b) it is possible to verify that those modules are constituted by genes from both species. The constitution of the modules appears to be balanced, having almost half the genes of each species.



(a) Modules Size Distribution

(b) Modules Constitution

Figure 5.2: Figure 5.2(a) - size distribution of the modules found with Infomap algorithm in the multilayer network of *S. cerevisiae* and *C. albicans*. Figure 5.2(b) - constitution of the modules found in the multilayer network of *S. cerevisiae* and *C. albicans*.

Once the modules are extracted, we proceed to their functional characterization. In Figure 5.3 we provide the results of the label assignment process for the modules found. In what concerns the classification of modules, we observe that *M0* has no labels associated with it. As has happened with other modules, this is a consequence of the difficulty of our approach in finding meaningful terms in modules that contain a large portion of the network genes. Regarding the remaining modules, most of them are well classified. However, some are just partially classified. For example, we have *M3*, *M9*, *M10*, *M15* and *M17* in which just a portion of the genes are associated with the obtained labels. Looking at the functionalities, most of them were previously detected in the label assignment for both species. Nonetheless, others were not, such as the molecular function regulator detected in *M12* or *M16*.

Going further with our analysis, we study the contribution of the genes of each species for the clas-

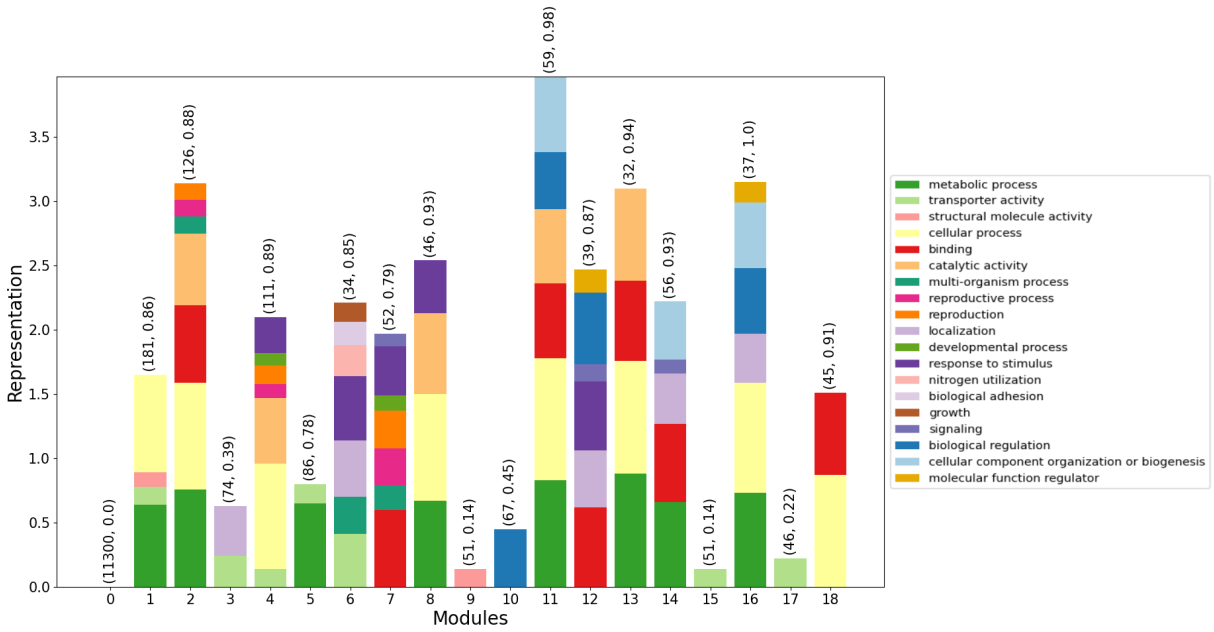


Figure 5.3: Gene Ontology terms for the different modules found in the multilayer network of *S. cerevisiae* and *C. albicans*. For each module are displayed the assigned labels. The bar size of each label depicts its representation in the module. The tuple at the top of each bar illustrates the size of the module and the proportion of the module classified with at least one of the labels.

sification of the modules in the multilayer network. To carry out this analysis, we divide each module into two groups of genes, those belonging to *S. cerevisiae* and those belonging to *C. albicans*. Then, we perform the label assignment process in those groups. The comparison between the labels of each module and those of the respective gene groups can be seen in Figure 5.4.

Looking at the first module, we can see that by dividing the module into two groups it is possible to detect some functionalities. However, these are unrelated, which indicates that this module cannot provide useful information about the similarity between species. Regarding 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. Therefore, almost all of these modules have structural elements conserved in the species. By inspecting the functions associated with those modules, it is noticed that some of these contain terms that are only associated with the genes of one species. Such as *M5* or *M18*. This way, we cannot consider these as functional elements conserved across species. Then, we have modules where there is conservation of some functions, yet, this is a residual conservation and does not represent the vast majority of the functionality of the module. This happens in *M14* with localization and in *M16* with molecular function regulator (a function not previously detected in none of the species). Finally, we have the modules that 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. In conclusion, this multilayer approach allowed us to identify functional structures conserved across species.

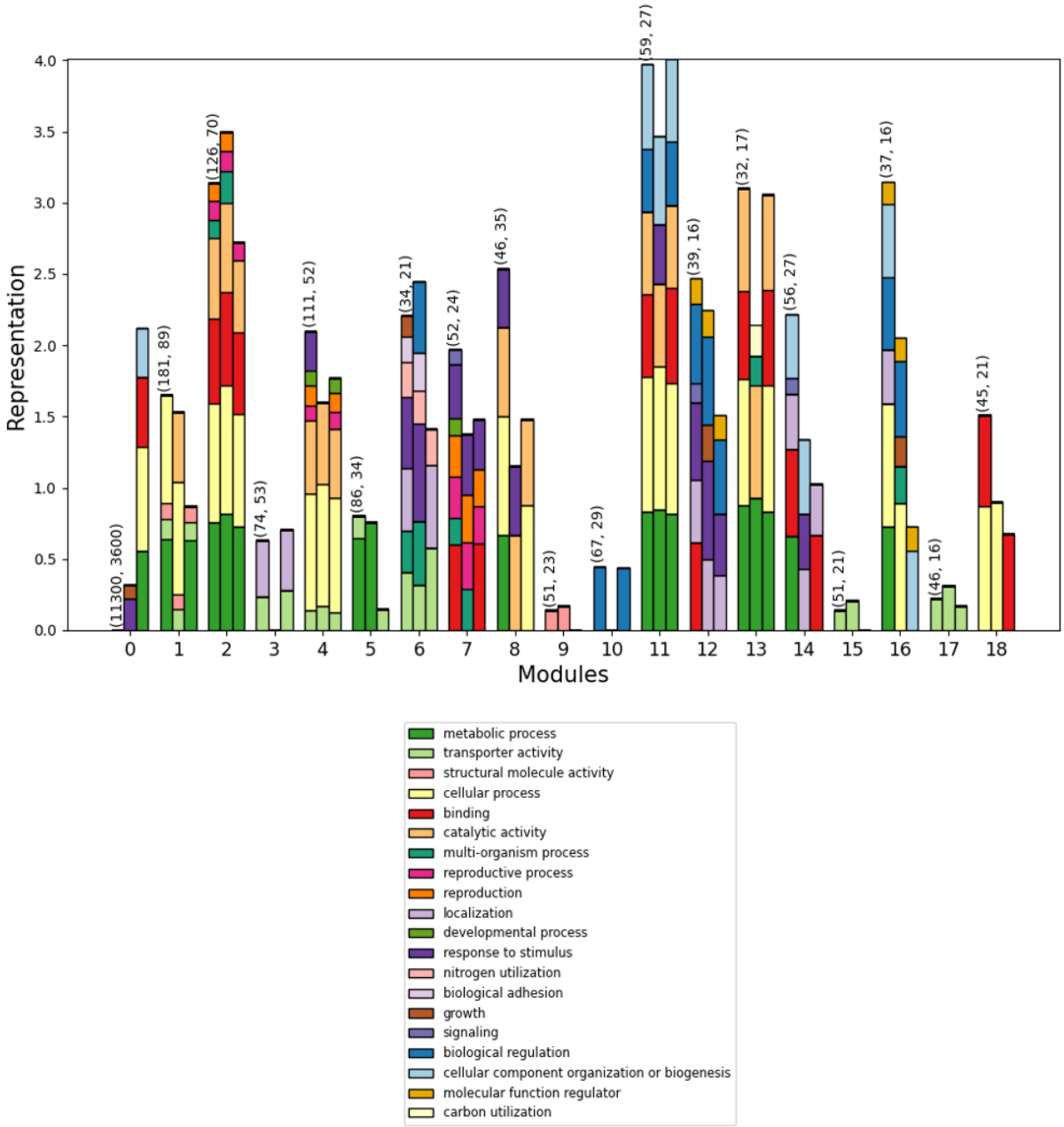


Figure 5.4: Comparison of labels between the modules of the multilayer and the respective groups of genes from *S. cerevisiae* and *C. albicans*. The three bars side-by-side respectively describe the labels of the module, of the genes from *S.cerevisiae* and the genes from *C. albicans*. At the top of the first bar of each module is shown the module size and the number of inter-layer links in the module.

6

Conclusion

Contents

6.1 Conclusions

In this thesis, we studied transcriptional regulatory networks, responsible for the genetic regulation of organisms. More specifically we analyzed the transcriptional network of closely-related yeast species. Our approach consisted of two phases. The first one corresponds to the study of functional elements in the species (Chapter 4). For this purpose, we analyzed the structure of the networks by performing a modules detection step. Then, we unveiled the functionalities associated with the genes of the extracted modules. In the second stage, we executed a cross-species comparison where we consider the homology mappings between species to find functional structures conserved among species (Chapter 5). With the results obtained in this work, we managed to contribute with important information about the species in study.

In Chapter 4. we began with the functional analysis of modules detected in the species. From the different algorithms used for module detection, 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. After extracting the modules, the functional analysis was performed using the label assignment process. 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, response to stimulus, or catalytic activity. Moving to Section 4.3, we verified that, although some modules seem functionally related, these may be responsible for distinct sub-processes. In the conclusion of the chapter, we compared the functional characterization of the different species. Here, we point out some interesting results. For instance, the identification of specific 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 Chapter 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 process, cellular process, biological regulation, among others. These connections were also fundamental to infer new functional elements in some modules that were not detected in Chapter 4. Finally, with the creation of the multilayer network, 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. This final step also allowed the identification of functions not previously detected in the species.

In conclusion, with our approach, we were able to functionally characterize the different yeast species. We identified the functions most represented in the species and responsible for most of their behavior. Beyond that, we also uncovered less represented functions responsible for specific processes in the species. Lastly, we identified the structures conserved across species and the functions associated with them.

6.2 Limitations and Future Work

Although we achieved good results with our approach, we have encountered some limitations. Starting with the detection of modules, some of the algorithms did not present acceptable results. This limited the number of algorithms used, so it was not possible to compare a wide variety of modules detection techniques. Nevertheless, the biggest constraints reside in the label assignment process. First, it is difficult to find meaningful terms with the p-value similarity 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 larger modules. A possible solution would be to increase the interval of p-value in which we consider a term as over-represented. However, this would likely lead to the detection of terms not specific to these large modules and already present in others. 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, as it happened with some specific processes (multi-organism and developmental processes) in *S. cerevisiae*. To overcome this problem, a possible solution would be to adapt the threshold value to the size of the modules. For this, we could test different values for the threshold in a set of modules with different sizes and see with which values the best performance is achieved. Then, we could verify if there is a relation between the threshold values and the size of the modules that would allow us to predict the threshold values for each module considering its size. Larger modules would have lower threshold values to facilitate the detection of more specific functions.

Lastly, additional future work is worth exploring. In the study of the results of the algorithms, we used different parameters to evaluate their performance. As future work, we could consider the creation of a measure that would allow us to evaluate the functional characterization of the modules. To create this measure, we could combine the diversity of functionality found in the modules and the proportion of genes in the modules that are covered by the functions assigned to them. Therefore, modules associated with functions covering almost all of their genes would be considered as well-classified. Moreover, 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.

Bibliography

[1] A.-L. Barabási *et al.*, *Network science*. Cambridge university press, 2016.

[2] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998. [Online]. Available: https://doi.org/10.1038/30918

[3] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of complex networks*, vol. 2, no. 3, pp. 203–271, 2014.

[4] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann, "Structure and evolution of transcriptional regulatory networks," *Current opinion in structural biology*, vol. 14, no. 3, pp. 283–291, 2004.

[5] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997.

[6] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart *et al.*, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular cell*, vol. 2, no. 1, pp. 65–73, 1998.

[7] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Molecular biology of the cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[8] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic expression programs in the response of yeast cells to environmental changes," *Molecular biology of the cell*, vol. 11, no. 12, pp. 4241–4257, 2000.

[9] H. C. Causton, B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True, E. S. Lander, and R. A. Young, "Remodeling of yeast genome expression in response to environmental changes," *Molecular biology of the cell*, vol. 12, no. 2, pp. 323–337, 2001.

[10] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, “Genome-wide mapping of in vivo protein-dna interactions,” *Science*, vol. 316, no. 5830, pp. 1497–1502, 2007.

[11] T. I. Lee and R. A. Young, “Transcription of eukaryotic protein-coding genes,” *Annual review of genetics*, vol. 34, no. 1, pp. 77–137, 2000.

[12] C. W. Garvie and C. Wolberger, “Recognition of specific dna sequences,” *Molecular cell*, vol. 8, no. 5, pp. 937–946, 2001.

[13] G. Orphanides and D. Reinberg, “A unified theory of gene expression,” *Cell*, vol. 108, no. 4, pp. 439–451, 2002.

[14] M. Ptashne and A. Gann, *Genes & signals*. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY:, 2002, vol. 402.

[15] M. Levine and R. Tjian, “Transcription regulation and animal diversity,” *Nature*, vol. 424, no. 6945, pp. 147–151, 2003.

[16] E. H. Davidson, *Genomic regulatory systems: in development and evolution*. Elsevier, 2001.

[17] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon *et al.*, “Transcriptional regulatory networks in saccharomyces cerevisiae,” *science*, vol. 298, no. 5594, pp. 799–804, 2002.

[18] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.-H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena *et al.*, “A genomic regulatory network for development,” *science*, vol. 295, no. 5560, pp. 1669–1678, 2002.

[19] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, “Genomic analysis of regulatory network dynamics reveals large topological changes,” *Nature*, vol. 431, no. 7006, pp. 308–312, 2004.

[20] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, “From molecular to modular cell biology,” *Nature*, vol. 402, no. 6761, pp. C47–C52, 1999.

[21] A. C. Lewis, N. S. Jones, M. A. Porter, and C. M. Deane, “The function of communities in protein interaction networks at multiple scales,” *BMC systems biology*, vol. 4, no. 1, pp. 1–14, 2010.

[22] K. Voevodski, S.-H. Teng, and Y. Xia, “Finding local communities in protein networks,” *BMC bioinformatics*, vol. 10, no. 1, pp. 1–14, 2009.

[23] S. Fortunato, “Community detection in graphs,” *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[24] A. M. Wiles, M. Doderer, J. Ruan, T.-T. Gu, D. Ravi, B. Blackman, and A. J. Bishop, "Building and analyzing protein interactome networks by cross-species comparisons," *BMC systems biology*, vol. 4, no. 1, pp. 1–16, 2010.

[25] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M. Vidal, "Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"," *Genome research*, vol. 11, no. 12, pp. 2120–2126, 2001.

[26] A. R. Borneman, T. A. Gianoulis, Z. D. Zhang, H. Yu, J. Rozowsky, M. R. Seringhaus, L. Y. Wang, M. Gerstein, and M. Snyder, "Divergence of transcription factor binding sites across related yeast species," *Science*, vol. 317, no. 5839, pp. 815–819, 2007.

[27] H. Zhang, Y. Li, Y. Liu, H. Liu, H. Wang, W. Jin, Y. Zhang, C. Zhang, and D. Xu, "Role of plant microrna in cross-species regulatory networks of humans," *BMC systems biology*, vol. 10, no. 1, pp. 1–10, 2016.

[28] L. Cantini, E. Medico, S. Fortunato, and M. Caselle, "Detection of gene communities in multi-networks reveals cancer drivers," *Scientific reports*, vol. 5, no. 1, pp. 1–10, 2015.

[29] A. Rai, P. Pradhan, J. Nagraj, K. Lohitesh, R. Chowdhury, and S. Jalan, "Understanding cancer complexome using networks, spectral graph theory and multilayer framework," *Scientific reports*, vol. 7, no. 1, pp. 1–16, 2017.

[30] P. T. Monteiro, J. Oliveira, P. Pais, M. Antunes, M. Palma, M. Cavalheiro, M. Galocha, C. P. Godinho, L. C. Martins, N. Bourbon, M. N. Mota, R. A. Ribeiro, R. Viana, I. Sá-Correia, and M. C. Teixeira, "YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts," *Nucleic Acids Research*, vol. 48, no. D1, pp. D642–D649, Oct. 2019. [Online]. Available: https://doi.org/10.1093/nar/gkz859

[31] A. Bavelas, "Communication patterns in task-oriented groups," *The journal of the acoustical society of America*, vol. 22, no. 6, pp. 725–730, 1950.

[32] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, p. 35, Mar. 1977. [Online]. Available: https://doi.org/10.2307/3033543

[33] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[34] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[35] P. Erdös and A. Rényi, “On random graphs i,” *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[36] E. N. Gilbert, “Random graphs,” *The Annals of Mathematical Statistics*, vol. 30, no. 4, pp. 1141–1144, Dec. 1959. [Online]. Available: https://doi.org/10.1214/aoms/1177706098

[37] J. Travers and S. Milgram, “An experimental study of the small world problem,” *Sociometry*, vol. 32, no. 4, p. 425, Dec. 1969. [Online]. Available: https://doi.org/10.2307/2786545

[38] J. S. Kleinfeld, “The small world problem,” *Society*, vol. 39, no. 2, pp. 61–66, Jan. 2002. [Online]. Available: https://doi.org/10.1007/bf02717530

[39] R. Albert, H. Jeong, and A.-L. Barabási, “Diameter of the world-wide web,” *Nature*, vol. 401, no. 6749, pp. 130–131, Sep. 1999. [Online]. Available: https://doi.org/10.1038/43601

[40] M. P. Stumpf, C. Wiuf, and R. M. May, “Subnets of scale-free networks are not scale-free: sampling properties of networks,” *Proceedings of the National Academy of Sciences*, vol. 102, no. 12, pp. 4221–4224, 2005.

[41] M. Boguná, R. Pastor-Satorras, and A. Vespignani, “Cut-offs and finite size effects in scale-free networks,” *The European Physical Journal B*, vol. 38, no. 2, pp. 205–209, 2004.

[42] S. N. Dorogovtsev, J. F. Mendes, and A. N. Samukhin, “Size-dependent degree distribution of a scale-free growing network,” *Physical Review E*, vol. 63, no. 6, p. 062101, 2001.

[43] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999. [Online]. Available: https://doi.org/10.1126/science.286.5439.509

[44] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network motifs: simple building blocks of complex networks,” *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[45] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin, “The structure and dynamics of multilayer networks,” *Physics reports*, vol. 544, no. 1, pp. 1–122, 2014.

[46] F. Battiston, V. Nicosia, and V. Latora, “The new challenges of multiplex networks: Measures and models,” *The European Physical Journal Special Topics*, vol. 226, no. 3, pp. 401–416, 2017.

[47] ——, “Structural measures for multiplex networks,” *Physical Review E*, vol. 89, no. 3, p. 032804, 2014.

[48] G. Menichetti, D. Remondini, P. Panzarasa, R. J. Mondragón, and G. Bianconi, “Weighted multiplex networks,” *PloS one*, vol. 9, no. 6, p. e97857, 2014.

[49] M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivelä, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas, "Mathematical formulation of multilayer networks," *Physical Review X*, vol. 3, no. 4, p. 041022, 2013.

[50] V. Nicosia and V. Latora, "Measuring and modeling correlations in multiplex networks," *Physical Review E*, vol. 92, no. 3, p. 032805, 2015.

[51] F. Battiston, "The structure and dynamics of multiplex networks," Ph.D. dissertation, Queen Mary University of London, 2017.

[52] E. Cozzo, M. Kivelä, M. De Domenico, A. Solé-Ribalta, A. Arenas, S. Gómez, M. A. Porter, and Y. Moreno, "Structure of triadic relations in multiplex networks," *New Journal of Physics*, vol. 17, no. 7, p. 073029, 2015.

[53] R. G. Morris and M. Barthelemy, "Transport on coupled spatial networks," *Physical review letters*, vol. 109, no. 12, p. 128703, 2012.

[54] L. Solá, M. Romance, R. Criado, J. Flores, A. García del Amo, and S. Boccaletti, "Eigenvector centrality of nodes in multiplex networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 23, no. 3, p. 033131, 2013.

[55] A. Halu, R. J. Mondragón, P. Panzarasa, and G. Bianconi, "Multiplex pagerank," *PloS one*, vol. 8, no. 10, p. e78293, 2013.

[56] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[57] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the national academy of sciences*, vol. 101, no. 9, pp. 2658–2663, 2004.

[58] S. Fortunato, V. Latora, and M. Marchiori, "Method to find community structures based on information centrality," *Physical review E*, vol. 70, no. 5, p. 056104, 2004.

[59] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical review E*, vol. 69, no. 6, p. 066133, 2004.

[60] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.

[61] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[62] V. A. Traag, L. Waltman, and N. J. Van Eck, “From louvain to leiden: guaranteeing well-connected communities,” *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[63] R. Guimera and L. A. N. Amaral, “Functional cartography of complex metabolic networks,” *nature*, vol. 433, no. 7028, pp. 895–900, 2005.

[64] L. Donetti and M. A. Munoz, “Detecting network communities: a new systematic and efficient algorithm,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2004, no. 10, p. P10012, 2004.

[65] A. Capocci, V. D. Servedio, G. Caldarelli, and F. Colaiori, “Detecting communities in large networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 352, no. 2-4, pp. 669–676, 2005.

[66] B. Yang and J. Liu, “Discovering global network communities based on local centralities,” *ACM Transactions on the Web (TWEB)*, vol. 2, no. 1, pp. 1–32, 2008.

[67] M. Rosvall and C. T. Bergstrom, “Maps of random walks on complex networks reveal community structure,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[68] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” *Physical review E*, vol. 76, no. 3, p. 036106, 2007.

[69] S. M. Van Dongen, “Graph clustering by flow simulation,” Ph.D. dissertation, Faculteit Wiskunde en Informatica, Universiteit Utrecht, 2000.

[70] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, “Cfinder: locating cliques and overlapping modules in biological networks,” *Bioinformatics*, vol. 22, no. 8, pp. 1021–1023, 2006.

[71] J. Baumes, M. K. Goldberg, M. S. Krishnamoorthy, M. Magdon-Ismail, and N. Preston, “Finding communities by clustering a graph into overlapping subgraphs.” *IADIS AC*, vol. 5, pp. 97–104, 2005.

[72] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, “Link communities reveal multiscale complexity in networks,” *nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[73] S. Gregory, “Finding overlapping communities using disjoint community detection algorithms,” in *Complex networks*. Springer, 2009, pp. 47–61.

[74] P. Anchuri and M. Magdon-Ismail, “Communities and balance in signed networks: A spectral approach,” in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 2012, pp. 235–242.

[75] F. Bonchi, E. Galimberti, A. Gionis, B. Ordozgoiti, and G. Ruffo, “Discovering polarized communities in signed networks,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 961–970.

[76] M. Cucuringu, P. Davies, A. Glielmo, and H. Tyagi, “Sponge: A generalized eigenproblem for clustering signed networks,” in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1088–1098.

[77] P. Esmailian and M. Jalili, “Community detection in signed networks: the role of negative ties in different scales,” *Scientific reports*, vol. 5, no. 1, pp. 1–17, 2015.

[78] J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[79] C. Castellano, F. Cecconi, V. Loreto, D. Parisi, and F. Radicchi, “Self-contained algorithms to detect communities in networks,” *The European Physical Journal B*, vol. 38, no. 2, pp. 311–319, 2004.

[80] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, and Y. Fan, “Clustering coefficient and community structure of bipartite networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 27, pp. 6869–6875, 2008.

[81] V. Latora and M. Marchiori, “Efficient behavior of small-world networks,” *Physical review letters*, vol. 87, no. 19, p. 198701, 2001.

[82] M. E. Newman, “Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality,” *Physical review E*, vol. 64, no. 1, p. 016132, 2001.

[83] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[84] J. MacQueen *et al.*, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[85] P. Bonacich, “Factoring and weighting approaches to status scores and clique identification,” *Journal of mathematical sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[86] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[87] P. D. G. I. J. Myung and M. A. Pitt, *Advances in minimum description length: Theory and applications*. MIT press, 2005.

[88] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[89] S. Gregory, "An algorithm to find overlapping community structure in networks," in *European conference on principles of data mining and knowledge discovery*. Springer, 2007, pp. 91–102.

[90] B. Karrer, E. Levina, and M. E. Newman, "Robustness of community structure in networks," *Physical review E*, vol. 77, no. 4, p. 046119, 2008.

[91] A. Lancichinetti, F. Radicchi, and J. J. Ramasco, "Statistical significance of communities in networks," *Physical Review E*, vol. 81, no. 4, p. 046110, 2010.

[92] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[93] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[94] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proceedings of the national Academy of sciences*, vol. 100, no. 3, pp. 1128–1133, 2003.

[95] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the national Academy of sciences*, vol. 100, no. 21, pp. 12 123–12 128, 2003.

[96] M. Blatt, S. Wiseman, and E. Domany, "Superparamagnetic clustering of data," *Physical review letters*, vol. 76, no. 18, p. 3251, 1996.

[97] J. Chen and B. Yuan, "Detecting functional modules in the yeast protein–protein interaction network," *Bioinformatics*, vol. 22, no. 18, pp. 2283–2290, 2006.

[98] V. Farutin, K. Robison, E. Lightcap, V. Dancik, A. Ruttenberg, S. Letovsky, and J. Pradines, "Edge-count probabilities for the identification of local protein communities and their organization," *Proteins: Structure, Function, and Bioinformatics*, vol. 62, no. 3, pp. 800–818, 2006.

[99] T. Z. Sen, A. Kloczkowski, and R. L. Jernigan, "Functional clustering of yeast proteins from the protein-protein interaction network," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–13, 2006.

[100] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical review E*, vol. 74, no. 1, p. 016110, 2006.

[101] P. Holme, M. Huss, and H. Jeong, "Subnetwork hierarchies of biochemical pathways," *Bioinformatics*, vol. 19, no. 4, pp. 532–538, 2003.

[102] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[103] D. M. Wilkinson and B. A. Huberman, "A method for finding communities of related genes," *proceedings of the national Academy of sciences*, vol. 101, no. suppl 1, pp. 5241–5248, 2004.

[104] G. de Anda-Jáuregui, S. A. Alcalá-Corona, J. Espinal-Enríquez, and E. Hernández-Lemus, "Functional and transcriptional connectivity of communities in breast cancer co-expression networks," *Applied Network Science*, vol. 4, no. 1, pp. 1–13, 2019.

[105] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young *et al.*, "Computational discovery of gene modules and regulatory networks," *Nature biotechnology*, vol. 21, no. 11, pp. 1337–1342, 2003.

[106] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker, "Conserved patterns of protein interaction in multiple species," *Proceedings of the National Academy of Sciences*, vol. 102, no. 6, pp. 1974–1979, 2005.

[107] B. P. Kelley, R. Sharan, R. M. Karp, T. Sittler, D. E. Root, B. R. Stockwell, and T. Ideker, "Conserved pathways within bacteria and yeast as revealed by global protein network alignment," *Proceedings of the National Academy of Sciences*, vol. 100, no. 20, pp. 11 394–11 399, 2003.

[108] J. H. Caufield, C. Wimble, S. Shary, S. Wuchty, and P. Uetz, "Bacterial protein meta-interactomes predict cross-species interactions and protein function," *Bmc Bioinformatics*, vol. 18, no. 1, pp. 1–14, 2017.

[109] X. Wang and Y. Jin, "Predicted networks of protein-protein interactions in stegodyphus mimosarum by cross-species comparisons," *BMC genomics*, vol. 18, no. 1, pp. 1–13, 2017.

[110] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich, "Fast and effective prediction of microrna/target duplexes," *Rna*, vol. 10, no. 10, pp. 1507–1517, 2004.

[111] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *science*, vol. 302, no. 5643, pp. 249–255, 2003.

[112] S. Bergmann, J. Ihmels, N. Barkai, and M. Eisen, "Similarities and differences in genome-wide expression data of six organisms," *PLoS biology*, vol. 2, no. 1, p. e9, 2004.

[113] M. Zitnik and J. Leskovec, "Predicting multicellular function through multi-layer tissue networks," *Bioinformatics*, vol. 33, no. 14, pp. i190–i198, 2017.

[114] P. Kapadia, S. Khare, P. Priyadarshini, and B. Das, “Predicting protein-protein interaction in multi-layer blood cell ppi networks,” in *International Conference on Advanced Informatics for Computing Research*. Springer, 2019, pp. 240–251.

[115] P. Shinde and S. Jalan, “A multilayer protein-protein interaction network analysis of different life stages in caenorhabditis elegans,” *EPL (Europhysics Letters)*, vol. 112, no. 5, p. 58001, 2015.

[116] B. Zhao, S. Hu, X. Li, F. Zhang, Q. Tian, and W. Ni, “An efficient method for protein function annotation based on multilayer protein networks,” *Human genomics*, vol. 10, no. 1, pp. 1–15, 2016.

[117] L. Liang, V. Chen, K. Zhu, X. Fan, X. Lu, and S. Lu, “Integrating data and knowledge to identify functional modules of genes: a multilayer approach,” *BMC bioinformatics*, vol. 20, no. 1, pp. 1–15, 2019.

[118] L. Yu, Y. Shi, Q. Zou, and L. Gao, “Studying the drug treatment pattern based on the action of drug and multi-layer network model,” *bioRxiv*, p. 780858, 2019.

[119] W. Zheng, D. Wang, and X. Zou, “Control of multilayer biological networks and applied to target identification of complex diseases,” *BMC bioinformatics*, vol. 20, no. 1, pp. 1–12, 2019.

[120] A. J. Gates and Y.-Y. Ahn, “Clusim: A python package for calculating clustering similarity,” *Journal of Open Source Software*, vol. 4, no. 35, p. 1264, 2019.