University of Verona

DEPARTMENT OF COMPUTER SCIENCE
Master's Degree in Medical Bioinformatics

MATCH&MOTIF: A CYTOSCAPE APP FOR STATISTICAL AND ANALYTICAL MOTIF SEARCH

Candidate:
Emanuele De Natale
Matricola VR403186

Thesis advisor:
Prof.ssa Rosalba Giugno

Research supervisor:
Prof. Dennis Shasha

Thesis submitted in 2017-2018

A mia madre

Ringraziamenti

Il mio percorso universitario è stato intenso, appassionante e produttivo. Non privo di difficoltà e ostacoli ma anche di soddisfazioni e traguardi. Ringrazio la mia relatrice, Rosalba Giugno, che mi ha pazientemente accompagnato e sostenuto durante il mio percorso di tesi, e tutto il gruppo di ricerca con cui ho collaborato, a cominciare da Giovanni Micale, Dennis Shasha, Vincenzo Bornici, Alfredo Pulvirenti e Fabio Rinnone. Niente sarebbe stato possibile senza il sostegno di mia madre, a cui dedico il mio lavoro di tesi, e di tutta la mia famiglia, specialmente i miei nonni, sempre presenti al mio fianco. Ringrazio anche Chiara, Alessio, Emiliano, Raffaele e tutte le altre persone importanti che mi hanno supportato quotidianamente in questi anni. Un ringraziamento particolare anche ai miei colleghi universitari, che ormai non sono più tali, bensì amici.

Grazie a tutti.

Abstract

One of the most important problems in bioinformatics and network analysis is to identify all the occurrences of a sub-graph contained within another graph. When this sub-graph is statistically overrepresented within the network, it is called motif. The problem of sub-graph matching finds application in various fields such as the analysis of the interaction between proteins, the design of new drugs and the correlation between molecular compounds. The greatest difficulties in dealing with this problem are due to the fact that the number of possible sub-graphs to search increases exponentially both as the size of this sub-graph increases and as the size of the target network increases. Different strategies have been developed for the problem of motif search, but they can be a concrete help to research only if they are used through tools implemented in a performing way. In this thesis, we will present a Cytoscape App called Match&Motif that allows the user to upload both target and query networks, draw queries and label them freely, search all the occurrences of the specified queries and finally validate the results according to two models, the statistical model and the analytical model.

Contents

List of Figures

List of Tables

Introduction

A complex system is a system composed of many interacting parts, such that the collective behaviour of those parts together is more than the sum of their individual behaviours. The collective behaviours are sometimes also called "emergent" behaviours, and a complex system can thus be said to be a system of interacting parts that displays emergent behaviour. The study of complex systems involves the analysis of the way in which their elements interact rather than only their individual roles.

The biology of complex systems regards the study of the complexity of structure and function of biological organisms, emphasising the intricate interactions and relational patterns that are essential for life. A complex system can be described as a network. Networks are a fundamental tool for understanding and modelling complex systems in physics, biology, neuroscience, engineering, and social sciences. Examples of biological and non-biological networks include transcriptional or gene regulation networks, protein–protein interaction networks (PPI network Figure 1), metabolic pathways, neural networks, social networks and technological networks.

Networks can be represented as graphs, composed by a wide variety of sub-graphs. Such sub-graphs may describe interactions between gene regulators or complex enzymatic processes, is the case of biology, or relationship among airlines and intriguing relationships involving people, in the case of non-biological problems. Different networks tend to have different sets of such frequent local structures, and each structure has a specific function in the network.

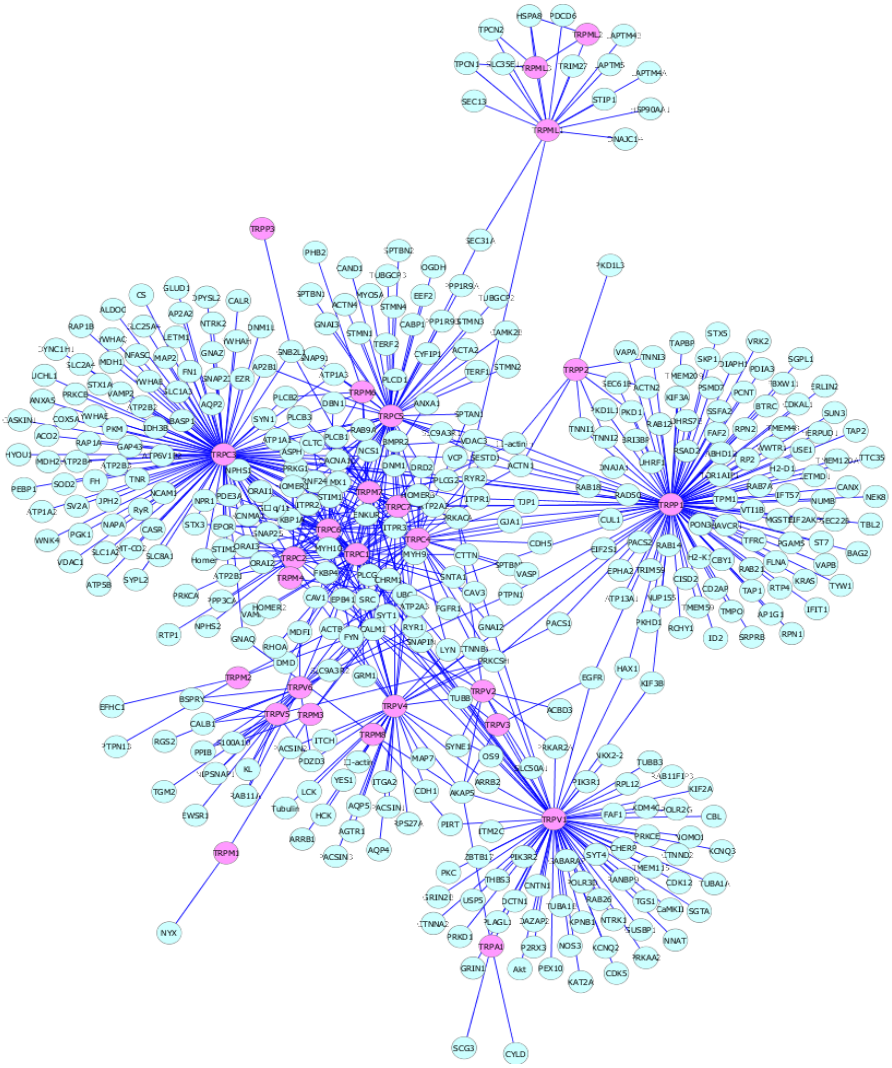

Figure 1: Example of PPI network generated with Cytoscape.

Network motifs can be defined as sub-graphs that arise unexpectedly often in a network [40]. They are recognized as 'the simple building blocks of complex networks'. Network motif search is the problem of finding sub-graphs of a network that occur more frequently than expected.

Discovering such motifs is important as it helps to understand the function of a network and also provides an alternative way to compare and analyse two or more biological networks. In fact, if two networks have common motifs, we can assume they perform some similar basic functions. Finding

similarity also helps to find phylogenetic relationships between organisms. It is interesting to know whether the functional behaviour of a motif can be predicted from its structural topology as well as whether the abundance of such motif necessarily implies biological significance. Some studies also investigated how network motifs might be shaped by evolution, in fact, when the same network is placed under varying environmental conditions where each condition demands different functional behaviour from the network, several network motifs emerge. Other studies have argued that overabundance of a network substructure might be a secondary result of some other phenomena.

The enumeration of motifs belonging to a network is a crucial point it is not only computationally heavy but also it must be taken into account that isomorphisms of the same topology must be counted only once because, in most cases, they correspond to a similar function within the network. Determining if two graphs are topologically equivalent requires graph isomorphism checking, which is a highly computation intensive problem with no known polynomial-time solution, meaning that it is an NP-Complete problem [24]. A fundamental point is to determine when a sub-graph appears unusually often within the network, so when one can refer to a motif and when it is simple sub-graph, it has no particular relevance. These problems are compounded by the fact that the number of sub-graphs of a given size in a network is exponential in both the network size and the sub-graph size and by the fact that Real-life networks tend to be large and dense. Different approaches and strategies have been developed that, depending on the use cases, find more or less efficient and precise solutions to these questions.

We will consider NetMatch* [42], a Cytoscape app [36], that allows to search all the occurrences of a query network within a target network and to check its significance with respect to some randomization models, and FlashMotif [31], which also allows to search for the occurrence of a query network within a target network, but paying attention to the labelling of the queries network and verifying the significance of the motif through an analytical model. The aim of this thesis is to unify in a cytoscape app,

which takes the name Match&Motif, the functionality of NetMatch* and FlashMotif giving the user free choice on how to verify the significance of a motif by choosing between analytical significativity or statistical significance, and, in this last case, select also the randomization model that user considers most appropriate.

In the first chapter of this thesis the main steps leading to the identification of a motif with the relative strategies will be illustrated, in the second chapter the two approaches for validation of the motif will be explored, in the third chapter the functioning of the Cytoscape App Match&Motif will be discussed with focus on the design of the query network, its labelling and the visualization of results.

Chapter 1

Preliminary definitions

In this first chapter some preliminary basic definitions will be presented concerning the theory of graphs [5, 48] and complex networks [28].

Definition 1.1. *(Graph). A graph is a pair* $G = (V, E)$ *where* $V = \{v_0, v_1, ...v_|V|\}$ *is the set of vertices (nodes) and* $E \subseteq V \times V$ *is the set of edges in* G*, that is the set of connections between the vertices of* G*.*

Definition 1.2. *(Undirected Graph). A graph* $G = (V, E)$ *is said to be non-oriented if* E *is a set of unordered pairs of vertices. Given two vertices* $v_i, v_j \in V$ *an edge of an undirected graph is indicated by* (vi, vj) *and is said incident to both vertices* vi *and* vj*.*

Definition 1.3. *(Oriented graph). A graph* $G = (V, E)$ *is called oriented if* E *is a set of ordered pairs of vertices. Given two vertices* $v_i, v_j \in V$ *an edge of an directed graph is indicated by* $\overrightarrow{(vi, vj)}$ *and is said incident from* vi *to* vj*.*

Definition 1.4. *(Multigraph). A multigraph is a graph* $G = (V, E)$ *containing multiple edges (edges connecting the same pairs of nodes) and loops (edges that connect nodes to themselves).*

Definition 1.5. *(Labelled graph). Given a graph* $G = (V, E)$*, an alphabet* $\sum$ *of labels, a function* $\alpha : V \to \sum$ *a function that assigns a label to each node*

of G and $\beta : V \to \sum$ a function that assigns a label to each edge of G, thus $G(V, E, \alpha, \beta)$ is called labeled graph.

Definition 1.6. *(Adjacency matrix). Adjacency matrix $A(a_ij)$ of a graph $G = (V, E)$ is a matrix $|V| \times |V|$ i.e.*

$$a_ij = \begin{cases} 1 & if \qquad (v_i, v_j) \in E \\ 0 & otherwise \end{cases}$$

The main diagonal of the adjacency matrix contains only values equal to zero.

Definition 1.7. *(Distances matrix). A distances matrix $D(d_ij)$ of a graph $G = (V, E)$ is a matrix $|V| \times |V|$ i.e. each element of it is equal to the distance of the minimum path between the vertices$v_i, v_j \in V$. The main diagonal of the distance matrix contains only values equal to zero.*

Definition 1.8. *(Sub-graph). A graph $G' = (V', E')$ is a sub-graph $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$*

Definition 1.9. *(Set of neighbors of a node). Give a graph $G' = (V', E')$ it is defined set of neighbors of a node $v \in V$ the set $N(G, v) = \{v' \in V : (v, v') \in E\}$. If the graph is oriented, the set of neighboring successors of a node $v \in V$ is defined as $N_{out}(G, v) = \{v' \in V : \overrightarrow{(v, v')} \in E\}$ and the set of neighbors predecessors as $N_{in}(G, v) = \{v' \in V : \overrightarrow{(v', v)} \in E\}$.*

Definition 1.10. *(Connected graph). A graph $G = (V, E)$ is called connected for each couple of distinct nodes $(v_i, v_j) \in V$ if exists a path from v_i to v_j. Otherwise the graph is said not connected.*

Definition 1.11. *(Complete graph). A graph $G = (V, E)$ is called complete for each couple of distinct nodes $(v_i, v_j) \in V$ if exists an edge between v_i and v_j.*

Definition 1.12. *(Isomorphism between graphs). Two graps $G = (V, E)$ and $G' = (V', E')$ are called isomor if there is a biunique correspondence $f : V \to V'$ i.e. $(u, v) \in E$ iff $(f(u), f(v)) \in E$. In other words it is possible to label the vertices of G with vertices of G', keeping the corresponding edges of G and of G'.*

Definition 1.13. *(Degree of a node). Given a graph $G = (V, E)$, the degree (or connectivity) k_{v_i} of a node $v_i \in V$ is the number of incident edges and is defined in terms of adjacency matrix $A(a_{ij})$ as:*

$$k_{v_i} = \sum_{v_j \in V} a_{ij}$$

If the graph is directed the indregree $k^{in}_{v_i}$ and the outdegree $k^{out}_{v_i}$ of a node $v_i \in V$ respectively $k^{in}_{v_i} = \sum_{v_j \in V} a_{ji}$ and $k^{out}_{v_i} = \sum_{v_j \in V} a_{ij}$. The total degree is defined as $k_{v_i} = k^{out}_{v_i} + k^{out}_{v_i}$.

Definition 1.14. *(Distribution of degree). Given a graph $G = (V, E)$ we diefined the distribution of degree $P(k)$ the probability that a certain randomly selected node has degree k. If the graph is direct, it is possible to define the probability distribution for the indegree and the outdegree, indicated respectively with $P(k^{in})$ and $P(k^{out})$.*

Definition 1.15. *(Local clustering coefficient). Given a graph $G = (V, E)$ and its node $v_i \in V$ we define local clustering coefficient [47] c_{v_i} of the node v_i the ratio between the number of edges $|E|$ and the maximum possible number of edges in G_i:*

$$c_{v_i} = \frac{2|E|}{k_{v_i}(k_{v_i} - 1)}$$

Definition 1.16. *(Clustering coefficient). Given a graph $G = (V, E)$ we define clustering coefficient C the average of clustering coefficient c_{v_i} of every node $v_i \in V$.*

$$C = \langle c \rangle = \frac{1}{|V|} \sum_{v_i \in V} c_{v_i}$$

Definition 1.17. *(Assortativity coefficient). Given a graph $G = (V, E)$ we define assortativity coefficient [34] the quantity:*

$$r = \frac{i}{\sigma_q^2} \sum_{jk} jk(e_{v_j v_k} - Q(j)Q(k))$$

where σ_q^2 is the variance of the distribution $Q(k)$, thus:

$$\sigma_q^2 = \sum_k k^2 Q(k) - \Big[\sum_k kQ(k)\Big]^2$$

Chapter 2

Motif search strategies

Intuitively, to define whether the frequency of a particular observation has relevance in a study, it is possible to define a frequency threshold. If this frequency exceeds the threshold, it will be considered relevant for the purpose of the study. However, choosing a threshold to evaluate observations means having some a priori information about what is being analysed, otherwise, it would be an arbitrary choice. Therefore, it is possible to define a null hypothesis and a probability distribution of the observed phenomenon. The probability of an outcome (i.e. the p-value) with respect to a reasonable null hypothesis is a principal way to determine unusualness. In the motif search, there is no prior knowledge about the motifs we are looking for or the networks we are analysing, for this reason, the conventional approach to determine if a sub-graph can be considered a motif is to follow these three steps:

1. Starting from the target network, generate a large set of random networks that share with it many of the most important features of a network. (roughly the same number of nodes and edges with similar degrees distributions);

2. For each distinct topology find the number of its occurrences in each of those networks;

3. Check if a sub-graph can be considered a motif, estimating the p-value by comparing the number of occurrences in the input network with the numbers of occurrences in the random networks.

2.1 Random graphs

There are various models for the generation of random networks, each of which applies different strategies. It is essential that the generated random graph is similar to the original graph in terms of global properties such as average grade, average path length, degree distribution, and many other relevant properties. The choice of the random model and the number of random networks to be created is a delicate aspect, in fact, in most cases, the biological networks are enormous and replicate them numerous times is computationally heavy.

2.1.1 The switching method

The most common random graph generation technique used in motif finding algorithms is the 'Switching Method' [32]. It repeatedly selects two random edges, and exchanges the ends to form two new edges. The random network has the same number of node and edge of the input network, and the in- and out-degrees of the nodes are preserved. Has also been widely applied in studying such features of biological networks as modularity and degree correlation. Many studies [38] have shown that, for many networks, $100*E$ times of switching appear to be adequate to achieve randomisation, where E is the number of edges.

2.1.2 Erdos-Renyi (ER) model

Erdos-Renyi [43, 10], also called ER model, is a model able to generate random graphs in which two nodes connect each other randomly and independently. In the variant of the model in which the graph is defined as

$G(|V|, |E|)$, the algorithm randomly creates a network uniformly over all networks that have $|V|$ nodes and $|E|$ edges. In another variant of this model, the generated graphs are denoted as $G(|V|, p)$, with $0 < p < 1$, where p is the probability of connecting each pair of nodes. The properties of the graphs generated according to this model vary considerably depending on the probability; therefore there is no certainty of having a uniformity of characteristics in the randomly generated networks (Figure 2.1).

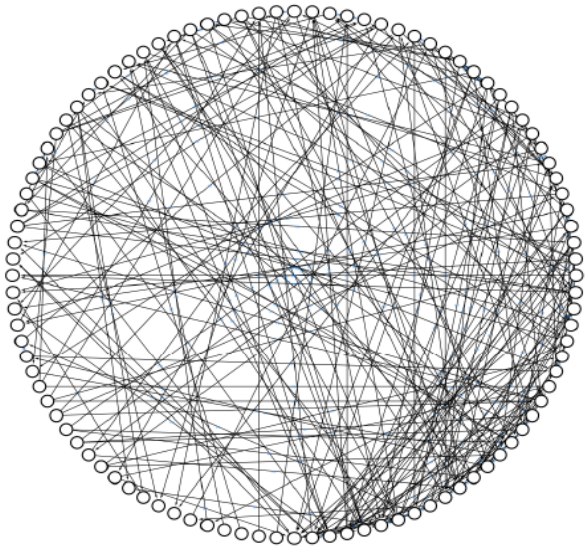

Figure 2.1: Network generated according to the Erdos-Reny model.

2.1.3 Watts-Strogatz model

The Watts-Strogatz [9] model generates networks that respect the small-world property. They are characterised by a small average length of the paths between the pairs of vertices and a high coefficient of clustering. Nodes are not close to each other. Nodes that are not directly connected to each other can be reached from any other node by a small number of hops or steps. Initially a set of $|V|$ is generated ring-shaped nodes in which each node is connected with an edge to its neighbour, so an edge to its right and one to its

left. The edges are randomly mixed with rewiring probability β. Low values of β produce quasi-regular graphs, where nodes have approximately the same degree, while high values of β produce networks which are very close to the ER model (Figure 2.2).

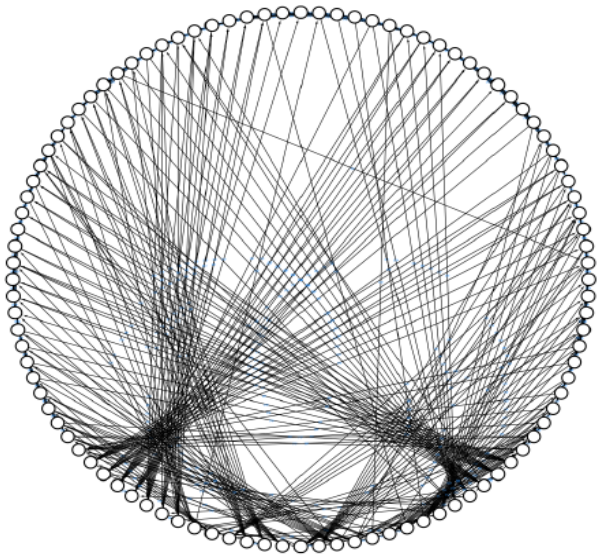

Figure 2.2: Network generated according to the Watts-Strogatz model.

2.1.4 Barabasi-Albert model

The Barabasi-Albert model [2] creates graphs where the more connected a node is, the more likely it creates new links. The generation of the random network starts from a small number of nodes. At each subsequent step, a node will be added and connected to the new network. The probability that an edge connects the new node with an already existing node will be linearly proportional to the current degree of this node (Figure 2.3).

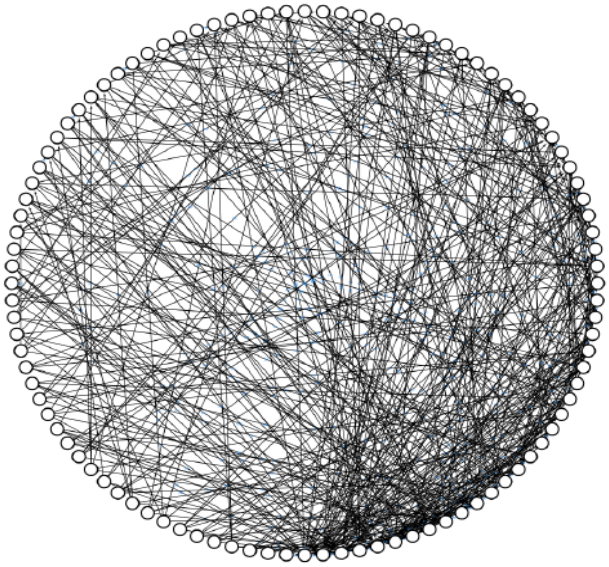

Figure 2.3: Network generated according to the Barabasi-Albert model.

2.1.5 Geometric model

The geometric model [27] is used especially to generate spatially oriented networks and for networks in which the information contained by the nodes influences the topologies of sub-graphs included in the network. In the geometric model, each node is represented as a point in a d-space. An edge between two nodes exists if the distance between corresponding points is within a threshold r (Figure 2.4).

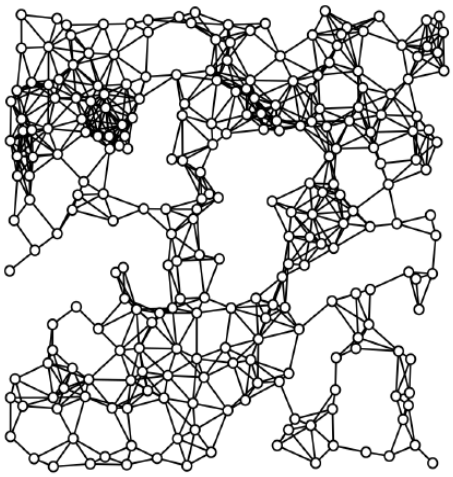

Figure 2.4: Network generated according to the Geometric model.

2.1.6 Forest-Fire model (FF)

In the Forest-Fire (FF) model [19] is used to describe continuously evolving networks. These networks evolve in a super-linearly way increasing the number of edges with respect to the number of nodes. Also, the distance between nodes shrinks as new nodes arrive. In other cases, it is possible that the average distance between the nodes decreases progressively, rather than growing slowly as a function of the increase in the number of nodes. A new node v attaches to the network by iteratively exploring existing edges starting from one or more anchor nodes, called ambassadors, which are chosen randomly. At each step of the exploration, v creates out-links with newly discovered nodes with a forward probability p and in-links with a backward probability r and continues exploration from those nodes.

2.1.7 Duplication model

The duplication model [12] is well suited to biological networks. In fact, they are subject to a different evolution compared to other networks,and in particular, their evolution model is characterised by process of duplication of

nodes and the connections between them. The main aspect that duplicating the nodes also duplicates the information they contain; it is considered a dominant evolutionary force for the growth of a network. There are two procedures for duplication. In the complete duplication of a network, a node is selected from the original graph with a fixed probability p. At each successive step, a new node v_j is added, connected to all neighbours of node v_i but not to v_j. In the process of complete duplication, the steps are the same, but the node v_j will be connected to the neighbours of v_i with a probability q. The lower is p, the more divergent is v_j as a copy of v_i (Figure 2.5).

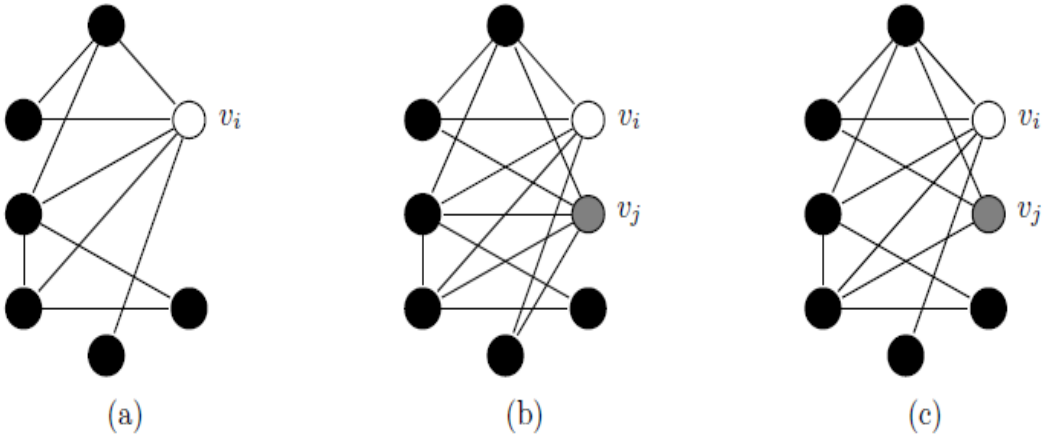

Figure 2.5: Total and partial duplication process in a graph: a node is selected with a probability p from the original graph (a); the process of total duplication involves the addition of a new node v_j connected to all neighbors of node v_i (b), instead the partial duplication process foresees that the vertex is connected to all neighbors of v_i with a probability q (c).

2.1.8 Chung-Lu model (EDD)

The Chung–Lu model [11], also known as Expected Degree Distribution (EDD), model generates graphs in which node degrees follow a given distribution. It was initially introduced for non-labelled graphs. Each edge is chosen independently with probability proportional to the product of the expected degrees of its endpoints. Given an undirected graph $G(V, E)$ with

$|V| = N$, let be a random variable f_D based on the degree distributions of G. $P(f_D = d)$ is the probability that a node has degree d in G. Assigning degrees to each node i in V' by sampling according to the f_D distribution and given f_D, we can generate a new graph $G = (V, E)$ with $|V'| = |V|$. An edge between two nodes i and j, with $i \neq j$, is generated with probability:

$$P(i, j|D(i), D(j)) = min(1, \gamma \times D(i) \times D(j))$$

where $\gamma = 1/[(N-1) \times \mathbb{E}[f_D]]$ and $D(i)$ is the degree of node i within the input graph. In Match&Motif an EDD model extension was used for labelled graphs. In this extension the node degrees depend, at least partially, on labels.

2.2 Motif occurences

The problem of counting the motifs is widely known and studied, which is why various strategies have been developed to optimise research. Regardless of which strategy is chosen, when the size of the sub-graph becomes > 10, the problem becomes intractable in fact the number of possible sub-graphs, considering that they can overlap between them, becomes enormous [35, 13, 1, 26]. Without the necessary precautions, to count all motifs occurrences can be time-consuming and computationally heavy, like the whole process of motif search. Recalling that a biological network can be made up of millions of nodes, looking for all the different possible topologies of a motif is not a trivial operation. For example, a motif with direct edges of size 5 (where 5 is the number of nodes that compose it) has 9364 different distinct topologies. For each topology, the millions of nodes of the target network must be checked to make matching, and one must avoid counting isomorphs of the topology. These operations must be repeated for all topologies. All this only for a network, in fact, the procedure described above must be repeated n times where n is the number of the random network that has been chosen for the significance of the motifs. The amount of comparisons is impressive.

2.2.1 Pattern growth tree

A reliable and efficient method to generate all possible sub-graphs starting from another sub-graph is to extend it one step at a time by modifying one of its nodes or edges and use the extended sub-graph to generate further variants. This sequence of activities can be viewed as a tree, where each node of the tree is one sub-graph, and its children are sub-graphs extended from that node. In this way, all the children of that node are an extension of that sub-graph. This strategy, called 'Pattern growth tree', can be used to systematically generate all possible size-k graphs starting with a size-k tree. It can also be used to systematically enumerate all occurrences of size-k sub-graphs in the target network. Considering a node of the tree, thus a sub-graph, if we suppose that its frequency falls below the frequency threshold, it is possible to use this information to prune the sub-trees rooted at that node, leading to increase computational efficiency. In fact, the biggest drawback in generating all possible topologies is that the frequencies of sub-graph not belonging to the target network could be calculated. Another advantage of pattern growth tree is that, if we search for size-k sub-graphs in the network, a pattern growth tree built for size-(k-1) sub-graphs can be reused. It may lead to significant improvements in computational cost.

2.2.2 Sub-graph census

Sub-graph census, which is usually used with Pattern growth tree, is the process of scanning the target network (node-by-node or edge-by-edge) and enumerating all occurrences of all sub-graphs of a given size. The two main strategies for the sub-graph census are the exhaustive census and the probabilistic census. The first, intuitively, consists in counting all the occurrences of a sub-graph. This approach is a time-consuming process, especially when the network or the sub-graph to be enumerated is large. The second strategy, called sampling strategy, consists of taking an adequate number of random size and sub-graph samples from the original network. With a good amount

of trials one can expect to get closely accurate yet quick results, but with a non-zero probability that some potential motifs will be missed. Sampling saves much computation, thus allowing us to discover larger motifs. Sample sub-graphs can be generated and extended from an edge, 'edge-sampling', or a node, 'node-sampling'. For each sub-graph in the pattern growth tree, an 'edge-sampling' strategy [33] an edge is randomly selected and extended until a sub-graph of size-k is selected from that edge. This sub-graph will be used for comparison. With this strategy, the sub-graphs that have more edge will be counted more often, so the probability of selecting sub-graphs of size-k is not uniform. The 'node-sampling' strategy [46, 45] is able to ensure the uniform probability of picking any size-k sub-graphs. It assigns to each level of its pattern growth tree a probability that the nodes (of the tree) at that level will be further explored. Passing through the tree in a probabilistic way, it is guaranteed that all the nodes that are the k-size sub-graph, will be explored with equal probability. Also, an accurate node extension in the pattern growth tree can ensure that a particular sub-graph will be encountered exactly once, which avoids redundant computation.

2.2.3 Isomorphism checking

Two graphs are 'isomorphic' if there exists a one-to-one mapping between their nodes such that each edge in one graph can be mapped to an edge in the other graph [44]. Verifying the isomorphism of two graphs means verifying they are topologically equivalent. This problem, known as isomorphism checking or sub-graph matching, is an NP-Complete [24] problem, in fact, does not exist an algorithmic solution able to solve it in polynomial times [25]. Isomorphism can be solved by using 'Canonical Labeling' of its nodes [22]. Practical algorithms such as the NAUTY [3] have been used in several motifs finding tools for isomorphism testing.

2.2.4 Mapping

The mapping strategy [20] can be used in the contraposition with the enumeration and estimating the frequency of all possible sub-graphs. It allows finding the instances of a sub-graph taking as input a sub-graph of size-k (query graph or candidate motive) and maps it onto the network in as many different places as it can. Based on the degree of the nodes and the degree of their neighbours, the mapping strategy classifies the nodes in the target network and then controls the nodes in the input network that have similar characteristics to the query network [20, 21]. Mapping, when used with symmetry breaking, resolves isomorphism without explicit checking [20].

2.2.5 Symmetry breaking

The symmetries of a graph are known as 'automorphisms'. The group of automorphisms are in the same 'equivalence class'. One can specify a set of symmetry breaking conditions [20] for each equivalence class. Checking if these conditions hold between two graphs is easier than checking the isomorphism. Algorithms that use symmetry breaking while generating candidate motifs can prevent many isomorphism computations, thus increasing efficiency.

2.3 Validation

When the frequencies of all the size-k sub-graphs have been calculated for both the target network and all random networks, different metrics and thresholds can be chosen to determine if a subprogram is significantly frequent, i.e. if it can be considered a motif. The two methods for calculating significance will be explored in the next chapter.

Chapter 3

Significance

There are two ways to calculate the significance of a motif: the statistical method, which through sampling of random networks similar to the target network verifies the significance of the motif, and the analytical model, which through the Pólya-Aeppli distribution accurately calculates the mean and the variance of a motif under any exchangeable random graph model.

3.1 Statistical significance

The most common way to determine when a sub-graph can be considered a motif is to establish a threshold. If the sub-graph repeats a number of times in the target network then it can be considered a motif, if instead, the sub-graph has a frequency below the threshold then it can not be a motif. Thus, the most intuitive metric is:

$$f_{input} \geq F$$

where F is the frequency thresholds and f_{input} is the frequency of the sub-graph in the target network. The 'uniqueness threshold' establishes that the frequency of a candidate motif in the input network is at least above its mean frequency in the set of randomly generated networks. Let a size-k sub-graph g_k occur $\bar{f}_{input}$ times in the input network. Let $\bar{f}_{random}$ be the mean of

frequencies of g_k in the random networks. Then, g_k is 'unique' if

$$(f_{input} - \bar{f}_{random}) > U \times \bar{f}_{random}$$

where U is the uniqueness threshold. Several other metrics are used for the significance of the motif [49].

Let a size-k motif g_k occur $\bar{f}_{input}$ times in the input network. Let $\bar{f}_{random}$ and σ^2_{random} be the mean and variance of frequencies of g_k in a sufficiently large set of random networks, respectively. The z-score is defined as the difference of finput and f random, divided by the standard deviation σ^2 random. The z-score is calculated as,

$$z(g_k) = \frac{f_{input} - \bar{f}_{random}}{\sqrt{\sigma^2_{random}}}$$

The P-value represents the equal or greater number of times in a random network than in the given input network. If the associated P-value is less than 0.01 or z> 2.0, a sub-graph is statistically significant and can be considered a motif.

Statistical significance can be measured in other ways. The z-score can be replaced by 'Abundance', Δ [39], that is defined as,

$$\Delta(g_k) = \frac{f_{input} - \bar{f}_{random}}{f_{input} + \bar{f}_{random} + \varepsilon}$$

where, ε is a small positive number preventing the ratio from approaching infinity when the frequencies are small. The value of Δ usually ranges between -1 (under-represented) and 1 (overrepresented).

Another metric, the motif 'significance profile' (SP), is defined as a vector of z-scores of a particular set of motifs, which is normalised to a length of one [39]. Let n be the number of motifs in the set, and z_i be the z-score of the i-th motif. The motif SP of the i-th motif in the set is thus calculated as,

$$SP_i = \frac{z_i}{\sqrt{\sum_{j=1}^{n} z_j^2}}$$

SP of the entire network gives a histogram of the normalised z-scores of all possible motifs [39].

The 'concentration' of a candidate motif denotes how frequent it is in the network compared to other sub-graphs of the same size [33, 46]. If there are n sub-graphs of size k in the network, the concentration of the i-th size-k sub-graph $g_{k,i}$ is defined as

$$C(g_{k,i}) = \frac{f_{k,i}}{\sum_{j=1}^{n} f_{k,j}}$$

where $f_{k,j}$ denotes the frequency of $g_{k,i}$ in the network.

3.2 Analytical significance

The reliability of the p-value is strictly related to the number of randomizations performed. To avoid such an expensive simulation, a key problem is to identify a proper distribution fitting the number of observations in the random reference model. Picard et al. [14] proposed a model to exactly compute the mean and variance of the count of a given pattern under any exchangeable random graph model. Exchangeability means that the probability of occurrence of topology does not depend on its position in the graph. The random model used is Expected Degree Distribution (EDD), which generates random graphs whose node degrees have the same expectation as the input network G. It is possible to extend the EDD model for non-labelled graphs, to those labelled. Ignoring the node's label, and considering only the topology, information related to it could remain hidden.

According to equation

$$P(i, j|D(i), D(j)) = min(1, \gamma \times D(i) \times D(j))$$

discussed in paragraph 2.1.8, used for generate random graphs under the EDD model, we can introduce some variables for graphs in which node degrees depend on labels. Let $f_D|c$ be a random variable defined as the degree distribution for nodes with label c in the input graph G . Let $P(f_D = x|c)$ be the probability of sampling a node in G with a degree x given the label c.

We define the occurence probability of the topology of a labelled motif m_C with k nodes, given a label assignment C, in the graph as:

$$\mu(m_C|C) = \gamma^{m_{++}/2} \prod_{u=1}^{k} \mathbb{E}[f_D{}^{m}_{u+}|c_u]$$

where $f_D|c_u$ is the degree distribution for nodes with label c_u in the input network, m_{++} is twice the total number of edges in m_C, m_{u+} is the number of out-going edges from node u in m_C and $\mathbb{E}[f_D^{m_{u+}}|c_u]$ is the m_{u+}-th moment of the conditional distribution $f_D|c_u$.

The mean and the variance of motifs under exchangeable random graph model can be calculated according to following equations. Let $\alpha = (i_1, ..., i_k)$ be a k-tuple of ordered indexes representing a potential position of m_C in G. We introduce a random indicator of occurrence $Y_\alpha(m_C)$ which equals one if the topology m_C occurs at position α and 0 otherwise. We need to consider a set of Non-Redundant Permutations (NRP) of m_C, which we call $R(m_C)$, because some permutation of the indexes produce the same motif. The total count $N(m_C)$ of labelled motif m_C is then:

$$N(m_C) = \sum_{\alpha} \sum_{m'_C \in R(m_C)} Y(m'_C)$$

Under the exchangeability assumption, the distribution of $Y_\alpha(m_C)$ does not depend on α and let us define:

$$\mu(m_C) = \mathbb{E}[Y_\alpha(m'_C)], \quad \forall \alpha, \forall m'_c \in R(m_C)$$

The expectation of $N(m_C)$ is:

$$\mathbb{E}[N(m_C)] = \binom{N}{k} \times |R(m_C)| \times \mu(m_C)$$

where $\binom{N}{k}$ is the number of all possible locations of m_C in G. We compute the variance of the number of occurrences of the labeled motif as:

$$\mathbb{V}[N(m_C)] = \mathbb{E}[N^2(m_C)] - \mathbb{E}[N(m_C)]^2$$

It is, however, possible to adapt the model for the analytical significance proposed by Picard et al. to any other random graph model, provided it is an exchangeable random graph model and respects other properties that will be mentioned soon.

Given $N(m_c)$, Picard et al. shown that the distribution of Pólya-Aeppli (also known as the Geometric-Poisson distribution) approximates well the distribution of $N(m_c)$ and is suitable to describe how the count of occurrences of the motif can vary. They also show that this is a good model for the distribution of the counts of sub-graph topologies since the fit is more accurate than a Gaussian model for the graphs of many applications.

The Pólya–Aeppli distribution supposes that objects (which are to be counted) occur in clusters, the number of clusters follow a Poisson distribution, while the number of objects per cluster has a geometric distribution. It holds when distinct topologies can share nodes and edges (i.e. clumps) [14]. Picard et al. show that when:

i) The number of clumps has a Poisson distribution with mean λ;

ii) The sizes of the clumps are independent of each other;

iii) The clumps have a Geometric distribution $G(1-\alpha)$;

the number of observed events X (topologies) has a Pólya–Aeppli distribution P(λ, α). These results lead to an estimate of the count of occurrences of a given topology. In this case, we have that $X \sim PA(\lambda, \alpha)$ is a random variable representing the number of observed events (i.e. motif occurrences in our case):

$$P(X = x) = \begin{cases} e^{-\lambda}\alpha^x \sum_{c=1\cdots x} \frac{1}{c!}\binom{x-1}{c-1}[\frac{\lambda(1-\alpha)}{\alpha}]^c & if \quad x > 0 \\ e^{-\lambda} & if \quad x = 0 \end{cases}$$

The mean and the variance of $PA(\lambda, \alpha)$ are defined as $\frac{\lambda}{1-\alpha}$ and $\frac{\lambda(1+\alpha)}{(1-\alpha)^2}$. By making use of the mean and variance obtained using the exchangeable random graph model we can deduce the parameters of the distribution as $\alpha = \frac{\mathbb{V}[N(m_C)]-\mathbb{E}[N(m_C)]}{\mathbb{V}[N(m_C)]-\mathbb{E}[N(m_C)]}$ and $\lambda = (1-\alpha) \times \mathbb{E}[N(m_C)]$.

Chapter 4

Match&Motif: a Cytoscape App for statistical and analytical motif search

This chapter presents a general overview of Match&Motif and its main features. Match&Motif is a Cytoscape app, compatible with all versions of the software starting from version 3.6. It is based on the plugin NetMatch* [15], compatible with the versions 3.1 of Cytoscape, which inherits some of its features, and FlashMotif [31] tool. Cytoscape [8] is open-source software for the visualisation and analysis of molecular interaction networks, whose peculiar characteristic is flexibility: in fact, Cytoscape allows researchers and developers to adapt the application to its specific needs thanks to the ability to implement external apps, developed in Java language, usable through an App Store [23] online.

4.1 Functionalities

To start the app, it is necessary to select the corresponding item from the "Apps menu". At startup, the app's main panel will be loaded at the Cytoscape control panel (Figure 4.1).

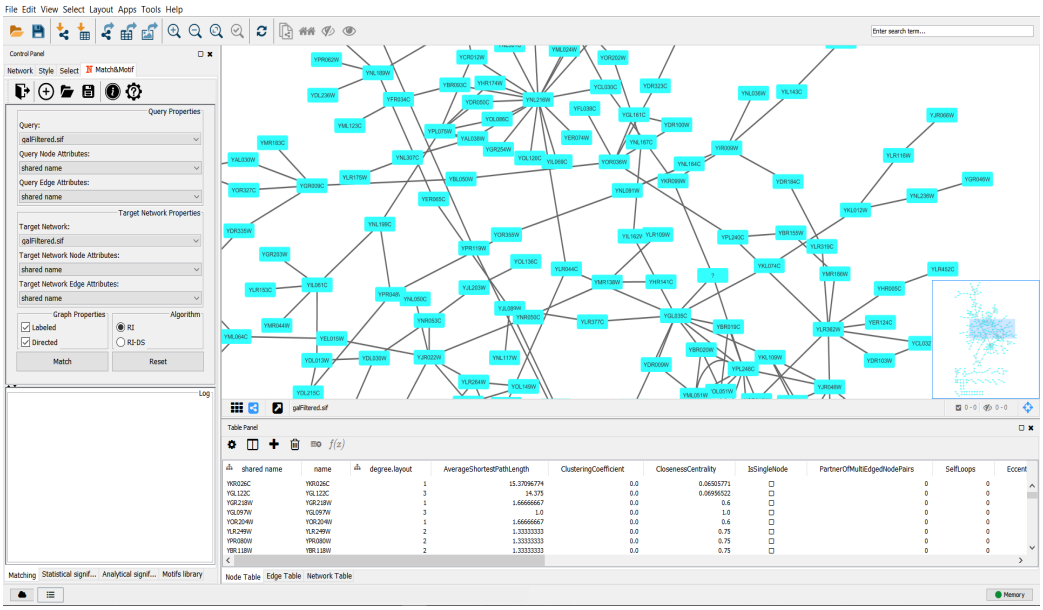

Figure 4.1: in the Cytoscape control panel, the main interface of Match&Motif is loaded on the left.

The main interface of Match&Motif consists of a box with four tabs:

1. The first tab, called "Matching", allows specifying the query networks and target networks and then it starts the search for matches (Figure 4.2 (a));

2. The second tab, called "Statistical Significance", allows to verify the statistical significance of the query according to a specific randomisation model (Figure 4.2 (b));

3. The third tab, called "Analytic Significance", allows to verify the analytical significance of the motif according to four different ways of labelling it (Figure 4.2 (c));

4. The fourth tab, called "Motif library", allows a set of pre-defined queries that can be passed to the search task (Figure 4.2 (d)).

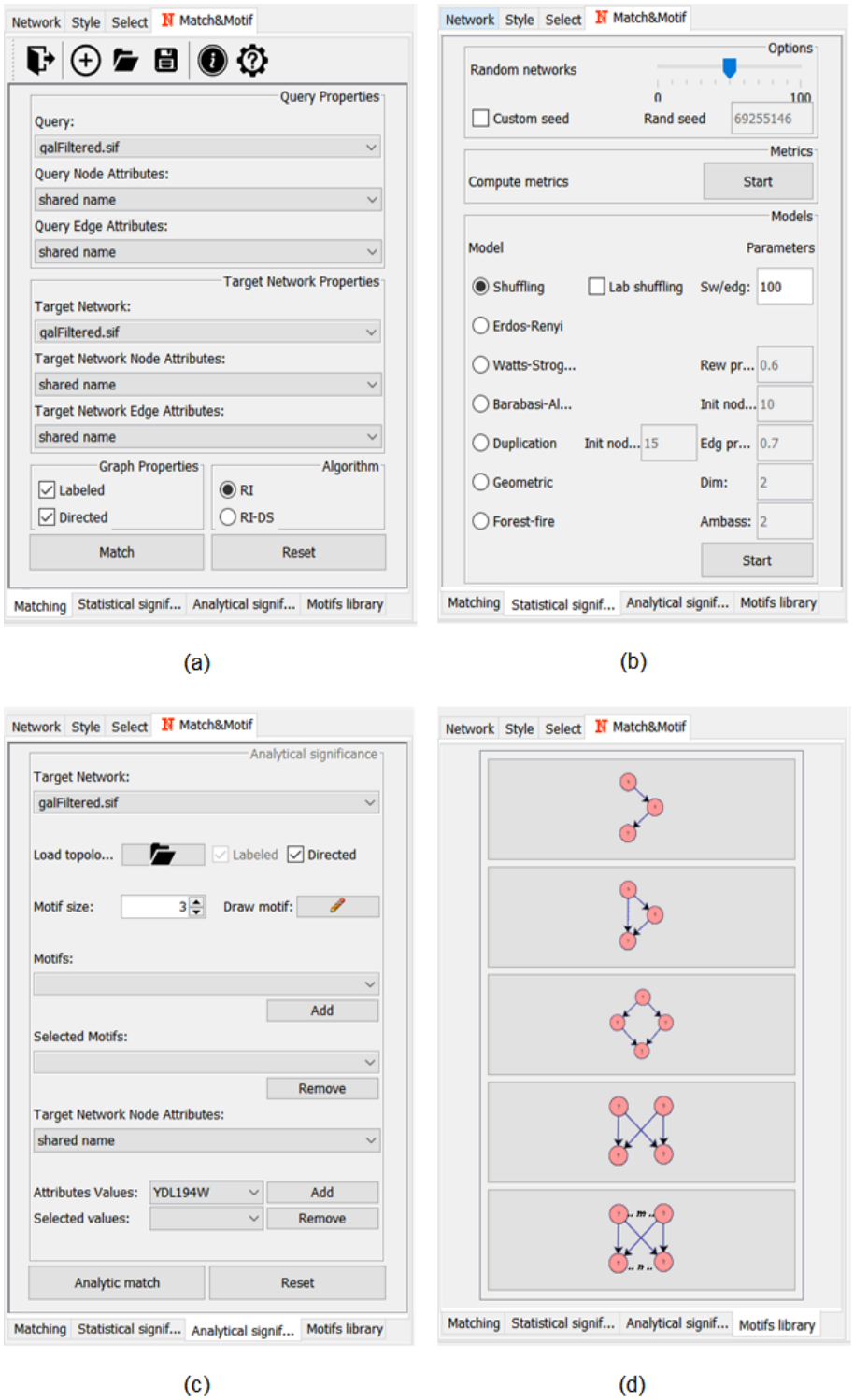

Figure 4.2: Four tabs of Match&Motif. Matching tab (a), Statistical Significance tab (b), Analytic Significance tab (c) and Motif library tab (d).

4.2 Load a query

From the main "Matching" panel it is possible to load the desired query and target networks. By selecting the "Load New Network" button you can load a network in various formats, including Simple Interaction Format (SIF) [8], GML [17], XGMML [37], GraphML [7] and many others. The loading of a network can also be done through the Cytoscape options. In any case, at the end of the upload, the networks will be added to the "Target networks" field both in the "Matching" field and in the "Analytical significance" panel.

4.3 Drawing a query

Match&Motif offers the possibility to create and edit new queries. To do this, from the main panel, select the "Create New Query Network" button. A new panel will be added to the main Cytoscape panel. Right-clicking on the panel will open the standard Cytoscape menu that allows the user to add, edit or remove elements of the graph, such as nodes (Figure 4.3) and edges (Figure 4.4).

By selecting the "Motifs library" tab (Figure 4.2 (d)), the user can access to a predefined set of quickly selectable queries that include some small topologies that can be identified as motifs [41] in many real networks. The motifs supported by Match&Motif are the three-chain, the feed-forward loop [29], the bi-parallel, the bi-fan and the m-to-n-fan (Figure 4.5).

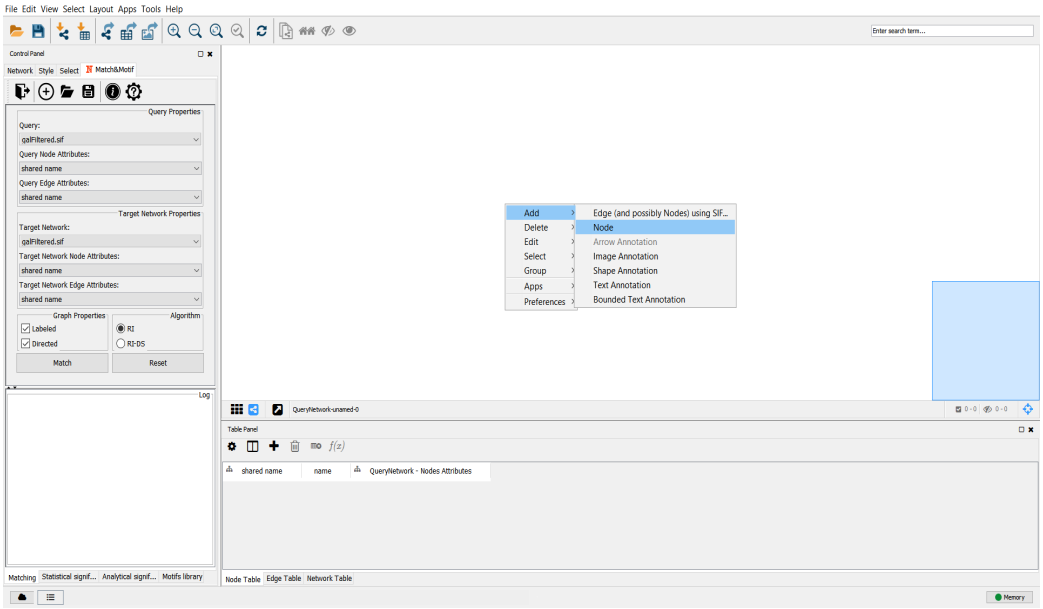

Figure 4.3: Menu for adding a new node to the query graph.

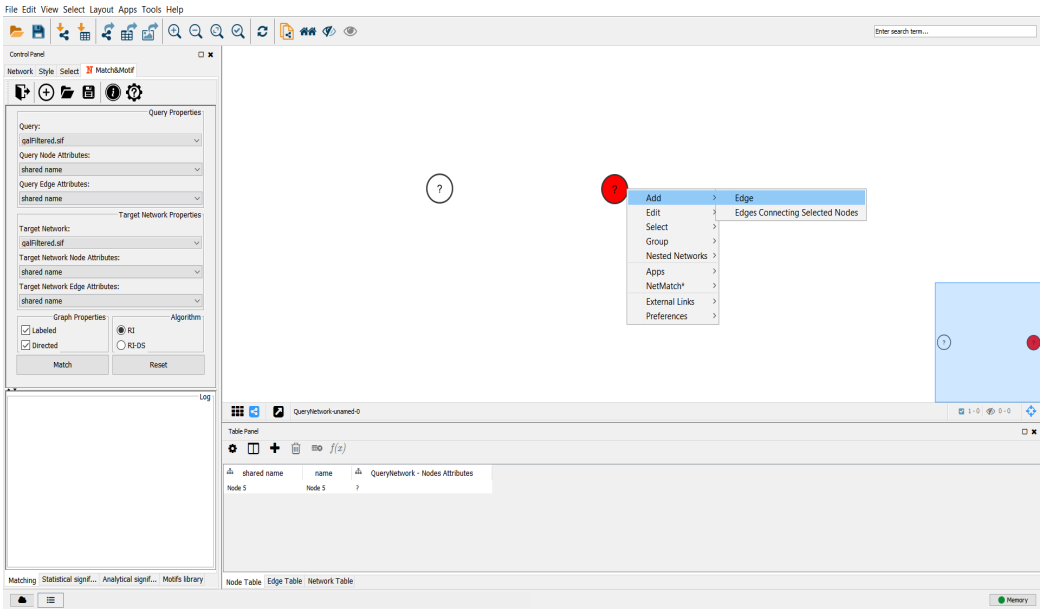

Figure 4.4: Menu for adding a new edge to the query graph.

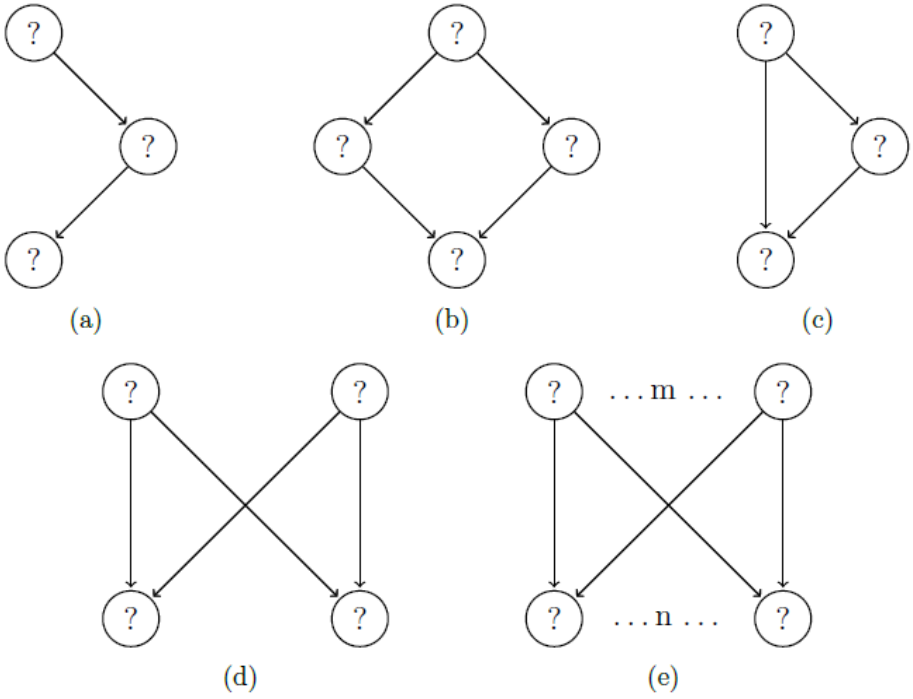

Figure 4.5: Match&Motif default query: three-chain (a); bi-parallel (b); feed-forward loop (c); bi-fan (d); m-to-n-fan (e).

A query can also be drawn by clicking on the "Draw" button, present in the panel of analytical significance. A query will be added to the main Cytoscape panel with a number of nodes equal to the size of the topology to be searched (Figure 4.6). This dimension is defined in the "Motif size" field. Then the user will be free to add the edges and create the topology (Figure 4.7). All the queries drawn can be selected in the "Motif" field if you want to proceed with the analytical significance (Figure 4.8) or in the "Query" field if you want to precode with the statistical significance.

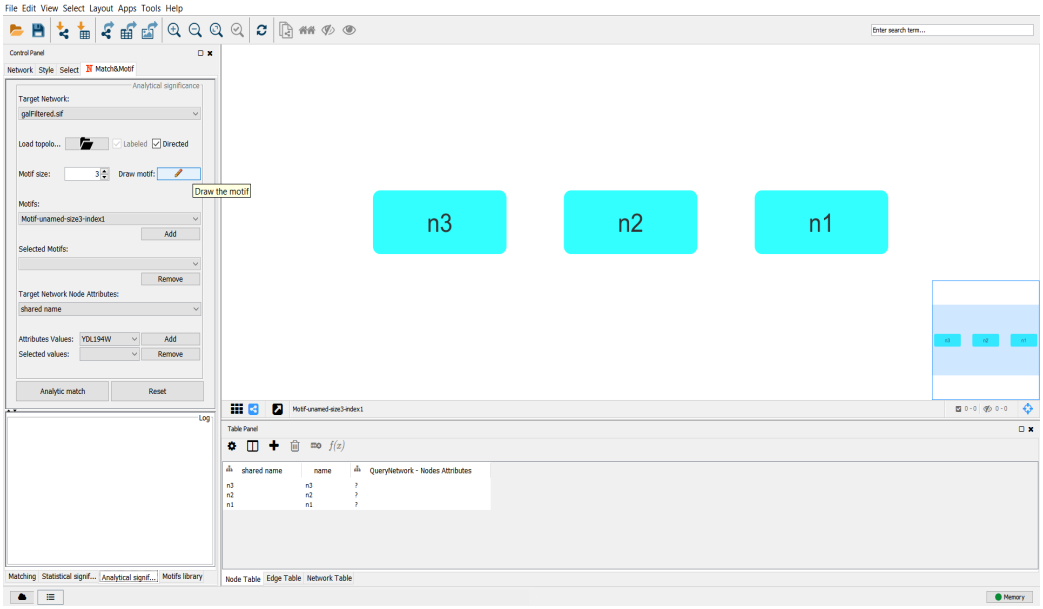

Figure 4.6: Query network drawn using the "Draw" button. The number of nodes in the query is equal to the size set in the "Motif size" field.

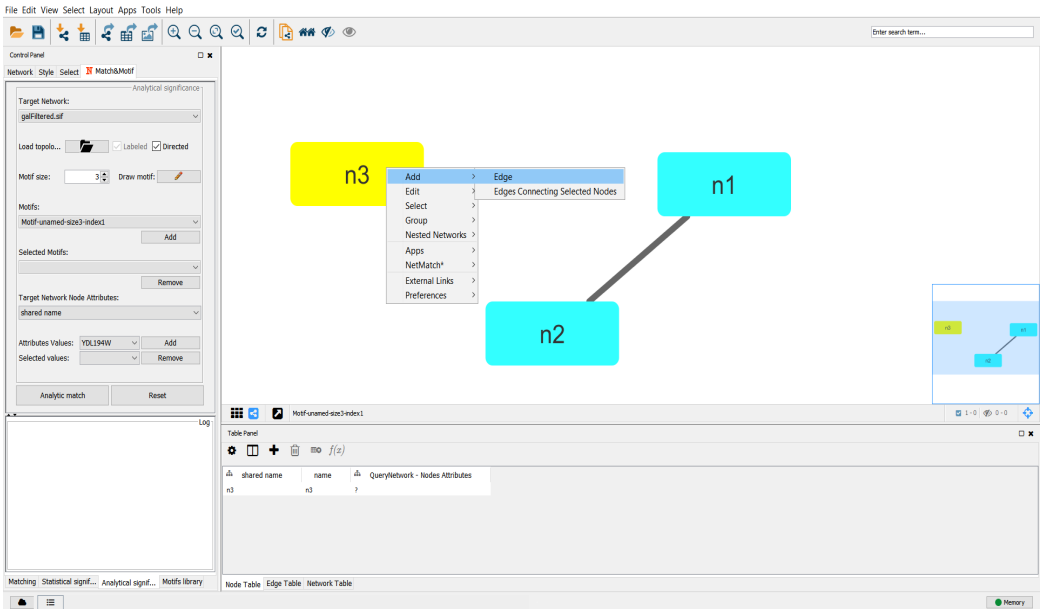

Figure 4.7: Adding an edge to the query network drawn using the "Draw" button.

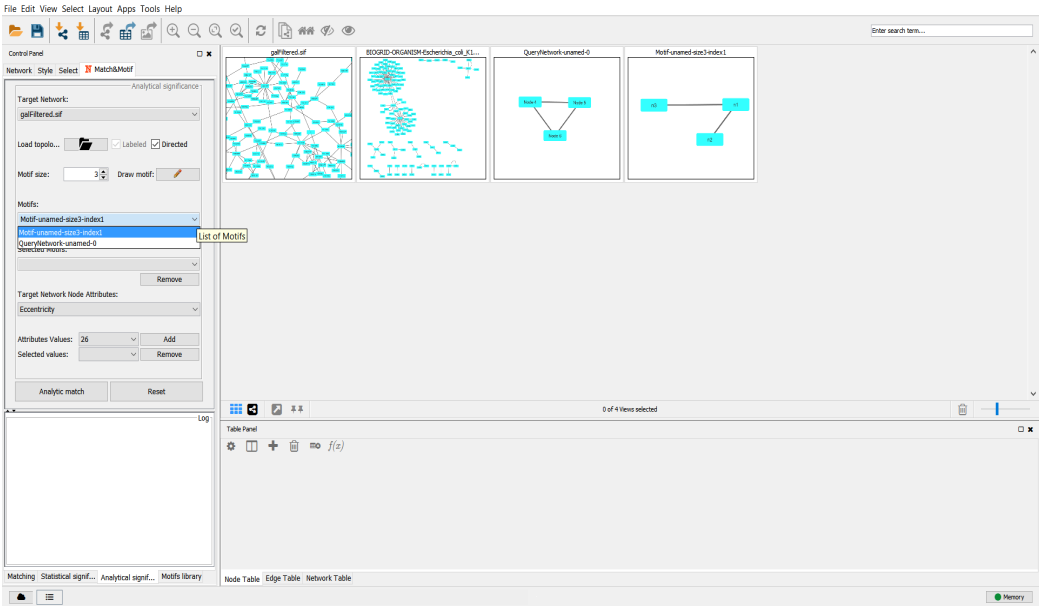

Figure 4.8: List of queries network in the "Analytical significance" tab.

In this last case, in addition to the queries drawn, all the target networks loaded at that moment in Cytoscape will be added to the field (Figure 4.9).

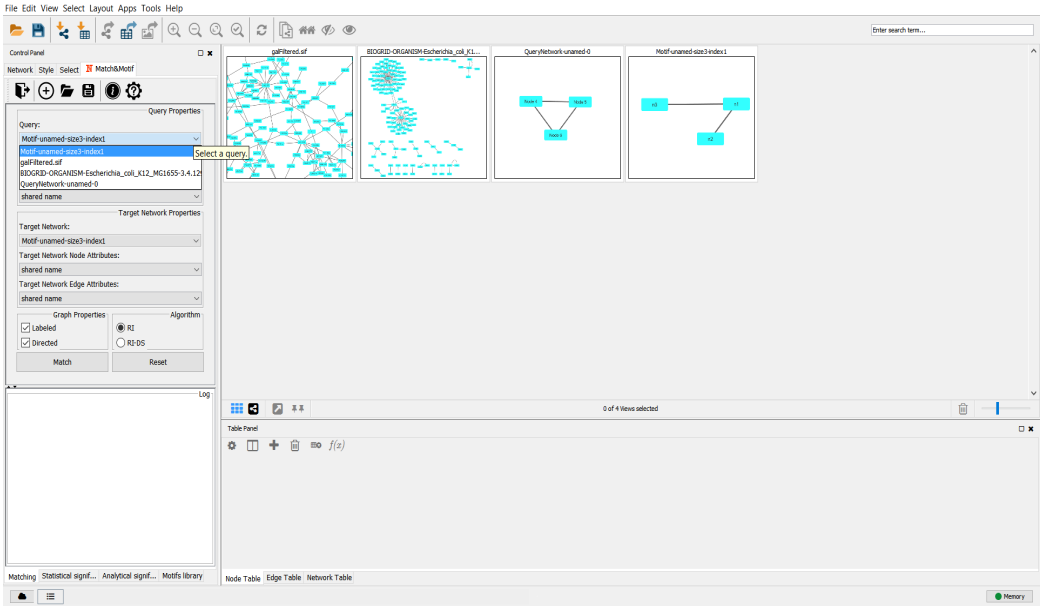

Figure 4.9: List of queries network in the "Matching" tab.

4.4 Labelling a query

Match&Motif gives the possibility to label the motifs in four different ways. It is possible to create injective topological labelled motif (Figure 4.10 (d)) through features already present in NetMatch*.

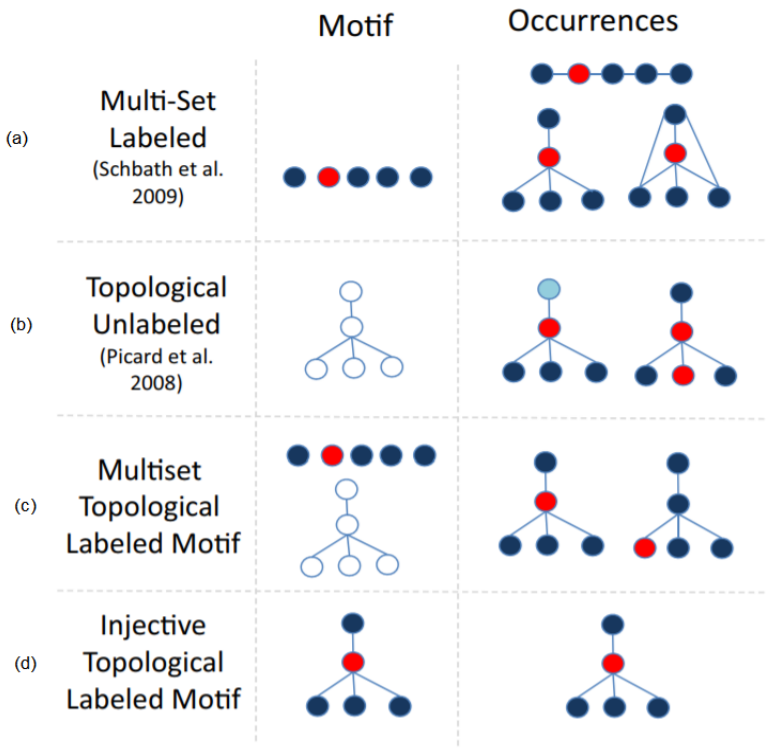

Figure 4.10: (a) Multi-Set Labeled motifs consider all connected structures containing four blues and one red. (b) Topological Unlabeled motifs motifs considers only the topology of the motif, regardless of labels. (c) Multiset Topological Labeled motifs count a chosen topology consisting of four blues and one red regardless of which of the five nodes have which labels. (d) An Injective Topological Labeled Motif count a chosen topology labeled of a single red node in the center surrounded by four blue nodes [31].

After adding the desired nodes and edges, selecting a node and clicking again with the right mouse button, the menu that will appear the allows user to set the options of the selected graph element. In this case, a sub-

menu named "NetMatch*" will appear that will allow the user to modify the labels of the selected node (Figure 4.11). By selecting the corresponding option, the user may decide to set a label for the selected node (Figure 4.12). Otherwise, the node will be considered unlabelled. By default, the nodes are created as unlabelled. In fact, to define an unlabelled node, a special wildcard attribute "?" is associated with it. Cytoscape interprets it as a character, but Match&Motif in the phase of matching will interpret it as absence of label. Any other character or string will instead be interpreted as a label.

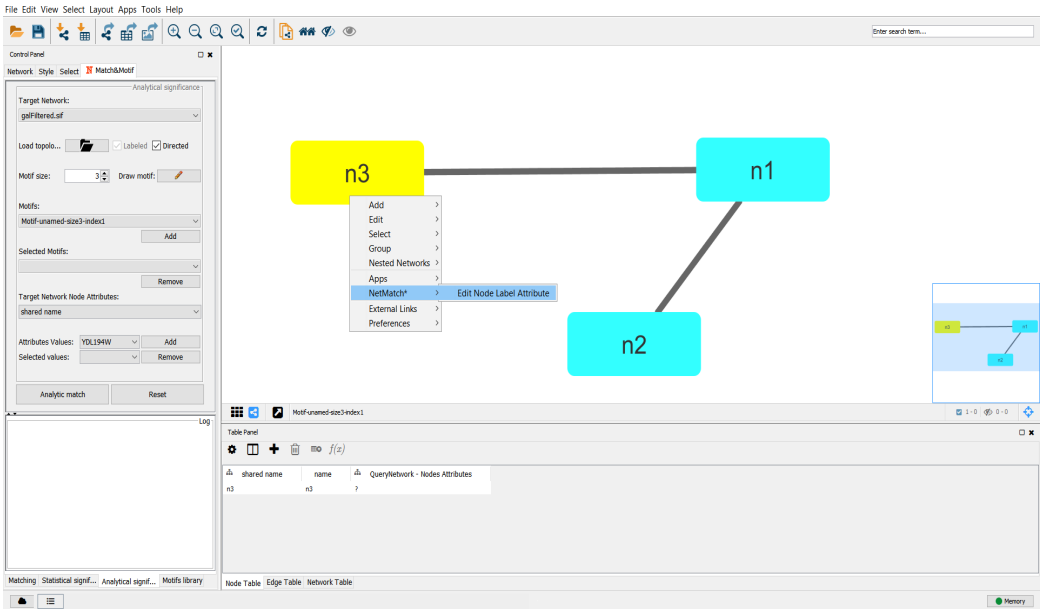

Figure 4.11: Open the menu for labelling the node

The other three types of labelling motifs (Figure 4.10 (a)(b)(c)) are managed in the panel of analytical significance. After selecting the target network, in the 'Target network node attribute' field all the possible attributes present in that network will be displayed. Also selecting the attribute, in the 'Attribute values' field, all the values of that attribute will be shown. Not selecting any query, but defining only its size through the 'Motif size' field, it is possible to search for all the possible topologies of that size, labelled with all the possible combinations of the values of the selected attribute.

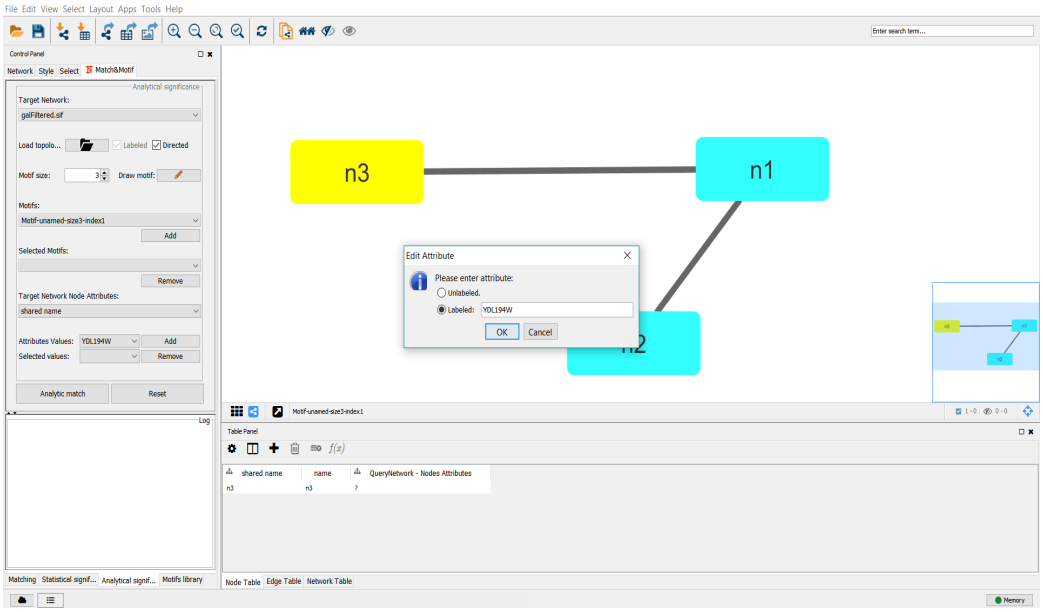

Figure 4.12: Labelling the node

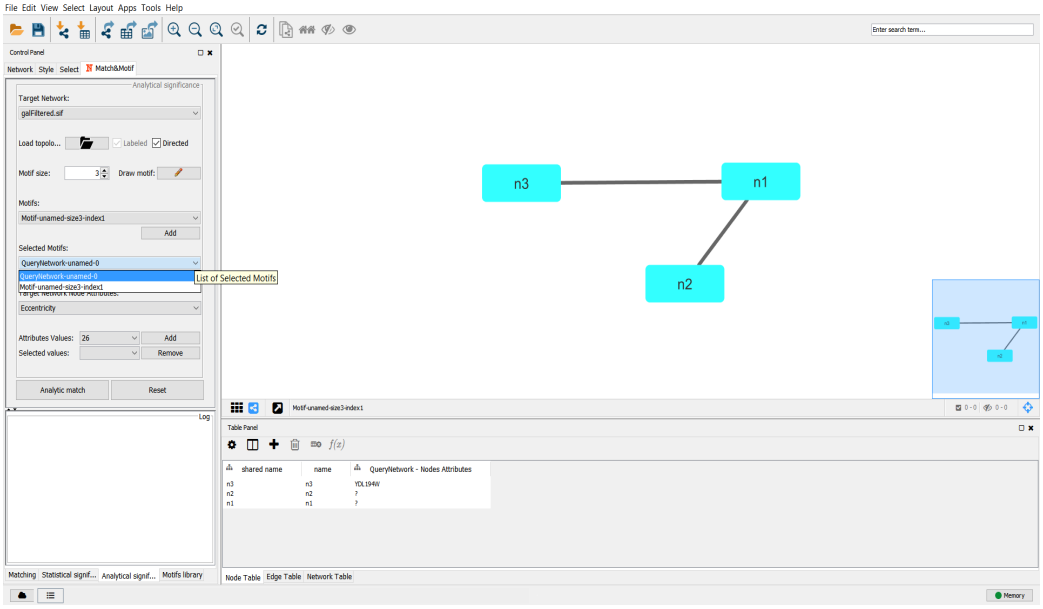

Figure 4.13: From the "Motif" field, using the "Add" button, all the topologies to search within the target network can be added to the "Selected motif" field. The Network target can be selected from the "Target networks" field [31].

Unlike the research carried out through the statistical significance panel, in the analytical significance panel, it is possible to search for multiple queries at the same time. By selecting only the queries, without selecting any labels, only the specified topologies, labelled with all the possible values of the selected attribute (Figure 4.10 (b)), will be searched within the target network (Figure 4.13). In addition, by not selecting any network queries but selecting only n values for an attribute, with n $\leq$ query-size, the search will be performed for all possible topologies with the specified labels (Figure 4.14).

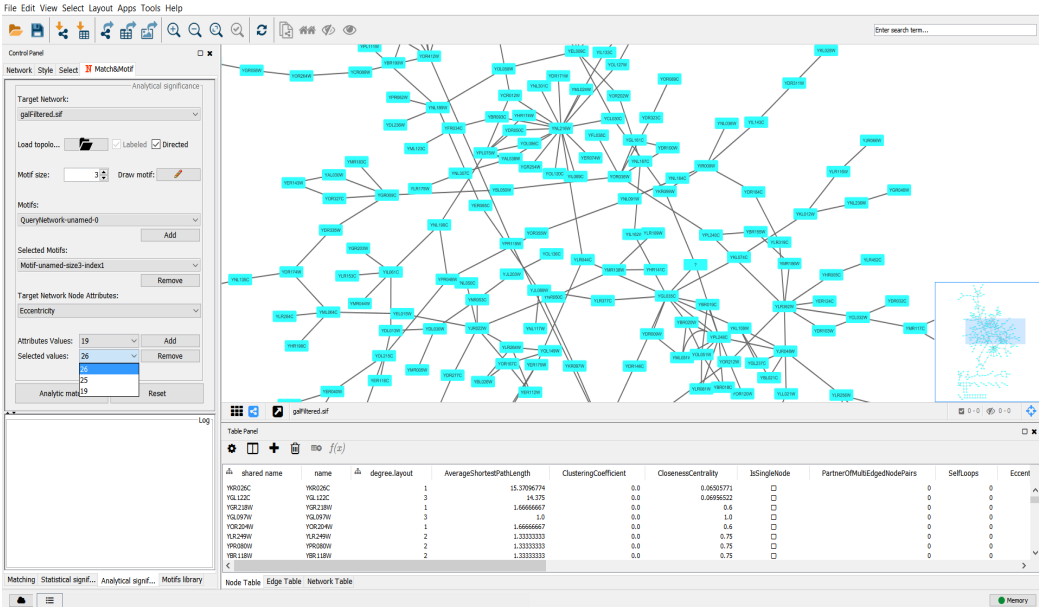

Figure 4.14: Choosing an attribute from the "Target network node attribute" field, all the values of the selected attribute will be shown in the "Attribute values" field. Through the "Add" and "Remove" buttons you can create the set of values with which to label the nodes, "Selected values" field. The "Selected motif" field is empty, so the search will be performed for all possible topologies of the dimension selected in the "Motif size" field.

In the case where n < query-size, (n - query-size) values will be labelled with all possible values of the attribute (Figure 4.10 (a)). The order of the labels is not important because it is not interesting that a specific node has a specific label but rather than a set of nodes is labelled with the set of

those values. In the case in which both the network query and the values with which to label the nodes are specified (Figure 4.15), the search will be performed only for the specified topologies and only for the selected attribute set (Figure 4.10 (c)).

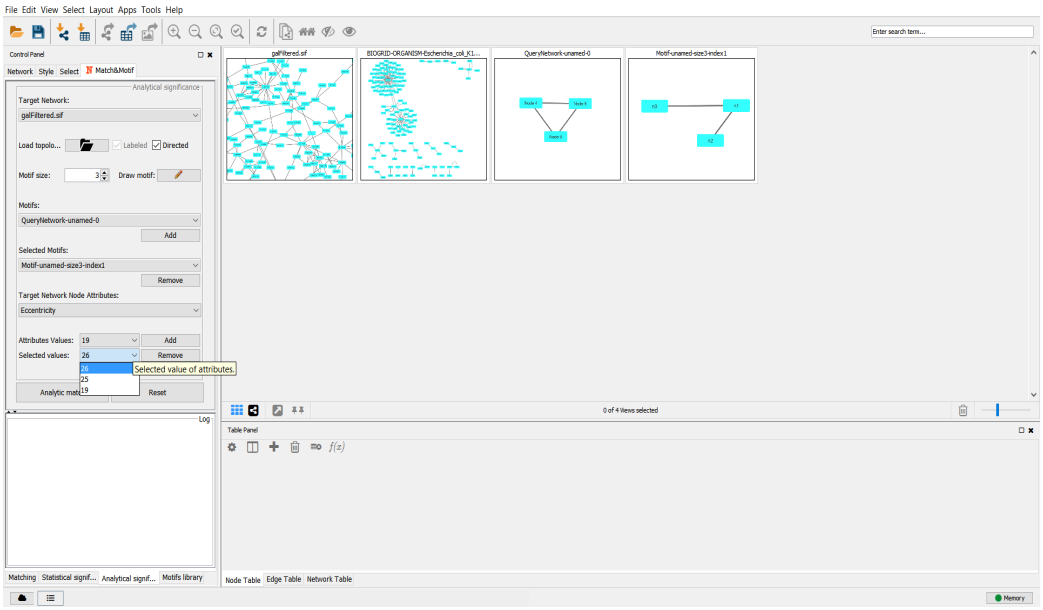

Figure 4.15: To search for a Multiset Topological Labeled Motif, the topologies to search and a set of values of an attribute have been selected

Another added feature is the possibility to import topologies with the related set of labels from a file. The format chosen is shown in the table 4.1. The first line represents the name of the query and will be preceded by the character "#". The second line specifies whether the query is direct or indirect. The third row represents the size of the query, which is the number of nodes. The n following lines, where n is the dimension of the query, the set of query labels. Finally, the lines in which there are two values, represent the edges that connect the nodes and define the topology of the motif.

#query1
undirected
3
?
?
?
0 2
1 2
#query2
undirected
3
?
?
?
0 1
0 2
1 2

Table 4.1: Example of two network queries loaded from files. Both are indirect and composed of three nodes. The character "?" indicates that the set of values with which the query is searched can assume any value of the selected attribute. The pairs of numbers represent edges between the nodes and define the two topologies of the input queries.

4.5 Visualization of results and significance

It is possible to start the search for motif both from the panel of statistical significance, and from that of analytical significativity. The preliminary operations to search, how to select the target network and establishes whether you are looking for a direct or indirect motif, are identical for both panels. Choosing whether the searched motifs should be labelled or not is an obligatory choice for the panel of analytical significance, in fact, the search is done

only for labelled motifs, while in the panel of statistical significance the choice is free. Drawing and labelling queries, as discussed above, depends on the significance and leaves the user wide choice. In the statistical significance panel, moreover, it is possible to choose the search algorithm between RI and RI-DS [6] (Figure 4.1), while for the analytical significance the chosen is implicit because the algorithm used for the matching is one, RI [6].

The main functional difference between the matching present in the two panels, in addition to the validation of the motifs, is that for the statistical significance a matching of the occurrences of the query network is made in the target network, then the occurrences are visualized in a result panel and then, choosing a randomization model, significance will be verified.The analytical significance is more immediate because the randomisation model (EDD) is implemented within the research task, so at the end of the matching a panel of results will be displayed with the motifs, the related occurrences and the p-value.

4.5.1 Results panel of statistical significance

The user can start the match search by clicking on the "Match" button so that the search task will be started. At the end of the search task, the results will be shown in the right side panel of Cytoscape ("Results Panel"). For each detected occurrence, its topology and the list of nodes and edges involved will be graphically shown (Figure 4.16). The user can select any match from the results panel. That occurrence will be highlighted in the target network (Figure 4.17). If the user selects the "Create a child network" option, when the match is selected, a new network with the topology of the selected match will be created and displayed.

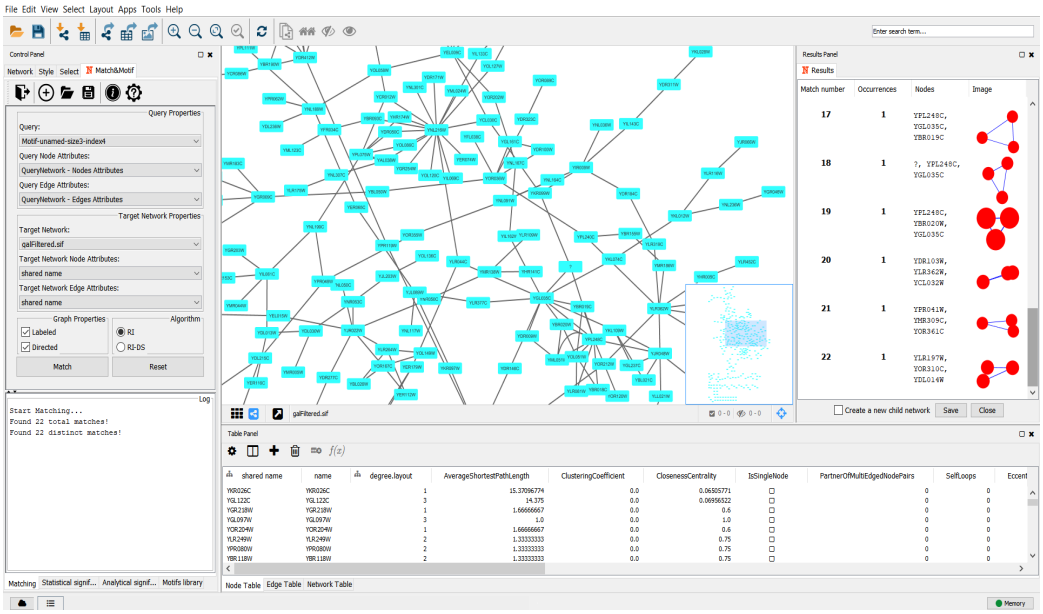

Figure 4.16: Display of results.

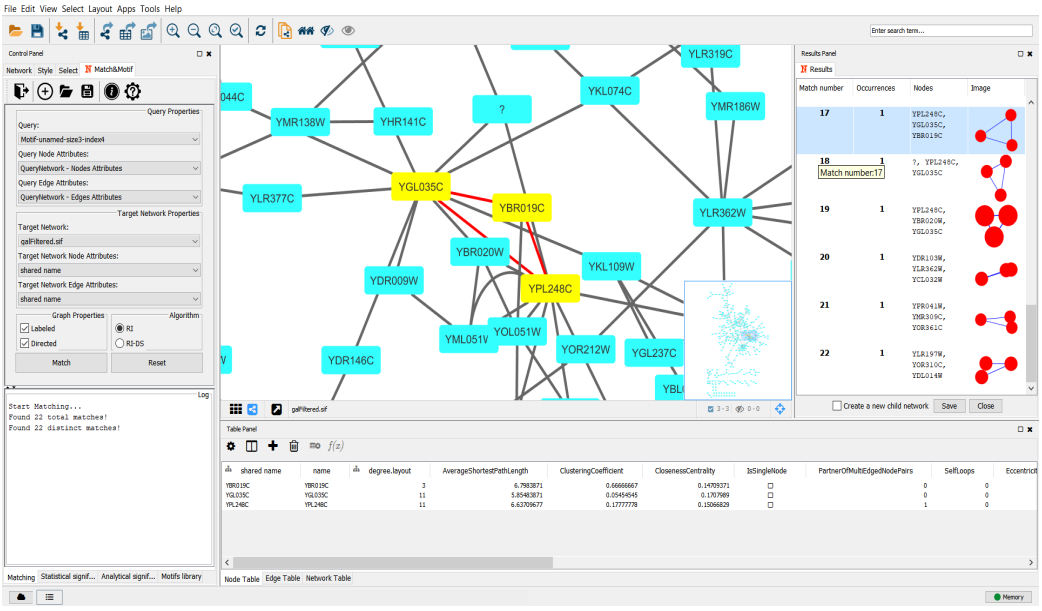

Figure 4.17: Highlight of a single match.

Selecting the "Significance" tab will display a tab containing the parameters for the evaluation of the statistical significance of a motif. The tab

consists of three subpanels:

1. The upper subpanel called "Options" allows you to select the number of random graphs to be generated to perform statistical tests (from 0 to 100). Furthermore, it is possible to indicate the seed for the generation of pseudo-random numbers optionally;

2. The central subpanel called "Metrics" allows the user to calculate some metrics relative to the target graph and the generated random graphs, for each randomisation model;

3. The lower sub-panel named "Models" allows to select the desired randomisation model and to set the main settings, as well as to start the statistical test by clicking on the "Start" button.

The calculation of the metrics includes as output the average degree, the clustering coefficient and the coefficient of assortment: the output is shown in dialogue and can be saved in CSV format optionally(Figure 4.18).

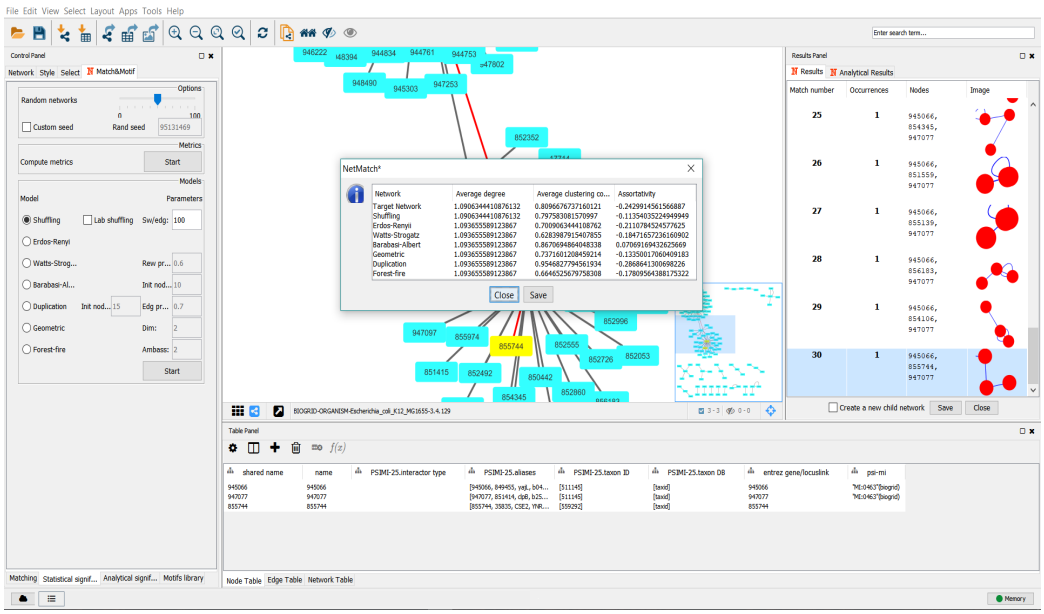

Figure 4.18: For the target network and for each network obtained through the 7 randomization models, the results are shown for average grade, clustering coefficient and assortativity coefficient.

Seven randomisation models are provided, and for each model, it is possible to set some options manually. For each model, parameters that are not set directly by the user, are estimated based on the number of nodes and edges of the selected target network. Initially, the computation is performed by generating a network with $|V|$ nodes having the same labels as the target network and without edges. Subsequently, edges are added in order to connect the various nodes until a new network, having $|V|$ nodes and $|E|$ edges, is obtained. By clicking on the "Start button", the statistical significance test can be started, the result will be shown in a dialogue message (Figure 4.19).

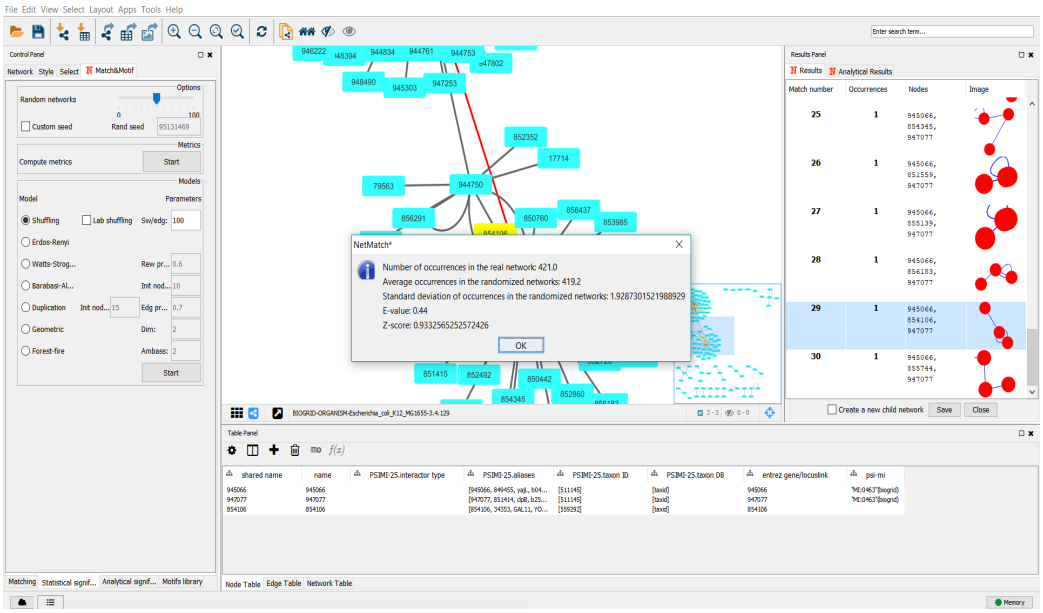

Figure 4.19: The number of occurrences in the real network, the average number of occurrences in the generated random networks, the standard deviation of occurrences in the generated random networks, the e-value and the Z-score are respectively shown.

4.5.2 Results panel of analytical significance

By clicking on the "Analytic match" the user will launch the match search task. At the end of the search task, the results will be shown in a right side panel of Cytoscape ("Results Panel"). In each row of the table will be re-

ported: the topology of the motif, the set of labels with which that topology is labelled, the occurrences of the motif and in the last three columns respectively the analytically calculated mean, variance and p-value (Figure 4.20). At the bottom of the panel, there is a "p-value selection" slider that allows the user to select only the rows below a certain p-value. After clicking on the 'Search' button, the lines corresponding to the search will be displayed in another panel. With a double-click on a line, a new panel called "Occurrences" will be displayed in which all the occurrences of that motif will be reported. For each occurrence it is reported: the topology, the set of labels, the names of the nodes and the SUID, that is a unique identifier assigned by Cytoscape during the creation of the network. The user can select any match from the "Occurrences" panel that will be highlighted in the target network (Figure 4.21). Finally, the results of each result panel can be saved to a .txt file.

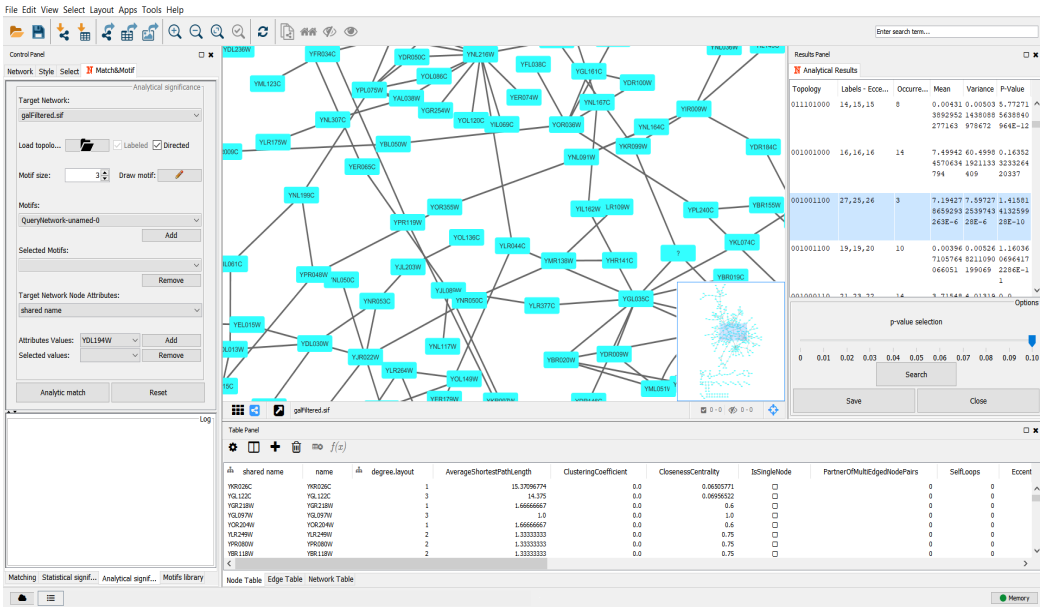

Figure 4.20: Result panel of the analytical significance. At the bottom of the panel the filter for the p-value and the save button is visible.

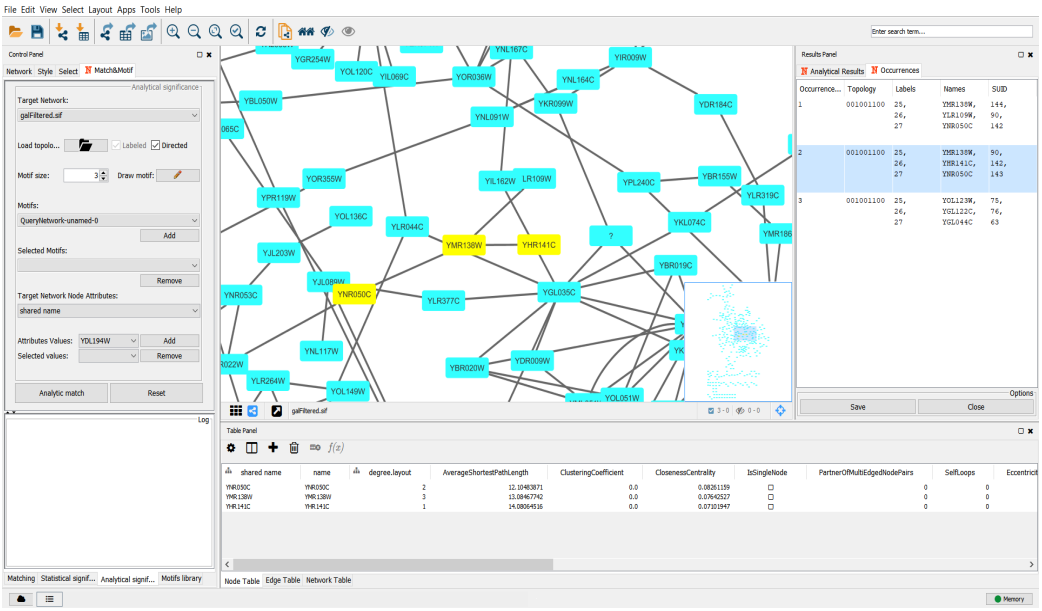

Figure 4.21: Display in another panel all the occurrences of a specific topology labeled with a specific set of values.

Chapter 5

Conclusion

In this thesis was presented Match&Motif, a Cytoscape app that allows to interrogate large biological networks in a performing way and to evaluate the statistical significance of the results obtained with respect to seven different randomisation models and the analytical significance obtained with respect to the EDD randomisation model. Match&Motif, based on two tools for the motif search detection (NetMatch* and FlashMotif), presents a graphical interface that is functional to the user's needs, which allows a quick and easy design of the network query. The user also has the ability to label the query network in four different ways, settable according to the research needs. Several tools and applications integrated into Cytoscape have been implemented that support the motif search and the evaluation of their significance such as CytoKavosk [30], GraMoFoNe [4] and Network Randomizer [16]. At the moment, however, no other app provides this freedom in the labelling of motifs and solves the problem of sub-graph matching in an efficient and performing way. Further developments could involve improving graphical interfaces, allowing node labelling only from query design, thus streamlining the selectable options from the panels and making the app more user-friendly. Finally, possible updates could concern the addition of other randomisation models for analytical significance.

Bibliography

[1] Inokuchi A, Washio T, and Motoda H. An apriori-based algorithm for mining frequent substructures from graph data. *Principles Data Mining Knowl*, 2000.

[2] Barabasi AL and Albert R. Emergence of scaling in random networks. *Science.*, 1999.

[3] McKay B. Number of graphs on n unlabeled nodes. *The NAUTY Page.*, 2011.

[4] G. Blin, F. Sikora, and S. Vialette. Gramofone: a cytoscape plugin for querying motifs without topology in protein-protein interactions networks. *In Hisham Al-Mubaid, editor, Bioinformatics and Computational Biology (BICoB'10), International Society for Computers and their Applications (ISCA)*, 2010.

[5] B. Bollobàs. Modern graph theory. *Springer Science & Business Media*, 2013.

[6] V. Bonnici, R. Giugno, A. Pulvirenti, D. Shasha, and A. Ferro. A subgraph isomorphism algorithm and its application to biochemical data. *Bioinformatics*, 2013.

[7] U. Brandes, M. Eiglsperger, J. Lerner, and C. Pich. raph markup language graphml. handbook of graph drawing and visualization. 2013.

[8] Cytoscape Consortium. Cytoscape user manual: Network formats. 2017.

[9] Watts DJ and Strogatz SH. Collective dynamics of 'small-world' networks. *Nature*, 1998.

[10] P. Erdős and A. Rèenyi. On random graphs i. *Publicationes Mathematicae*, 1959.

[11] Chung F. and Lu L. The average distances in random graphs with given expected degrees. *Proc Natl Acad Sci*, 2002.

[12] Chung F, Lu L, Dewey TG, and et al. Duplication models for biological networks. *J Comput Biol.*, 2003.

[13] Harary F and Palmer EH. Graphical enumeration. *Academic Press*, 1973.

[14] Picard F, Daudin JJ, and Koskas M. Assessing the exceptionality of network motifs. *J Comput Biol*, 2008.

[15] A. Ferro, R. Giugno, G. Pigola, A. Pulvirenti, D. Skripin, G.D. Bader, , and D. Shasha. Netmatch: a cytoscape plugin for searching biological networks. *Bioinformatics*, 2007.

[16] Tosadori G., Bestvina I., Spoto F., C. Laudanna, and Scardoni. Creating generating and comparing random network models with network randomizer. *F1000Research*, 2016.

[17] M. Himsolt. Gml: A portable graph file format. 1997.

[18] Wanhua Hu, Xiaodong Lin, and Kelong Chen. Integrated analysis of differential gene expression profiles in hippocampi to identify candidate genes involved in alzheimer's disease. *Mol Med Rep.*, 2015.

[19] Leskovec J, Kleinberg J, and Faloutsos C. Graphs over time: Densification laws, shrinking diameters and possible explanations. *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining.*, 2015.

[20] Grochow JA and Kellis M. Network motif discovery using sub-graph enumeration and symmetry-breaking. *Res Comp Mol Biol*, 2007.

[21] Ullmann JR. An algorithm for subgraph isomorphism. *JACM*, 1979.

[22] Babai L and Luks EM. Canonical labeling of graphs. *Proceedings of the Fifteenth Annual ACM symposium on Theory of Computing*, 1983.

[23] S. Lotia, J. Montojo, Y. Dong, G.D. Bader, and A.R. Pico. Cytoscape app store. *Bioinformatics*, 2013.

[24] Garey M and Johnson D. Computers and intractability: A guide to thetheory of np-completeness. *New York: W. H. Freeman and Co.*, 1979.

[25] Garey M and Johnson D. Computers and intractability: A guide to thetheory of np-completeness. *W. H. Freeman and Co.*, 1979.

[26] Kuramochi M and Karypis G. An efficient algorithm for discovering frequent sub-graphs. *IEEE Tran Knowl Data*, 2014.

[27] Penrose M. Random geometric graphs. *Oxford University Press*, 2004.

[28] Van Steen M. Graph theory and complex networks: An introduction. *OnDemand Publishing LLC-Create Space*, 2010.

[29] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 2003.

[30] A. Masoudi-Nejad, M. Ansariola, M.K.Z. Razaghi, A. Salehzadeh-Yazdi, and S. Khakabimamaghani. Cytokavosh: A cytoscape plug-in for finding network motifs in large biological networks. *PLoS ONE*, 2012.

[31] Giovanni Micale, Rosalba Giugno, Alfredo Ferro, Misael Mongiovì, Dennis Shasha, and Alfredo Pulvirenti. Fast analytical methods for finding significant labeled graph motifs. *Data Mining Knowl*, 2017.

[32] R. Milo, N. Kashtan, S. Itzkovitz, M.E.J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *Condensed Matter*, 2004.

[33] Kashtan N, Itzkovitz S, and et al. Milo R. Efficient sampling algorithm for estimating sub-graph concentrations and detecting network motifs. *Bioinformatics*, 2004.

[34] M.E.J. Newman. Assortative mixing in networks. *Physical Review Letters*, 2002.

[35] Sloane NJA and Plouffe S. Number of graphs on n unlabeled nodes. *The Online Encyclopedia of IntegerSequences.*, 2011.

[36] A.R. Pico, G.D. Bader, B. Demchak, O. Guitart Pla, T. Hull, W. Longabaugh, C. Lopes, S. Lotia, P. Molenaar, J. Montojo, J.H. Morris, K. Ono, B. Schwikowski, D. Welker, and T. Ideker. The cytoscape app article collection. *F1000Research*, 2014.

[37] J. Punin and Krishnamoorthy M. Xgmml extensible graph markup and modeling language 1.0 draft specification. 2001.

[38] Milo R, Kashtan N, Itzkovitz S, and et al. On the uniform generation of random graphs with prescribed degree sequences. *Arxiv preprint*, 2007.

[39] Milo R, Itzkovitz S, Kashtan N, and et al. Superfamilies of evolved and designed networks. *Science*, 2004.

[40] Milo R, Shen-Orr S, Itzkovitz S, and et al. Network motifs: simple building blocks of complex networks. *Science*, 2002.

[41] S R. Milo, Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 2002.

[42] F. Rinnone, G. Micale, V. Bonnici, G.D. Bader, D. Shasha, A. Ferro, A. Pulvirenti, and R. Giugno. Netmatchstar: an enhanced cytoscape network querying app. *F1000Research*, 2015.

[43] Boccaletti S., Latora V., Moreno Y., Chavez M., and Hwang D.U. Complex networks: Structure and dynamics. *Physics Reports*, 2016.

[44] Fortin S. The graph isomorphism problem. *Technical report*, 1996.

[45] Wernicke S. A faster algorithm for detecting network motifs. *Algorithms Bioinformatics*, 2005.

[46] Wernicke S and Rasche F. Fanmod: A tool for fast network motif detection. *Bioinformatics*, 2006.

[47] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *Nature*, 1998.

[48] D.B. West. Introduction to graph theory, volume 2. *Prentice Hall*, 2001.

[49] Elisabeth Wong, Brittany Baur, Saad Quader, and Chun-Hsi Huang. Biological network motif detection: principles and practice. *Briefings in Bioinformatics.*, 2011.