Functional Characterization of Transcriptional Regulatory Networks of Yeast Species

Paulo Dias[1(✉)], Pedro T. Monteiro[1,2], and Andreia Sofia Teixeira[2,3]

[1] Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
paulo.a.c.dias@tecnico.ulisboa.pt
[2] INESC-ID, Lisboa, Portugal
[3] LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

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 work, we study transcriptional regulatory networks of closely related yeast species with the aim of revealing which functions or processes are encoded in the regulatory network topology. The first phase of this work consists of the detection of modules followed by their functional characterization. Here, we unveil the functionality of the species by capturing it in functional modules. In the second phase, we move towards a cross-species analysis where we compare the functional modules of the different species to settle the similarities between them. Lastly, we use a multilayer network approach to combine the genetic information of different species. We seek to identify the functional elements conserved across the different organisms by applying a detection of modules in the multilayer network.

Keywords: Complex Networks · Transcriptional Regulatory Networks · Multilayer Networks · Community Detection · Functional Modules

1 Introduction

Gene expression is the biological process that allows a cell to respond to its changing environment. Each cell is the product of specific gene expression events involving the transcription of thousands of genes. The transcription factors (TFs) are the core elements in the control of gene expression. These genes are responsible for activating or inhibiting the genes under their regulation, the target genes (TGs). Normally, the expression level of a target gene is the result of the combinatorial regulation of multiple transcription factors. The hundreds of interactions

P. Ribeiro et al. (Eds.): NetSci-X 2022, LNCS 13197, pp. 1–15, 2022.
https://doi.org/10.1007/978-3-030-97240-0_11

between transcription factors and target genes define a transcriptional regulatory network that underlies cellular identity and function. The morphological differences between species/organisms arise from the gene's differential regulation encoded in the transcriptional regulatory networks. Thus, these networks are of great biological importance since their analysis is fundamental to understanding differential gene expression [1,2]. Therefore, insights from the structure and function of these networks are essential to the study of organisms. However, despite their central role in biology, the structure and dynamics of these type networks are still not completely understood.

In biological networks, communities can share common biological functions, and they are studied in the investigation of cellular systems of organisms. The study of communities have allowed the identification of important protein complexes in protein-protein interaction (PPI) networks [3,4]. In gene regulatory networks, we highlight the discovery of functionally related groups of genes [5] and of groups of genes associated with functions that drive cancer [6].

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. In PPI networks, cross-species have been used to predict protein-protein interactions (interologues) conserved across species [7,8]. Moreover, the characterization of interspecies differences in gene regulation has already proven to be fundamental for understanding the diversity and evolution of species [9,10]. Multilayer network approaches are useful in studies involving different types of data since it allows its representation and comparison. As examples, already helped to make predictions in protein functions in yeast [11] or to recognize candidate driver cancer genes [12,13].

In this work, we characterize transcriptional regulatory networks of closely related species. In particular, we consider data from YEASTRACT+[14], 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. These networks are represented as graphs, the transcription factors and target genes are represented by the nodes and the interactions between them by the edges. 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 within each species and also between species.

In the first phase, we perform a detection of modules, applying several community detection techniques to understand which one is the most suitable for the considered networks, followed by their functional characterization to divide the networks into functional elements that may represent the different functions of the species. The functional characterization of communities is done using the Gene Ontology[1] [15]. Considering that transcription factors may be associated with multiple regulatory processes, we include the study of overlapping communities, as they allow genes to belong to different functional groups. Moreover, since the regulatory associations are negative (inhibition) or positive (activation), we also consider the division of the network in polarized communities.

[1] http://geneontology.org/.

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

Network	#Nodes	#Edges	#TFs	#TGs	$\langle k \rangle$	CC	D
S. cerevisiae	6 886	195 498	220	6 886	56.60	0.47	4
S. cerevisiae B	6 478	45 209	176	6 475	13.93	0.22	5
C. albicans	6 015	35 687	118	6 015	11.83	0.28	5
Y. lipolytica	5 288	9 238	5	5 288	3.49	0.36	4
C. parapsilosis	3 381	6 986	11	3 380	4.13	0.25	4
C. glabrata	2 133	3 508	40	2 116	3.29	0.09	6*
C. tropicalis	665	698	16	663	2.08	0.01	5
K. pastoris	561	581	4	559	2.07	0.01	5
K. lactis	111	126	10	106	2.25	0.15	2*
Z. bailii	32	31	1	31	1.94	0.00	2
K. marxianus	4	3	1	3	1.50	0.00	2

Regarding community detection algorithms, we highlight some of the most recognized. The Girvan-Newman [16] is the most commonly used divisive algorithm. About modularity-optimization-based methods, we underline the Louvain [17], the Clauset-Newman-Moore [18] and Leiden [19] algorithms. The spectral algorithms, such as the Donetti-Muñoz algorithm [20], are also a well-known class of techniques. Enumerating other techniques, we have the Infomap [21], the Label Propagation [22] and the Markov Cluster algorithm [23]. In the detection of overlapping communities, we point to the CFinder algorithm [24]. For more details, we refer to the review from Fortunato *et al.* [25].

Moving to the second stage, we start by settling the similarities among species by comparing the functional modules between these. We also use the connections between species to infer functional elements not previously detected in some organisms. Finally, we use a multilayer network approach combining the genetic information of the species in which we apply a modules detection algorithm to find functional elements conserved across species.

2 Identification of Functional Modules

2.1 Data

We consider data from the YEASTRACT+[2] portal which provides the transcriptional regulatory networks of 10 closely-related yeast species [14]. The characteristics of these networks are presented in Table 1. We can observe that the different species have different levels of documentation, as reflected by the number of nodes and edges. The gene associations may be classified into two major

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

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

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	-

groups: (1) DNA binding evidence; (2) expression evidence. Due to the high level of information of *S. cerevisiae*, we consider a new network to our set denoted *S. cerevisiae B*, which 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 for these species. Comparing the characteristics of the original and filtered network, we observe that the number of nodes, transcription factors, and target genes remains close to the original. This indicates that filtering the original network managed to retain most of the genetic evidence of *S. cerevisiae*. Unlike the species mentioned above, there are species whose networks are small and sparse – *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 discarded these networks from the current analysis.

2.2 Comparative Analysis of Modules

For the detection of modules, we select a collection of algorithms that exploit the diverse ideas and techniques of Network Science developed over the years. The set is composed of the following algorithms: Girvan-Newman (GN), Louvain, Leiden, Clauset-Newman-Moore (CNM), Label Propagation (LP), Markov Clustering (MC), Infomap, CFinder (CF), and a spectral clustering technique (SC) for modules detection on signed networks. To execute the introduced algorithms, we used libraries where they are already implemented. Some of the considered algorithms are stochastic, i.e., the result may change in each run because their procedure depends on random events. The Louvain, the Label Propagation, and the Infomap are the non-deterministic algorithms we use in our approach. To compare the different outputs of the algorithms, we run these algorithms 1 000 times. Next, to study the different partitions obtained, we compare each pair of different partitions having the number of modules equal to the value of the mode. To make this comparison, we use the package *clusim* [26] that allows us to compare different partitions using similarity measures, in our case we use *Rand Index* [27]. Despite the stochasticity of the algorithms, we obtain high values for the measure of similarity of the considered pairs of partitions and low variance between them, showing that the structural differences between the

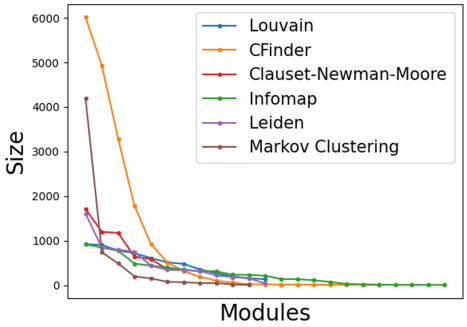


Fig. 1. Modules size distribution for *C. albicans*

partitions are minimal. Thus, regarding stochastic algorithms, we adopt one of the results having the number of modules equal to the mode. Due to the temporal complexity of Girvan-Newman and CFinder algorithms, it was not possible to run them on some of the biggest networks in a reasonable time. Table 2 displays the number of modules obtained for the networks using the different algorithms.

The results in Table 2 show that some of the algorithms fail to detect distinct modules, such as Label Propagation, Girvan-Newman, and Spectral Clustering algorithms in signed networks, which lead us not to choose to study these results. In the *S. cerevisiae*, we detected more modules in the filtered network than in the original network. The applied filter reveals to be essential in the study of the species, the large number of modules found suggests the possibility of discovering a greater diversity of behaviors in the species. Therefore, we decided to use the results of the filtered network to study the respective species. In *Y. lipolytica* few modules were detected, a consequence of the low number of transcription factors. Regarding the other species, it was possible to extract some modules, indicating that these species may contain genetic information about more processes than *Y. lipolytica*. To better understand the division into modules, we decided to study the distribution of their sizes for the different algorithms. In the Fig. 1 we present the distributions for *C. albicans* as example.

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. The distribution shows that the modularity-based algorithms (Louvain, Leiden and, Clauset-Newman-Moore) have a more balanced division than the others. Infomap, despite having some very small modules, produced others with equivalent size to those mentioned above. CFinder, although it has modules which include almost the entire network, the smaller ones can help us understand if the species benefit from an overlapping communities study. Lastly, the Markov Clustering algorithm gives

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

	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	-	-	-	-

us a unbalanced division, having only two modules of same magnitude of those found with the other algorithms, therefore, we decided to discard these results.

To close the first phase of our analysis, we calculated the C-score and B-score [28] for the modules obtained with the modularity-based algorithms. These measures allow us to evaluate the significance of those modules by testing their robustness and stability against random perturbations of the graph structure. These results are presented in Table 3. Looking at the C-score values, none of the algorithms could identify significant modules, consequence of the restrictive null model of the method. The B-score, which uses a less restrictive null model, identifies some modules as significant. According to the B-score values, the Louvain algorithm only produced one significant module, which may be a consequence of its stochasticity. Regarding the other two algorithms, both produced significant modules. Combining the significance of some modules and the balanced division, at that point, Leiden showed to be the one that best captures the structure of the species. Nevertheless, in the functional analysis, we take into account the results of Infomap, CFinder, Louvain, and Clauset-Newman-Moore, which also presented interesting results.

2.3 Functional Analysis of Modules

In this section we provide the functional characterization of the modules previously detected through the label assignment process. These labels represent specific functionalities of the species. The idea is to associate the modules to the most represented and significant Gene Ontology terms among their genes. Given the whole set of terms associated with a module, we perform a three-step filtering of the terms to find the most representative and significant terms: (1)

select only the most global terms (level 2 terms of the Gene Ontology hierarchy); (2) keep only the most over-represented terms of the module using the hypergeometric test (we consider a term as over-represented if is p-value $\in [0; 0.05]$); (3) retain the terms represented in at least 10% of the module.

Algorithms Performance

Using *S. cerevisiae* network as a reference, we compare the performance of the considered algorithms. Beginning with the modularity-based methods, Fig. 2. A first look shows that most modules have more than one label, exposing the functional diversity within these. However, it is observable that not all genes in the modules are linked to functionalities that characterize the modules they belong to. By applying the p-value filtering, we obtain only the most specific terms from 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 by other modules.

In Fig. 2, we observe that some functions appear with high representation in the modules. Such as the metabolic process, cellular process, biological regulation, or response to stimulus. In contrast, others seem to be less represented. Being specific functions, these are associated with a smaller set of genes. Reproduction, reproductive process, and transporter activity are good examples of specific functions detected in the modules. The Clauset-Newman-Moore algorithm captured a smaller diversity of functions, failing to identify some functions present in the modules originated by the Louvain and Leiden algorithms. Comparing the results from Louvain and Leiden we can observe that some modules are very similar in terms of functionality. However, Leiden was able to identify functions that Louvain could not, such as the cellular process (usually heavily represented in modules) or reproductive process. Moreover, Leiden was the algorithm in whose modules it was possible to identify more functions of the species, indicating that the division of the species obtained with this algorithm is the one that better reflects the division of functionality of the species.

Regarding the study of overlapping communities, it was possible to retain some new information about the species, such as the presence of functions not previously detected: transcription regulator activity, developmental process, and signaling. However, the study of overlapping communities is not enough to functionally characterize the species, since most communities are small components of larger communities. This results in most communities to be associated with the same behaviors. The performance of the Infomap algorithm lacked consistency. Although it managed to classify some modules of relevant size, it failed to classify the vast majority of modules.

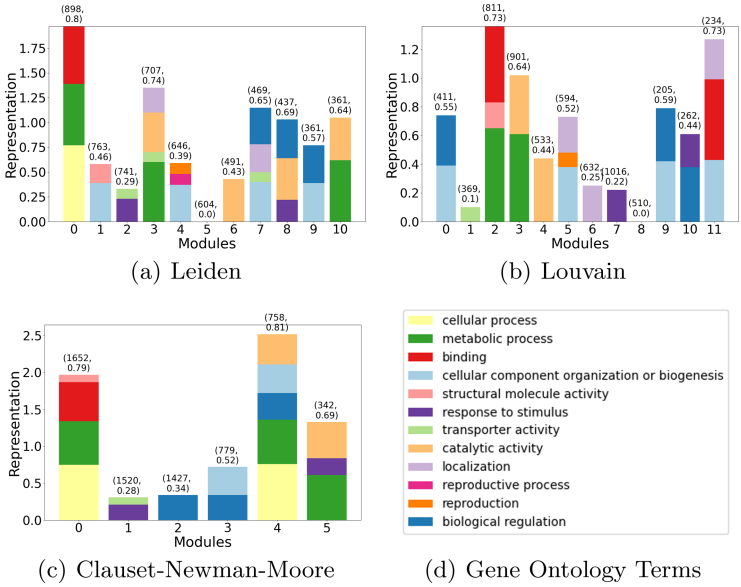
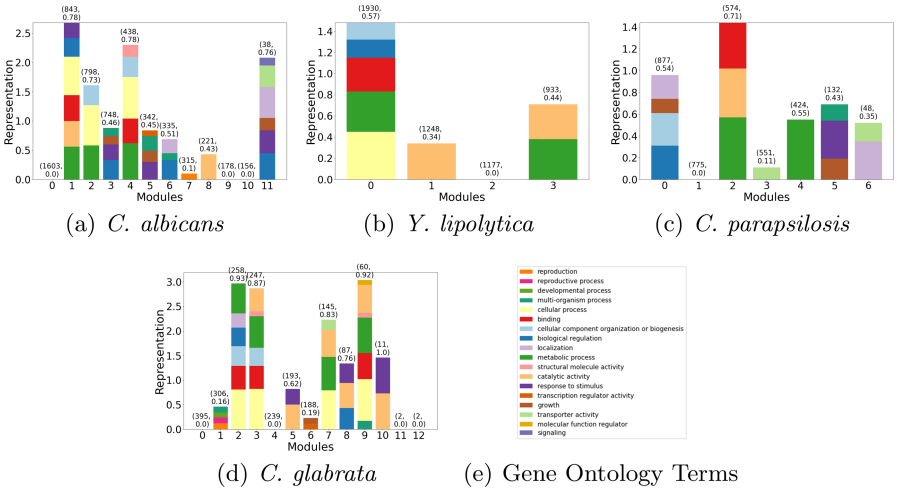


Fig. 2. Modules and respective functions for modularity-based methods on *S. cerevisiae*. The bar of each term symbolizes its representation in the module. The pair of values at the top of each bar are respectively the size of the module and the fraction of genes of the module related with at least one term (in the module).

Functional Analysis of Remaining Species

Additionally, we analyze the results of the label assignment process for the remaining species. We use the results obtained with the Leiden algorithm, Fig. 3, since it is the algorithm that best captures the functions of *S. cerevisiae*.

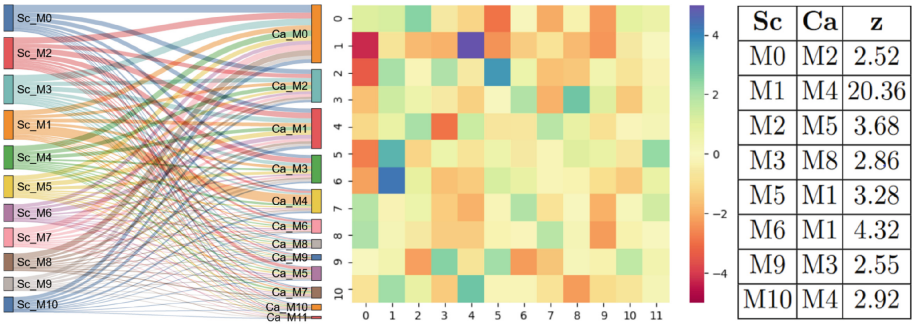
Starting with *C. albicans*, we notice the absence of terms in modules *M0*, *M9*, *M10*. In *M0*, since the module encompasses a large part of the species, it is difficult to detect most over-represented terms using the p-value. All the remaining modules are associated with at least one function. Many of those are associated with three or more terms, capturing many of the functions of the species. An interesting point is the association of some modules to functions such as multi-organism process and growth, which are not sufficiently representative/significant to be associated with a module in *S. cerevisiae*. Also in *C. parapsilosis* and *C. glabrata*, some modules are associated with functions not detected in *S. cerevisiae*. Due to the large sizes of *S. cerevisiae* modules, it is difficult for specific terms to have a good representation in these, since they are associated with few genes. In all of these species, general functions already captured in *S. cerevisiae* were also detected, such as metabolic process, response to stimulus, or biological regulation. Revealing once again the central role these have in the functionality of different organisms. It is noticed that the modules of *C. glabrata* are associated with more functionality than the modules of *C.*



(a) *C. albicans* (b) *Y. lipolytica* (c) *C. parapsilosis*

(d) *C. glabrata* (e) Gene Ontology Terms

Fig. 3. Label Assignment results for the different species using Leiden algorithm.



(a) Sankey Diagram (b) Heat map

Sc	Ca	z
M0	M2	2.52
M1	M4	20.36
M2	M5	3.68
M3	M8	2.86
M5	M1	3.28
M6	M1	4.32
M9	M3	2.55
M10	M4	2.92

(c) Z-scores

Fig. 4. (a) - Sankey diagram representing the connections between the modules of *S. cerevisiae* and *C. albicans*. (b) - Heat map representing the level of connectivity between the modules of *S. cerevisiae* (y axis) and *C. albicans* (x axis). (c) - Table with the highest Z-score values.

parapsilosis and *Y. lipolytica*, although we have more generic evidence on the last two. Whereas that *C. glabrata* has more transcription factors, we assume that the information about this species contains genetic evidence about more biological processes. This results in a more diversified classification of modules in comparison to *C. parapsilosis* and *Y. lipolytica*.

3 Cross-Species Comparison

3.1 Functional Comparison of Modules

To compare the modules we resort to the homology mappings between species to establish the connections between species. Each link in a homology mapping denotes the connection between two homologous genes. In biology, it is established that the DNA sequence of two homologous genes derives from a common ancestor (may or may not have the same function). For this work, the homology mappings are also obtained from YEASTRACT+ [14].

S. cerevisiae vs *C. albicans*

We now focus on the comparison between *S. cerevisiae* and *C. albicans*. For this purpose, we explore the level of connection between the functional modules obtained with the Leiden algorithm. In Fig. 4(a) we present a Sankey diagram representing the connections between the modules for both species.

To understand the level of connection between modules, we perform an analysis to assess the quality of the mappings. First, we calculate the number of links shared between every pair of modules of the two species. Then, we compare these distributions with 1 000 realizations of the same process in a null model, which consists of maintaining the community structure of both networks but with randomization of the nodes. Consequently, this procedure results in different mappings between species. In Fig. 4(b) we present the heat map of the z-scores representing the level of connection between modules. The analysis of the heat map reveals the existence of some pairs of modules with strong connections (green and blue colors), these pairs are listed in Fig. 4(c).

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

Table 4. Strongly connected pairs of modules from different species. For each module, we can consult the percentage of genes that have homologous with the same function in the other module that is part of the connection. 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).

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

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

Detailed Analysis of Connections

We examine the terms associated with the links of the connections between modules of different species. A term is associated with a link if the term is common to the homologous genes in it. In Table 4 we present some of the most relevant connections among species. The detailed analysis of the connections demonstrates that there are functional groups of considerable size in different species formed by homologous genes with the same functions. This evidence reveals the conservation of functional elements across different organisms. Also, using the information of Table 4, we can diagnose functional elements in some modules that were not detected with the previous analysis. Such as the metabolic and cellular processes in *M1* of *S. cerevisiae*. Finally, we look at the connection between *M0* of *C. albicans* and *M0* of *Y. lipolytica*. In this cross-species analysis, we unveil some functional elements present in *M0* of *C. albicans*. With this new information, it is clear that the absence of labels assigned to this module in the label assignment process results from its large size.

3.2 Multilayer Approach for Cross-Species Comparison

In the previous section, we found functional elements conserved across species. However, we did not check if these elements have other associated functions or even if they overlap, since each gene can have more than one function associated to it. Therefore, in this final step, we build a multilayer network between species in which we perform a module detection task using the Infomap algorithm, which 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.

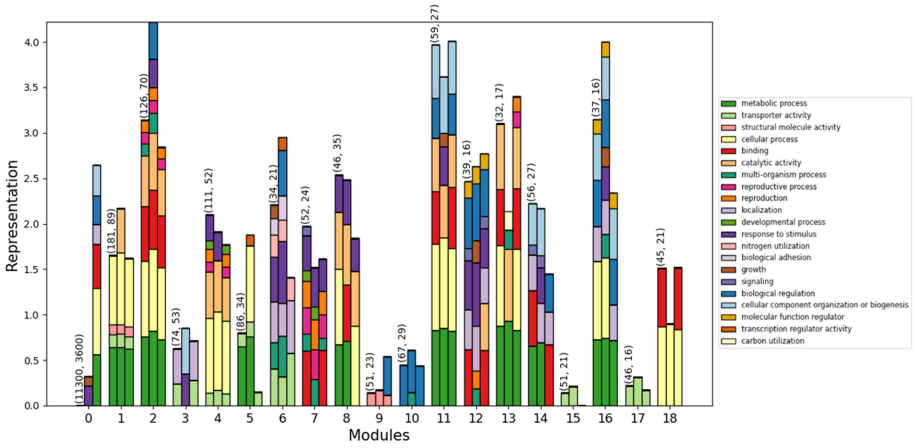


Fig. 5. Comparison of labels between the modules of the multilayer and the respective groups of genes from *S. cerevisiae* and *C. albicans*. The three bars side-by-side respectively describe the labels of the module, of the genes from *S.cerevisiae* and the genes from *C. albicans*. At the top of the first bar of each module is shown the module size and the number of inter-layer links in the module.

Using *S. cerevisiae* and *C. albicans*, we create the multilayer network. From the detection of modules, we could find 19 modules containing genes from both species. Going further with our analysis, we study the contribution of the genes of each species for the classification of the modules in the multilayer network. The comparison between the functions of each module and those of the respective gene groups can be seen in Fig. 5.

We observe that the first module represents a large part of the multilayer network, suggesting that this one may not provide useful information about small functional elements conserved between species. Looking at the classification of this module, we confirm that this one does not have GO terms associated with it, not contributing with relevant results for the analysis. Regarding the rest of the modules, we verify that the number of pairs of homologous genes corresponds to about half the module size, indicating that these modules are mostly composed of pairs of homologous genes. We verify that in some modules the functionalities result from the mutual contribution of the species, such as in *M1*, *M2*, *M4*, *M7*, *M8*, *M11*, *M12*, *M13* and *M16*. These modules result from the combination of homologous genes that are functionally identical and that constitute functional structures conserved among species. Some functions in the modules are equally represented, such as the metabolic and cellular process in *M1* or reproduction and reproductive process in *M7*. This is a consequence of these functions being associated to the same set of genes.

4 Conclusions

In this work, we contribute with relevant information about transcriptional regulatory networks of the considered yeast species. From the algorithms used in the detection of modules, the methods based on optimization of the modularity achieved a better performance. Of these, we highlight Leiden, which best managed to combine a balanced division of modules with a good functional classification. The functional characterization of modules revealed that there are biological functions more represented than others among modules of different species. From these processes, we highlight the metabolic process, cellular process, biological regulation, or response to stimulus. Furthermore, we observed that in species *C. glabrata*, although it has less genetic evidence, it was possible to detect a greater diversity of functions than in species *C. parapsilosis* and *Y. lipolytica*. The transcription factors are the main agents responsible for regulating the behavior of species, the set of interactions between these and their target genes constitute the regulation of certain behaviors. Since *C. glabrata* contains more transcription factors, it contains genetic evidence about more functions.

The cross-species comparison allowed us to establish some similarities between species. As an example, we found that modules from different species contain identical functions due to the presence of functionally identical homologous genes between them. With the creation of the multilayer, we were able to verify that there are functional structures conserved across species that carry identical genetic information.

We highlight some limitations of our approach. Firstly, the difficulty of finding meaningful terms with the p-value approach in large modules. Therefore, in an unbalanced division of the network, it will be difficult to label the large modules. Secondly, the threshold used to consider a term as relevant in a module (10%) may be too restrictive. To overcome this problem, we could test different values for the threshold in a set of modules with different sizes. Then, we could use the relation between the threshold values and the size of the modules to predict the threshold values for each module considering its size.

As future work, we could consider the creation of a measure that would allow us to evaluate the functional characterization of the modules. This one 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, we would like to consider subprocesses, i.e., GO terms at a level greater that 3, to uncover specific regulatory processes within the identified modules. Also, we found some genes in modules not associated with any Gene Ontology terms, we could use the functions of the modules in which these genes belong to predict their functionality.

Acknowledgements. This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references PTDC/BII-BIO/28216/2017 and PTDC/CCI-BIO/29676/2017, UIDB/50021/2020 and UIDP/00408/2020 (INESC-ID and LASIGE multi-annual funding, respectively).

References

1. Davidson, E.H., et al.: A genomic regulatory network for development. Science **295**(5560), 1669–1678 (2002)
2. Luscombe, N.M., Madan Babu, M., Yu, H., Snyder, M., Teichmann, S.A., Gerstein, M.: Genomic analysis of regulatory network dynamics reveals large topological changes. Nature **431**(7006), 308–312 (2004)
3. Rives, A.W., Galitski, T.: Modular organization of cellular networks. Proc. Natl. Acad. Sci. **100**(3), 1128–1133 (2003)
4. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. **100**(21), 12123–12128 (2003)
5. Wilkinson, D.M., Huberman, B.A.: A method for finding communities of related genes. Proc. Natl. Acad. Sci. **101**(suppl 1), 5241–5248 (2004)
6. de Anda-Jáuregui, G., Alcalá-Corona, S.A., Espinal-Enríquez, J., Hernández-Lemus, E.: Functional and transcriptional connectivity of communities in breast cancer co-expression networks. Appl. Netw. Sci. **4**(1), 1–13 (2019). https://doi.org/10.1007/s41109-019-0129-0
7. Matthews, L.R., et al.: Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res. **11**(12), 2120–2126 (2001)
8. Sharan, R., et al.: Conserved patterns of protein interaction in multiple species. Proc. Natl. Acad. Sci. **102**(6), 1974–1979 (2005)
9. Borneman, A.R., et al.: Divergence of transcription factor binding sites across related yeast species. Science **317**(5839), 815–819 (2007)
10. Stuart, J.M., Segal, E., Koller, D., Kim, S.K., et al.: A gene-coexpression network for global discovery of conserved genetic modules. Science **302**(5643), 249–255 (2003)
11. Zhao, B., Sai, H., Li, X., Zhang, F., Tian, Q., Ni, W.: An efficient method for protein function annotation based on multilayer protein networks. Hum. Genomics **10**(1), 1–15 (2016)
12. Cantini, L., Medico, E., Fortunato, S., Caselle, M.: Detection of gene communities in multi-networks reveals cancer drivers. Sci. Rep. **5**(1), 1–10 (2015)
13. Yu, L., Shi, Y., Zou, Q., Gao, L.: Studying the drug treatment pattern based on the action of drug and multi-layer network model. bioRxiv, p. 780858 (2019)
14. Monteiro, P.T., et al.: YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. Nucleic Acids Res. **48**(D1), D642–D649 (2019)
15. Ashburner, M., et al.: Gene ontology: tool for the unification of biology. Nature Genet. **25**(1), 25–29 (2000)
16. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
17. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. **2008**(10), P10008 (2008)
18. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**(6), 066111 (2004)
19. Traag, V.A., Waltman, L., Van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. **9**(1), 1–12 (2019)
20. Donetti, L., Munoz, M.A.: Detecting network communities: a new systematic and efficient algorithm. J. Stat. Mech. Theor. Exp. **2004**(10), P10012 (2004)

21. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. **105**(4), 1118–1123 (2008)
22. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**(3), 036106 (2007)
23. Van Dongen, S.M.: Graph clustering by flow simulation. PhD thesis, Faculteit Wiskunde en Informatica, Universiteit Utrecht (2000)
24. Adamcsek, B., Palla, G., Farkas, I.J., Derényi, I., Vicsek, T.: Cfinder: locating cliques and overlapping modules in biological networks. Bioinformatics **22**(8), 1021–1023 (2006)
25. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**(3–5), 75–174 (2010)
26. Gates, A.J., Ahn, Y.Y.: Clusim: a python package for calculating clustering similarity. J. Open Source Softw. **4**(35), 1264 (2019)
27. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971)
28. Lancichinetti, A., Radicchi, F., Ramasco, J.J.: Statistical significance of communities in networks. Phys. Rev. E **81**(4), 046110 (2010)