

RESEARCH

# Guideline for comparing functional enrichment of biological network modular structures

Guillermo de Anda-Jáuregui\*

Correspondence:  
gdeanda@inmegen.edu.mx  
Computational Genomics,  
Instituto Nacional de Medicina  
Genómica, Periférico Sur 4809,  
Ciudad de México, México  
Full list of author information is  
available at the end of the article  
\*Corresponding author

## Abstract

The use of networks to analyze biological data, such as large gene or protein expression datasets, is on the rise. Often, there is an interest of identifying modules (or communities) of biological molecules that may be associated to known functions. This functional modularity analyses usually revolve around a workflow that combines 1) a method for network reconstruction from biological data, 2) a community or clustering algorithm on a network, and 3) an enrichment analysis to associate modules to known biological categories. With this, it is possible to identify sets of functions associated to modules in networks of distinct biological conditions, allowing for the comparison of such different phenotypes.

Currently there is no set of recommendations for such analyses, which can lead to problems in assessing these results for a given biological context. Furthermore, without properly identifying the methodological scopes and limitations at each stage for a given functional modularity analysis, it is not immediately possible to compare the biological implications of analyses in different phenotypes.

In this work, critical points in a functional modularity analysis for biological networks are identified, and methods are proposed for assessing the topological and biological results of functional modularity analyses in biological networks, and to calculate topological and functional similarity between comparable phenotypes. These methods are demonstrated on biological networks artificially constructed from known biological pathways.

**Keywords:** Communities; Modularity; Biological Networks; Functional enrichment

## Background

The emergence of high-throughput technologies for the study of biological systems has lead to the use of several data-centric strategies for their study. Among these, techniques derived from network science are increasingly being adopted [1, 2, 3]. A well described property of biological systems is their modular structure[4], with different biological functions being controlled by different sets of molecular interactions [5].

As biological network analysis becomes an increasingly important tool in the study of biological systems, the detection of modular structures in this networks, and *module enrichment* (the association of these modules or communities to known biological functions) becomes more commonplace [6, 7, 8, 9]. A PubMed query for networks, clusters, modules, and enrichment returns over 900 hits, with 222 from the year 2018 alone (see Additional File 1 for an illustration). As this trend continues,

it will be important for biological and biomedical researchers to identify critical points in the workflows used for such analyses.

#### Critical points in a typical modular analysis of a biological network

There are many different methodological strategies available for the construction of networks from biological data, for the identification of biological communities or modules in these networks, and for the association of known biological functions to these modules of biomolecules). In general, these analysis pipelines will consist of three steps:

- 1 A methodology for network construction: these include the integration of known biomolecular interactions from available literature [10] or public databases such as The Kyoto Encyclopedia of Genes and Genomes (KEGG) [11] or the Search Tool for the Retrieval of Interacting Genes database (STRING-db) [12], or probabilistic methods of gene or protein co-expression inference [7, 13, 14].
- 2 An algorithm for the identification of communities or modules in the network: there is an abundant body of published algorithms for the detection of communities in large complex networks based on different approaches [15].
- 3 An enrichment analysis to associate the modules to known biological processes or functions: There is a variety of methods using different approaches [16] that may be used to associate, the sets of biomolecules that form each module to sets of biomolecules that are known to be involved in biological functions, such as those described in databases like the aforementioned KEGG or the Gene Ontology [17].

With many different combinations of methodologies available for these module enrichment studies, it becomes difficult to assess whether the results obtained through one pipeline are comparable to results obtained from another. In this work, we will find parameters that allow for the quantitative comparison of modular structures of biological networks from a topological and functional perspective.

## Data

In this work, we use four different biological networks, with nodes representing genes. These networks are constructed by merging different pathways obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [11] into undirected, unweighted graphs. A visualization of this networks is found in [figure 2](#). [Figure 1](#).

We will describe different measures to compare these networks and their modular structures from a topological and functional perspective. For this, we will identify each network as follows:

- 1 The first network is the *Main network*. It is constructed by merging 5 pathways (as mentioned in Methods). This will be the network against which other networks will be compared.

- 2 The second network is the *Alternative network*. It was constructed from pathways with almost no overlap to the ones used to build the Main network.
- 3 The third network is the *Mixed network*, constructed from two pathways from Main and two pathways from Alternative.
- 4 The last network is a *Rewired network* derived from the Main network.

### Modular structure comparison pipeline

Consider a pipeline for the comparison of modular structures in biological networks, containing the aforementioned critical points. In [figure 1](#)[Figure 2](#), a graphical representation of this proposed pipeline is shown, indicating the quantitative parameters that may be used for the comparison of networks at each level. In what follows, we will explore each of these comparison levels, using a set of artificial networks generated from known biological pathways, which attempt to represent likely structures of networks generated from biological data.

### Network topological comparison

The first critical step is to assess the comparability of biological networks. To do so we must describe the topological features of these networks. Since a common use-case for network analyses in the biological sciences is data exploration, these descriptors are readily available. Importantly, these features will be directly related to the way the networks are originated: whether they are curated from literature, inferred using a probabilistic approach from high throughput data, reconstructed through experimental methods, etc. The decision on whether networks are comparable based on their structural patterns can only be assessed based on the biological question of interest. For instance:

- Two sets of genes associated to two different biological conditions are used as the bases for network construction; these gene sets may have different sizes, and therefore the number of nodes will be different.
- A method for the inference of whole-genome co-expression network is used for two different biological conditions; the set of associated co-expression relationships may differ between the two conditions.

A researcher shall consider, based on their originating data and the biological question at hand, at what level the networks are comparable. Furthermore, the results and biological insights derived from these networks shall be explicitly described in this context.

**Table 1 Basic Network Descriptors**

	Main	Alternative	Mixed	Rewired
No.Nodes	276	228	189	276
No.Edges	692	723	596	692
No.Components	1	11	3	3

By construction, we generated two networks (Alternative and Mixed networks) that are different from our Main Network in terms of number of nodes and edges. Meanwhile, we generated a Rewired network that has the same number of both nodes and edges. Some basic descriptors of these networks are found in Table 1.

**Table 2 Node similarity, compared to the Main network**

	Jaccard Index, nodes
Main	1.00
Alternative	0.03
Mixed	0.32
Rewired	1.00

Furthermore, none of the networks have the same degree distribution, as illustrated in Figure 3.

For the purposes of functional comparability of nodes, perhaps the most immediate parameter to consider is the similarity in number of nodes, as this will be a decision point in further comparison. Differences in the number of edges are however, quite important in the context of module analysis, as these modules are defined not only by nodes but by the connections between them. This consideration could be extended to the influence of a difference in component number, as a network with more than one component will, by definition, have at least one pair of nodes without a linking path.

The biological network structures associated to different phenotypes will hopefully reflect the differences between the underlying biological phenomena. The modular structure will inherit from these network structures. This shall be kept in mind while discussing the results of further analyses.

#### Node set identity comparison

It is important to assess not only the topological similarity of networks, but whether the biological elements that are represented in the networks are comparable. In order to do this, a similarity measure is used to compare the identity (ie, labels) of nodes in different graphs. For instance, we could use the Jaccard index ( $J$ ), as seen in Table 2. [Jaccard index and other measures used throughout the manuscript are defined in the \*Methods\* section.](#)

The rewired network has exactly the same set of nodes representing genes, as it was generated by reconnecting the Main network, whereas the others are dissimilar to the original one. It can be noticed that the alternative network is not completely dissimilar to the original one (as one may naively guess), since there is a small overlap in the genes composing the pathways used for the construction of the Main and alternative networks (13 genes in total). Edge similarity can also be evaluated, as seen in [table-Table 3](#).

**Table 3 Edge similarity, compared to the Main network**

	Jaccard Index, edges
Main	1.00
Alternative	0.00
Mixed	0.26
Rewired	0.37

Again, a similarity threshold for nodes and edges must be defined in terms of the biological question that is being studied. For instance, a biological question in itself could be if networks reconstructed from two different conditions using the

same methodology is composed of the same biological molecules. In other cases, it is necessary for the networks to be composed by the same set of nodes.

For the purposes of module comparison, whether the node set of each network to be compared is a critical point that defines which methods are available to compare the modular structures of two different networks.

### Comparison of modular structures

Once the structural comparability of the networks has been assessed, it would be possible to analyze and compare the modular structures that are identified.

#### Election of the module detection algorithm

Selecting the module detection algorithm for a given analyses is an open problem [18]. Ideally, the selected algorithm would be able to recover an organization of the constituent elements in the network that has its origin in a biological phenomenon. Assessing how suitable an algorithm is would then consist on evaluating how well it recovers such structures. Unfortunately, in the case of biological networks such *ground truth* is seldom available.

Comparing partitions generated by an algorithm to the ground truth organization is a problem that can be approached with different well described techniques. Several methods have been developed and implemented: these include Variation of Information [19], Normalized Mutual Information [20], Adjusted Mutual Information [21], split-join distance [22], and (adjusted) Rand index [23, 24]. A description of these methods is beyond the scope of this manuscript (see [25] and [26]), but are noted for such cases in which such *ground truth* was available. When such information is unavailable, it will be necessary to resort to heuristics, or to identify consensus between algorithms.

For this work, different algorithms were tried and the partitions generated compared to each other (see Additional file 2, in which different module detection algorithms are compared using Adjusted Mutual Information). For simplicity, throughout the manuscript the results of the Infomap algorithm will be used. The decision to use this algorithm was arbitrary, since the intention is to illustrate the critical points of the functional module analysis workflow.

#### Evaluation and comparison of modular structures

Having identified modular structures in each biological network, the question is how similar are these modular structures to each other. A first approach is to obtain and compare descriptors of these modular structures. The modularity value  $Q$  [27, 28] measures whether a proposed division of a network is adequate capturing communities, in the sense of recovering groups with a higher proportion of edges within the group than to other groups.

In the case of comparing different biological networks and their associated modular structures, by comparing the value of  $Q$  for each network partition it is possible to assess if there is one network in which the module detection algorithm, applied

to each network, identified a more modular structure. In Table 4, the results for the example networks are shown.

**Table 4 Modularity of the Infomap Partition**

	Q
Main	0.75
Alternative	0.85
Mixed	0.77
Rewired	0.48

The measure of modularity  $Q$  is readily available in network analysis packages. While widely used, there are limitations in the use of  $Q$  as a sole measure of the modular structure of a network identified by a module detection algorithm: particularly, it should be noted that since this metric is used as an optimization criterion for certain module detection algorithms, it could bias comparisons [18]. Further descriptors should be used in order to more adequately capture the differences in modular structures.

The node similarity of the networks is critical to decide how to compare modular structures. The methods available for networks with identical node sets will be different to those for networks composed of non-identical node sets.

#### *Comparison of modular structures in networks with identical node sets*

If the node sets of the networks to be compared are identical, then the problem of comparing their modular structures is not entirely unlike the previously described problem of comparing modular partitions on the same network derived from different algorithms: instead of comparing a partition of graph  $G_1$  obtained by algorithm  $A$  with the partition of  $G_1$  obtained by algorithm  $B$  (or perhaps to a *ground truth* partition), it would be a comparison of the partitions obtained using algorithm  $A$  of graphs  $G_1$  and  $G_2$ .

To do such comparison, the methods previously mentioned when discussing the election of module detection algorithms are available. For the purposes of this manuscript, two of the constructed networks were generated to satisfy the condition of containing the same set of nodes: the Main and Rewired networks. In [table-Table 5](#), the use of the different comparison methods is illustrated, as applied to these networks.

**Table 5 Comparison of modular structures detected using the Infomap algorithm Main vs Rewired networks**

	Method	Value
	variation of information	2.28
	normalized mutual information	0.63
	split-join distance	222.00
	Rand index	0.93
	adjusted Rand index	0.34

By using these results it is possible to provide an overall descriptor of how different is the organization of elements in the network, derived from the differences in network structure, as identified by the module detection algorithm. However, it is important to point out that the methods previously described consider only

the *membership of nodes* into different modules. Since the module structure in this networks contains also topological information, it is possible to approach this comparison from a topological perspective.

To illustrate such comparison, three topology-based comparisons between the Main and Rewired networks are presented. The first one consists on comparing the distribution of the module sizes identified in each network. The second is the comparison of the distribution of embeddedness [29], a measure of the amount of neighbors of each node which belong to its own module. A third one is the comparison of the distribution of modular degree [30], which extends on the classical notion of degree centrality to describe both the local and global influence of a node in the network. In ~~figure~~ Figure 4 we observe that, based on any of these parameters, the modular structure of each network exhibits differences.

It should be noted that the topology-based comparisons presented here are only a few of the possible options to incorporate this dimension of analysis. In [26] a more extensive discussion on the subject is found. Nevertheless, the reader should consider using such level of description to more thoroughly evaluate (and convey) the differences in modular structures, and the implications regarding the differences in the underlying biological phenomena.

#### *Comparison of modular structures in networks with non-identical node sets*

In the case of networks where node sets are not identical, it may not be feasible to compare using the aforementioned metrics, as the algorithm implementations are written with identical node sets in mind.

Let us focus on the similarity of modules in each network, in terms of nodes and edges. A matrix of a similarity measure (such as Jaccard index) may be computed, in which each module of the first network is compared to each module of the second network. We provide such matrices for the comparison of the Alternative and Mixed networks against the Main network as Additional File 3. Similarity matrices allow to answer the following questions:

- Whether there are modules in the first network that have some overlap with modules in the second network.
- Whether there are modules in the first network that are perfect matches to modules in the second network.

A global parameter of similarity can be defined to have a descriptor for each network. In this work, we have defined a Similarity Score in terms of the similarity matrix for each pair of networks (see Methods).

In ~~table~~ Table 6 we find the comparison of the Alternative or Mixed networks against the Main network in terms of module node composition. It shows the number of modules in Main that had at least one non-zero overlap with module in the test networks, and the number of modules of Main that perfectly match with a module in the test networks.

In ~~table~~ Table 7 we find again a comparison of the Alternative or Mixed networks against the Main network, this time in terms of edge composition. We may notice

**Table 6 Comparison of modular structures detected using the Infomap algorithm between the Main network and the Alternative or Mixed networks, based on module node similarity**

	Similarity_Score	Matching.Modules	Perfect.Matches
Alternative	0.020	7	0
Mixed	0.377	11	9

that in terms of edges, the similarity between the Alternative and Main network is much more penalized.

**Table 7 Comparison of modular structures detected using the Infomap algorithm between the Main network and the Alternative or Mixed networks, based on module edge similarity**

	Similarity_Score	Matching.Modules	Perfect.Matches
Alternative	0.001	1	0
Mixed	0.367	11	9

By comparing the similarities of these partitions, we are able to identify how much are the biological differences of each condition being captured in the modular structures of networks. Furthermore, we gain further information for the discussion of the functional associations identified for each module: for instance, whether functions are being associated to similar or dissimilar modules. The comparison of these modules at the functional level is the topic of the next section.

## Functional comparison of modules

Now the modules will be compared in terms of the associated biological functions identified by an enrichment analysis. This comparison involves answering three complementary questions:

- How similar are the sets of biological functions that are associated to the whole network, through the enrichment of individual modules?
- How similar are the modules found in each network, in terms of the sets of associated biological functions?
- In how many modules is represented each biological process?

When comparing these biological processes, it is important to remember that it is possible for networks composed of different genes to be associated to the same processes and functions. In other words, networks could be functionally comparable even if their node identities are dissimilar. On the other hand, if the genes that integrate the biological functions being evaluated are not present in the networks, these biological functions will not be identified by an enrichment analysis.

### Comparison of biological function sets associated to the overall network

The sets of all biological functions identified in each network are to be compared in terms of similarity. By analyzing the sets of identified biological functions for each network, as shown in Table 8, we can compare and contrast which functions are shared by both networks or uniquely found in each. From the biological perspective, this provides a good starting point to discern which functions or processes may be associated to each condition of interest.

While this comparison is useful as a first approximation, it is soon evident that this is not necessarily considering the modular nature of the network. As such, we will now perform a comparison at the level of individual modules.



**Table 8 Similarity of enriched processes associated to the overall network**

Similarity of Enriched Processes , Jaccard Index	
Main	1.00
Alternative	0.50
Mixed	0.30
Rewired	0.88

Similarity of modules in terms of associated biological functions

It is possible to compare each module of the network of interest to the modules of another network, in terms of their associated biological functions. A similarity matrix can be calculated for the pair of networks, allowing the identification of modules that are functionally similar in each network, as seen in Table 9.

**Table 9 Number of modules in networks that have Jaccard index  $J = 1$  with a module of the Main network**

Module of Main Network	Alternative														15-Mixed	16-Rewired		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16		
1							1	7	0						19	3	20	2
Alternative 2							0								0		0	
3							0								0		0	
4							0								0		0	2
5							0								0		0	1
6							0								0	2	0	3
7							0								0	2	0	3
8							0	Mixed							2		3	
9							0								0		0	1
10							0								2	0	2	1
11							2	0							0		0	
12							0								3		3	2
13							0								0	3	0	2
14							0								3	0	0	
Rewired 2-15							0								0		2	1
16							3	3	0						3	0	1	
17							1	0							0		2	1
18							2	0							0		1	0
19							1	0							1	3	2	
20							0								0		1	

It is important to consider that, since more than a biological function could be associated to more than one module in a network, it is possible to find, for a given module in a network, more than one functionally similar module in another network.

Number of modules associated to a given biological function

As it has been mentioned before, it is possible for more than one module to be associated to a given biological function. Therefore, another comparison parameter that can be used to assess the similarities between two biological networks is to evaluate, for each identified biological function, the number of modules that were

found to be enriched in that particular function.

It is possible to calculate a distance (for instance, Euclidean) between two networks based on the number of modules associated to each biological function found in the networks, as it is shown in Table 10.

**Table 10 Functional Euclidean Distance of network modules**

	Functional Euclidean Distances
Main	0.00
Alternative	10.20
Mixed	8.77
Rewired	6.00

### Assessing similarity of modular structures

The final integration of all comparison parameters in order to assess the similarity of the modular structures in biological networks must be guided by the research question that is to be answered by this comparison. In other words, it would not be possible to establish general thresholds for what is to be accepted as similar or dissimilar.

Furthermore, the objective of the comparison will determine the weight given to a parameter for the definition of comparability. In Table 11, we observe the overall comparison of our three test networks against our Main network. As it has been shown throughout the manuscript, it is possible to determine a quantitative (or semi-quantitative) value of comparison for each network, which can be informative of the points of similarity and divergence between the networks and their associated modular structures.

The alternative network is very dissimilar in terms of node composition. We observe, however, that even then there is some similarity in terms of the overall functions associated to the network. Nevertheless, we find that there are no modules in the Alternative network that precisely reflect the functionality associated to the modules in the Main network. Meanwhile, the mixed network has some overlap in terms of node and edge composition. It exhibits a degree of module topological similarity, and is also functionally closer to the Main network. In relative terms, it is possible to establish that the Mixed network is both topologically and functionally closer to the Main network than the Alternative, which is consistent with the construction of these networks.

The Rewired and Main networks are identical in terms of number of nodes and edges, as well as the identity of nodes, by construction. They were also guaranteed by construction to be different in terms of edges. Since these networks were composed by the same set of nodes, it was possible to compare their modular structures using a well defined parameter of similarity, such as Normalized Mutual Information. Furthermore, we observe that these two networks are very similar in functionality terms.

**Table 11 Comparison of the modular structure of the Main network**

Parameter	Alternative	Mixed	Rewired
No. Nodes	Lesser	Lesser	Equal
No. Edges	Greater	Lesser	Equal
Node Similarity (J)	0.03	0.32	1.00
Edge Similarity (J)	0.00	0.26	0.37
Modularity (Q)	Greater	Greater	Lesser
Module Topological Similarity	$SimScore_{edge} = 0.01$	$SimScore_{edge} = 0.30$	$nmi = 0.63$
Network Functional Similarity (J)	0.50	0.30	0.88
Main modules with functionally similar counterparts	0/20	7/20	15/20
Functional Euclidian Distance	10.20	8.77	6.00

### *On the subject of biological comparison versus validation*

An open question remains on how to decide whether a network model is comparable to another network model for the purposes of modular structure comparison. As we have discussed, the two most common situations where a comparison would be necessary are A) the comparison of biological conditions and B) the validation of a biological network model.

Consider that networks that are derived from experimental datasets may or may not be necessarily composed of the same set of nodes, either by technical or biological conditions. In this case, if the experimental settings of each condition are comparable, and the methods for network inference, module detection, and enrichment analysis used in each case are the same, then a comparison in terms of functional similarity is feasible. This setting could also be used for the validation of functional findings, for instance if using a discovery and validation set. An example of such use-case is found in [31].

Meanwhile, if the objective is the validation of any of the steps of the modular analysis pipeline, it would be important to evaluate the comparability of modular structures and functional enrichment analyses in the context of a network with the same set of nodes and number of edges. Importantly, if a rewiring algorithm is used, the limitations of the implementation should be discussed; for instance, the rewiring algorithm used for the generation of the Rewired network in this work does not preserve the degree distribution, and as such it would not be an adequate model to generate null models for the assessment of functional associations.

## Conclusions

The study of modularity in biological networks provides opportunities to understand the organization of biological phenomena and how these structures shape functionality. In this work, we provide a guideline to quantitatively compare modular structures in biological networks, in topological and functional terms. This provides a basis to identify aspects of network modular structures that may guide the discussion regarding phenotype comparison from a network perspective, as well as some critical points for the validation of network models.

## Methods

All methods used for this work are available at <https://github.com/guillermodeanda.jauregui/BiologicalModuleComparison>

### Network Generation

We generated four networks by acquiring network representations of pathways in KEGG using the Graphite package [32]. For each of the networks Main, Alternative, and Mixed, the pathways listed in Table 13-12 were used. In each case, pathways were merged into a single undirected, unweighted network to be analyzed using Igraph [33].

**Table 12 Source pathways for the model networks**

Main	Alternative	Mixed
Hedgehog signaling pathway	Pentose phosphate pathway	Hedgehog signaling pathway
NF-kappa B signaling pathway	Notch signaling pathway	Notch signaling pathway
VEGF signaling pathway	mRNA surveillance pathway	Pentose phosphate pathway
p53 signaling pathway	TGF-beta signaling pathway	p53 signaling pathway
RIG-I-like receptor signaling pathway	IL-17 signaling pathway	

The Rewired network was generated by taking the Main network and using a rewiring algorithm as implemented in Igraph [33], in which the endpoints of edges from network 1 were reconnected randomly, with a uniform rewiring probability of 0.25.

### Module detection

Modules for the four networks were detected using the Infomap algorithm [34] as implemented in Igraph. Additionally, modules were detected for the Main network using the Girvan-Newman edge-betweenness algorithm [35], the Fast-Greedy algorithm [28], the Louvain method [36], the Walktrap method [37], the spin-glass method [38], the leading eigenvector method [39], and the label propagation method [40]. [A comparison of these algorithms, including a description of each along with suggestions for algorithm selection may be found in \[41\]; further discussion may be found in \[42\].](#)

### Module enrichment

Over-Representation Enrichment analyses for the gene sets of each detected module was performed using hypergeometric testing. This test is equivalent to the one-tailed Fisher's exact test, which assesses the probability of drawing  $k$  elements belonging to a set of  $K$ , by drawing  $n$  elements out of a population  $N$ .

It is widely used as a gene set enrichment tool for gene clustering methods, where the genes in the cluster (or module) form the set of  $n$  elements drawn from the  $N$  population (usually, the whole set of measured genes). The genes that belong to a given pathway or biological function gene set represent  $K$ . Finally, the members of the cluster that belong to the pathway or biological function are represented by  $k$ .

The testing was performed using as implemented in the HTSanalyzer package [43]. The significance threshold for enrichment was set to be a Benjamini-Hochberg adjusted  $p$  - value  $< 0.05$ . For the sake of simplicity in the analysis, it was decided that the list of pathways used for the enrichment analyses would consist of the 10 pathways used to construct networks 1 and 2, plus two additional pathways "Sphingolipid signaling pathway", and "Insulin signaling pathway" that were not originally used for the construction of any network.

### Topological network comparison

The networks were compared in terms of the number of nodes, edges and connected components that composed them. The degree distribution of each network was also obtained. Finally, the similarity of node sets and edge sets of the networks were compared using the Jaccard index.

#### *Jaccard index*

The Jaccard index is a measure of similarity between sets. It considers the sizes of the intersection and union of two sets  $A$  and  $B$ , as follows:

$$J = \frac{|A \cap B|}{|A \cup B|}$$

### Modular structure comparison

#### *Evaluation of modular structures*

The modularity score  $Q$  [28] was calculated as implemented in the igraph package for R. The topological measures of embeddedness [29] and modular degree [30] were implemented following their descriptions in the original references, using the igraph package for R.

#### *Networks with identical node sets*

The modular structures of networks composed of identical node sets were compared using the modularity comparison methods implemented in Igraph. These included the Variation of Information [19], Normalized Mutual Information [20], split-join distance [22], and (adjusted) Rand index [23, 24].

#### *Networks with non-identical node sets*

For a pair of networks  $G_1$  and  $G_2$ , with a modular partition  $M_1$  and  $M_2$ , we calculate a similarity matrix *SimMatrix*. For each pair of  $M_i$  in  $G_1$  and  $M_j$  in  $G_2$ , each value  $SimMatrix(i, j) = Similarity(M_i, M_j)$ .

In this work, the measure of similarity will be the Jaccard index.

We define a Similarity Score in terms of the similarity matrix for each module  $M_i$  in  $G_1$  such that

$$SimScore(M_i) = \sum_{j=1}^q \frac{J(M_i, M_j)}{|(M_i \cap M_j) \neq \emptyset|},$$

with  $M_i = 0$  if  $|(M_i \cap M_j) \neq \emptyset| = 0$ .

Furthermore, we then define a global Similarity Score as

$$\frac{\sum_{i=1}^k SimScore(M_i)}{k} = \frac{1}{k} \sum_{i=1}^k SimScore(M_i).$$

### Functional comparison

#### *Network functional similarity*

Consider  $Functions_G$  to be the set of all biological functions associated to each module of a network  $G$ . Functional similarity of two networks is to be calculated as  $FunctionalSimilarity = J(Functions_{G_1}, Functions_{G_2})$

### *Functional Similarity of Modules*

Consider  $Functions_i$  to be the set of biological functions associated to each module  $M_i$  in a network  $G_1$ , and  $Functions_j$  to be the set of biological functions associated to each module  $M_j$  in a network  $G_2$ . A Functional Similarity Matrix is defined such that  $FuncSimMatrix(i, j) = J(Functions_i, Functions_j)$

### *Number of modules associated to a given biological function*

Consider  $Functions_{G_1}$  and  $Functions_{G_2}$  to be the set of functions associated through enrichment to  $G_1$  and  $G_2$  respectively. Let  $Functions_{both} = Functions_{G_1} \cap Functions_{G_2}$ . Let  $E_1$  and  $E_2$  be  $k$ -dimensional vectors where each element  $k$  of  $E_1$  and  $E_2$  is the number of modules in  $G_1$  or  $G_2$  to which  $Functions_{both_k}$  is associated. An Euclidian distance between  $E_1$  and  $E_2$  can be calculated.

### **Declarations**

Availability of Data and Material

Code used for the generation and analysis of networks is available at <https://github.com/guillermodeandajauregui/BiologicalModuleComparison>

### Competing interests

The author declares to have no competing interests.

### Funding

No funding was allocated for this work.

### Author's contributions

GDJ developed the work in the present manuscript.

### **Acknowledgements**

The author would like to thank Enrique Hernández-Lemus and Jesús Espinal-Enríquez, for the discussions leading to this manuscript.

### **References**

- Ma'ayan, A.: Introduction to network analysis in systems biology. *Science signaling* **4**, 5 (2011). doi:10.1126/scisignal.2001965
- Dobrescu, R., Purcărea, V.: Network based models for biological applications. *Journal of medicine and life* **2**, 176–184 (2009)
- Barabási, A.-L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. *Nature reviews. Genetics* **5**, 101–113 (2004). doi:10.1038/nrg1272
- Lorenz, D.M., Jeng, A., Deem, M.W.: The emergence of modularity in biological systems. *Physics of life reviews* **8**, 129–160 (2011). doi:10.1016/j.plrev.2011.02.003
- Ames, R.M., Macpherson, J.I., Pinney, J.W., Lovell, S.C., Robertson, D.L.: Modular biological function is most effectively captured by combining molecular interaction data types. *PloS one* **8**, 62670 (2013). doi:10.1371/journal.pone.0062670
- Alcalá-Corona, S.A., de Anda-Jáuregui, G., Espinal-Enríquez, J., Hernández-Lemus, E.: Network modularity in breast cancer molecular subtypes. *Frontiers in physiology* **8**, 915 (2017). doi:10.3389/fphys.2017.00915
- Langfelder, P., Horvath, S.: Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics* **9**, 559 (2008). doi:10.1186/1471-2105-9-559
- Liu, Y., Brossard, M., Roqueiro, D., Margaritte-Jeannin, P., Sarnowski, C., Bouzigon, E., Demenais, F.: Sigmod: an exact and efficient method to identify a strongly interconnected disease-associated module in a gene network. *Bioinformatics (Oxford, England)* **33**, 1536–1544 (2017). doi:10.1093/bioinformatics/btx004
- Adamcsek, B., Palla, G., Farkas, I.J., Derényi, I., Vicsek, T.: Cfinder: locating cliques and overlapping modules in biological networks. *Bioinformatics (Oxford, England)* **22**, 1021–1023 (2006). doi:10.1093/bioinformatics/btl039
- Hur, J., Schuyler, A.D., States, D.J., Feldman, E.L.: Sciminer: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics (Oxford, England)* **25**, 838–840 (2009). doi:10.1093/bioinformatics/btp049
- Kanehisa, M., Goto, S.: Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30 (2000)
- Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J., von Mering, C.: The string database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research* **45**, 362–368 (2017). doi:10.1093/nar/gkw937

13. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A.: Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* **7 Suppl 1**, 7 (2006). doi:10.1186/1471-2105-7-S1-S7
14. Madhamshettiar, P.B., Maetschke, S.R., Davis, M.J., Reverter, A., Ragan, M.A.: Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Medicine* **4**(5), 41 (2012). doi:10.1186/gm340
15. Fortunato, S.: Community detection in graphs. *Physics Reports* **486**(3), 75–174 (2010). doi:10.1016/j.physrep.2009.11.002
16. García-Campos, M.A., Espinal-Enríquez, J., Hernández-Lemus, E.: Pathway analysis: State of the art. *Frontiers in physiology* **6**, 383 (2015). doi:10.3389/fphys.2015.00383
17. Consortium, G.O.: Gene ontology consortium: going forward. *Nucleic acids research* **43**, 1049–1056 (2015). doi:10.1093/nar/gku1179
18. Jebabli, M., Cherifi, H., Cherifi, C., Hamouda, A.: Community detection algorithm evaluation with ground-truth data. *CoRR abs/1711.09472* (2017). 1711.09472
19. Meilă, M.: Comparing clusterings by the variation of information. In: Schölkopf, B., Warmuth, M.K. (eds.) *Learning Theory and Kernel Machines*, pp. 173–187. Springer, Berlin, Heidelberg (2003)
20. Danon, L., Duch, J., Diaz-Guilera, A., Arenas, A.: Comparing community structure identification. doi:10.1088/1742-5468/2005/09/P09008. cond-mat/0505245v2
21. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In: *Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09*, pp. 1073–1080. ACM, New York, NY, USA (2009). doi:10.1145/1553374.1553511. <http://doi.acm.org/10.1145/1553374.1553511>
22. Dongen, S.: Performance criteria for graph clustering and markov cluster experiments. Technical report, Amsterdam, The Netherlands, The Netherlands (2000)
23. Rand, W.M.: Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**(336), 846–850 (1971). doi:10.1080/01621459.1971.10482356. <https://www.tandfonline.com/doi/pdf/10.1080/01621459.1971.10482356>
24. Hubert, L., Arabie, P.: Comparing partitions. *Journal of Classification* **2**(1), 193–218 (1985). doi:10.1007/BF01908075
25. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **11**, 2837–2854 (2010)
26. Orman, G.K., Labatut, V., Cherifi, H.: Comparative evaluation of community detection algorithms: A topological approach. *CoRR abs/1206.4987* (2012). 1206.4987
27. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004). doi:10.1103/PhysRevE.69.026113
28. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. doi:10.1103/PhysRevE.70.066111. cond-mat/0408187v2
29. Lancichinetti, A., Kivela, M., Saramaki, J., Fortunato, S.: Characterizing the community structure of complex networks. *PLOS ONE* **5**(8), 1–8 (2010). doi:10.1371/journal.pone.0011976
30. Ghalmane, Z., Hassouni, M.E., Cherifi, C., Cherifi, H.: Centrality in modular networks. *CoRR abs/1810.05101* (2018). 1810.05101
31. Alcalá-Corona, S.A., Espinal-Enríquez, J., de Anda-Jáuregui, G., Hernández-Lemus, E.: The hierarchical modular structure of her2+ breast cancer network. *Frontiers in physiology* **9**, 1423 (2018). doi:10.3389/fphys.2018.01423
32. Sales, G., Calura, E., Cavalieri, D., Romualdi, C.: graphite - a bioconductor package to convert pathway topology to gene network. *BMC bioinformatics* **13**, 20 (2012). doi:10.1186/1471-2105-13-20
33. Csardi, G., Nepusz, T.: The igraph software package for complex network research. *InterJournal Complex Systems*, 1695 (2006)
34. Rosvall, M., Axelsson, D., Bergstrom, C.T.: The map equation. *The European Physical Journal Special Topics* **178**(1), 13–23 (2009)
35. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. *Physical review. E, Statistical, nonlinear, and soft matter physics* **69**, 026113 (2004). doi:10.1103/PhysRevE.69.026113
36. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. doi:10.1088/1742-5468/2008/10/P10008. 0803.0476v2
37. Latapy, M., Pons, P.: Computing communities in large networks using random walks. cond-mat/0412368v1
38. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. doi:10.1103/PhysRevE.74.016110. cond-mat/0603718v1
39. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. doi:10.1103/PhysRevE.74.036104. physics/0605087v3
40. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. doi:10.1103/PhysRevE.76.036106. 0709.2938v1
41. Yang, Z., Algesheimer, R., Tessone, C.J.: [A comparative analysis of community detection algorithms on artificial networks. \*Scientific Reports\* \*\*6\*\* \(2016\)](#)
42. Poulin, V., Théberge, F.: Ensemble clustering for graphs. In: Aiello, L.M., Cherifi, C., Cherifi, H., Lambiotte, R., Lió, P., Rocha, L.M. (eds.) *Complex Networks and Their Applications VII*, pp. 231–243. Springer, Cham (2019)
43. Wang, X., Terfve, C., Rose, J.C., Markowitz, F.: Htsanalyzer: an r/bioconductor package for integrated network analysis of high-throughput screens. *Bioinformatics* **27**(6), 879–880 (2011)

#### List of abbreviations

KEGG: Kyoto Encyclopedia of Genes and Genomes STRING-db: Search Tool for the Retrieval of Interacting Genes database

**Figures**

**Figure 1** ~~Methodological pipeline for the comparison of modular structures in biological networks from a topological and functional perspective.~~  
Network visualizations. Panel A, *Main Network*. Panel B, *Alternative Network*. Panel C, *Mixed Network*. Panel D, *Rewired Network*.

**Figure 2** ~~Methodological pipeline for the comparison of modular structures in biological networks from a topological and functional perspective.~~

**Figure 3** Degree frequency for the four networks analyzed in this work.

**Figure 4** Empirical cumulative density functions (ecdf) of modular structure descriptors based on topological features, for the Main and Rewired networks. Panel A) ecdf of module size. Panel B) ecdf of node embeddedness (see [29]). Panel C) ecdf of the (tangent of the phi argument of the) modular degree, as defined in [30] )

**Additional Files**

Additional file 1 — Figure (PNG) showing the results of a PubMed search for works on networks, clustering, and enrichment, 2000-2019

Additional file 2 — Table comparing the partitions of the Main network generated by different algorithms, using Adjusted Mutual Information

Additional file 3 — Rdata file containing networks, modular structures and enrichment results