Department of
INFORMATICS, SYSTEMS AND COMMUNICATION

PhD Program: Computer Science Cycle XXX

Protein Interaction Networks: module detection integrating topological information and Gene Ontology

---

Danila Vella 798697

Tutor: Prof. Alberto Leporati
Supervisor: Prof. Giancarlo Mauri
Co-Supervisor: Prof. Riccardo Bellazzi

Coordinator: Prof. Stefania Bandini

ACADEMIC YEAR 2018

"Gli dei non hanno certo svelato ogni cosa ai mortali fin da principio, ma, ricercando, gli uomini trovano a poco a poco il meglio.

Senofane"

Abstract

The reductionist approach of dissecting biological systems into their constituents was proven successful in the first stage of molecular biology, in order to elucidate the chemical basis of several biological processes. This knowledge has helped biologists to understand the complexity of biological systems, furthermore pointing out that most biological functions do not arise from individual molecules. Therefore, the emergent properties of biological systems cannot be explained or predicted by investigating individual molecules without taking into consideration their relations. Thanks to the improvement of the current -omics technologies and the increasing understanding of molecular relationships, more and more studies are evaluating biological systems through approaches based on graph theory. In this context, Protein-protein interaction (PPI) networks are viable tools in understanding cell functions, disease machinery, and drug design/repositioning. As PPI networks involve hundreds to thousands of components, the use of these models, for clinical and biological applications, is strictly dependent on the Bioinformatics field. Bioinformatics sustains research, on one hand, making available infrastructures to collect both protein interactions and proteomics data; on the other hand, it provides softwares and tools to build and analyze these networks.

PPI networks are based on physical/functional interactions deriving from experimental and computational techniques and collected in public repositories. However, PPIs often lack reliability and do not cover all the interactions of an organism. Moreover, because of their biological nature, they are condition-specific. Thus, the PPIs detected in a specific biological context may not be valid to build a model of a system under different conditions. To overcome these issues, an alternative to building protein interaction network models consists in using large-scale quantitative proteomic data, i.e. the levels of expression of protein sets detected in condition-specific organic samples. While correlation within gene expression (i.e. co-expression analysis) is normally used to build gene co-expression networks, this technique is rarely used on proteomics data. However, it represents a complementary procedure which gives the opportunity to evaluate a biological context at system level, including organisms that lack information on PPIs.

PPI network structures are routinely analyzed by algorithms and tools so as to identify key proteins known as hubs/bottlenecks as well as functionally linked protein groups, called modules. Interpreting a PPI, however, is a particularly challenging task due to network complexity and limitations owing to their biological nature, such as the detection of small/sparse modules. Several algorithms were proposed for an automatic PPI interpretation, at first by solely considering the network topology, and later by integrating Gene Ontology (GO), in order to take into account the biological nature of the problem. However, to date, these methods provide only a topological interpretation of the networks, and further analysis is needed so as to infer biological knowledge regarding the phenomenon represented.

Based on these premises, this dissertation introduces the reader to protein interaction networks, and in particular to PPI based and co-expression based networks, in order to address different aspects of reconstruction and analysis.

Regarding the reconstruction of these networks, the new concept of evaluating large-scale proteomic data by means of co-expression networks has been investigated, focusing on several state-of-the-art studies. As a result, an analysis pipeline, specific to amyloidosis diseases based on protein co-expression networks, has been developed.

Concerning the analysis of these networks, a special attention has been devoted to topological and module analysis. Firstly, the most used metrics and mathematical models have been revised. Secondly, the problem of module identifications in PPI networks has been faced, considering characteristics and limitations of state-of-the-art techniques. As a result of this study, a novel algorithm has been developed known as *MTGO*, which stands for Module detection via Topological information and GO knowledge.

MTGO let emerge the biomolecular machinery underpinning PPI networks by leveraging on both biological knowledge and topological properties. In particular, it directly exploits GO terms during the module assembling process and provides a set of GO terms as output, thus easing network biological interpretation. A software version of MTGO, freely available at https://gitlab.com/d1vella/MTGO, has been produced and some examples of application have been explored, including the use on an experimentally-derived PPI network of Myocardial infarction. Moreover, for method validation, MTGO has been compared with state-of-the-art algorithms (including recent GO-based ones), using four different PPI Networks and three gold-standard target sets. MTGO shows largely better results than others when searching for small or sparse modules, while providing comparable or better results in all other cases. To conclude, the stability of the algorithm has been investigated, considering both the random components, on

which it relies on, and the presence of noisy PPI interactions on network models.

Acknowledgements

First and foremost, I would like to thank my Supervisors and Tutor, Prof. Riccardo Bellazzi, Prof. Giancarlo Mauri, and Prof. Alberto Leporati. Their expertise, excellent understanding and patient guidance had a major influence on this thesis, and this research work would not have been possible without their fundamental support.

I owe special gratitude to the Proteomics and Metabolomics Laboratory of the Institute of Biomedical Technology for their crucial contributions that have been of great value in this study. I want to thank all members of the Laboratory of Informatics and Systems Engineering for Clinical Research of the IRCCS Fondazione Salvatore Maugeri, for providing an excellent and inspiring working atmosphere.

I gratefully acknowledge Dr. Dario Di Silvestre and Ph.D Simone Marini, I enjoyed their interest in my research as well as the fruitful advises and support, I was delighted to interact with them.

My deepest gratitude goes to my parents and my sister for their love and support throughout my life; this dissertation would have been simply impossible without them.

Contents

To my family, Francesca, Gino e Lara...

Chapter 1

Introduction

A goal of the biomedical sciences is to transfer new findings in tools in order to enhance the current clinical practices, ranging from an accurate early diagnosis to the selection of proper therapeutic strategies [1]. The possibility to realize this idea is strictly related to the knowledge of the patho-physiological processes. Furthermore, their investigation by advanced strategies may result in a more preventive, predictive and personalized medicine [2]. The knowledge of the patho-physiological processes is supplied by various sciences, Biology, Chemistry, Mathematics, Informatics and Physics, applied to the investigation of complex biological systems.

The claim made by Francis Crick (1966) that "The ultimate aim of the modern movement in biology is to explain all biology in terms of physics and chemistry" well depicts the main concept that has dominated this research field for half a century, commonly known as reductionism. According to this term, an effective way to explain living processes consists in dissecting biological systems into their constituent parts, focusing attention on isolated molecules and their structure, so as to provide an understanding of the whole system [3].

Today, many findings suggest this approach has reached its limit to give rise to a new holistic concept already conceived by Aristotele: "The whole is something over and above its parts and not just the sum of them all". According this vision, the behaviour of a biological system does not arise from the specificity of the individual molecules that are involved, but rather through the way in which these components assemble and function together [4]. Living systems are perceived as an ensemble of highly interconnected networks, where the most cellular functions occur from a concerted action of multiple molecules [5, 6].

The success of this new concept has been sustained by the development of new high-throughput technologies over recent years, such as microarray and mass-spectrometry (MS), leading to the availability of a large amount of complex data. The mere size of these datasets requires specialized analytical tools, able to deal with large lists of objects. This holistic vision, alongside the availability of a

big quantity of biological data, has posed new challenges for scientific research, pursuing the aim to analyse cells, tissues or microorganisms at systems level [7, 8].

This scenario has lead to the development of Systems Biology, a set of interdisciplinary techniques and approaches aimed at interpreting a complex biological system through the analysis of relations between its component parts, which are studied not in isolation but as elements of the entire system. According to Systems Biology, one of the most used approaches consists in the use of graph theory to model the relations between molecular components. The main idea behind graph theory was introduced in 1736 by Euler [9]: "Graphs are simple mathematical objects used to describe interactions between different actors: a vertex, also called node, is an abstract and featureless object describing an entity in a certain context, while an edge is a connection, e.g. a link, between two vertices". This formalism has the ability to sum up complex high dimensional data into a net model.

In contrast to the study on individual molecules, biological networks aim to investigate how biological phenotypes, i.e. the biological traits characterizing a living organism, arise from cellular systems, conceived as networks involving bio-molecules, such as genes, proteins and metabolites [10]. In particular, the occurrence and development of a pathology can be viewed as the spatio-temporal change of the respective bio-molecular network. Therefore, the study based on these network models offers a new conceptual framework in order to investigate pathologies.

Early biological experiments revealed that proteins interacting with each other as well as with other bio-molecules, are the main agents of biological functions, therefore directly responsible in determining the phenotype of an organism. On the basis of this concept, the complexity of a disease can be viewed as the result of an intricate net of proteins involved in the process. This key concept, alongside the spread of technologies to measure the interaction among molecules [11], has lead to the increasing interest towards Protein-Protein Interaction (PPI) Networks. To date, there are many public repositories collecting known protein interactions in addition to many analysis specific tools which encourage the use of PPI networks, in order to pursue numerous objectives, such as protein function discovery [12], disease mechanism understanding [13] and drug target identification [14]. A common approach of PPI network modelling consists in building the model by retrieving the protein interactions from public repositories, starting from a set of proteins, identified in organic samples by

high-throughput technologies. These databases collect protein interactions obtained both by experiment and by computational prediction algorithms. However, the effectiveness of this model is limited by the accuracy and reliability of the information contained therein and the ability of the analysis tools to infer useful knowledge. The PPI networks built via public databases may not well represent the specific studied biological state, due to the fact that protein relationships are strictly dependent on the biological conditions. PPI networks are a result of different experiments performed on different conditions and therefore represent possible rather than active interactions (static rather than dynamic information). Therefore, following this modelling approach, the network interactions may not be specific of the biological context under study.

In recent years, the spread of Mass-Spectrometry technologies made available quantitative data of thousands of proteins from a biological tissue or sample, providing a snapshot of the studied biological system. To fully exploit these technologies, new procedures and methods should be developed in order to integrate this information into the network model, aiming to build a PPI network able to describe a specific biological state.

The use of PPI network is also influenced by the ability of dedicated analysis tools to provide reliable results. Indeed, due to the complexity of the biological system studied and the biological nature of the elements considered, these tools are often proved to be scarce in the extent of elucidating the organization principles of cellular functions and disease mechanisms.

The work presented in this thesis is based on this line of research. The contribution of my research considers mainly two fields. Firstly, the reconstruction of protein network models starting from the expression profiles of proteins from organic samples; following the approach of the newly proposed applications of the co-expression networks to proteomics data. Secondly, the development of a new algorithm for protein interaction network analysis, MTGO (Module detection via Topological information and Gene Ontology (GO) knowledge); which has been created with two main aims (I) to overcome certain issues linked to the biological nature of the problem and (II) to provide biological insight regarding the system represented by the network model. This double aim is obtained through the combination of information both from graph theory and a-priori biological knowledge concerning proteins involved.

1.1 Preliminary concepts

1.1.1 Bio-Molecular Interaction Networks

The main molecules involved in living organisms are DNA, RNA and proteins; DNA is a long macro-molecule, consisting in a sequence of smaller units, the nucleotides; RNA is a DNA-derived molecule involved in protein synthesis; and proteins are large macro-molecules consisting of one or more long chains of smaller units, the amino-acids. Much of twentieth-century biology has been an attempt to reduce biological phenomena to the behaviour of molecules. Following this research line, biologist inferred the existence of genes, functional parts of the DNA molecule, and their properties; as a result, several principles have been established, such as: each gene controls the synthesis of one protein, DNA contains genetic information, the genetic code links the sequence of DNA to the structure of proteins. In order to better understand these concepts, Section 1 in Chapter 2 is devoted to a brief description of elements of Molecular Biology.

To investigate biological phenomena, several techniques have been developed over the years to identify (I) which types of bio-molecular components are present in samples coming from a specific biological context (qualitative analysis), and (II) their respective abundances, commonly called expression profile (quantitative analysis). A briefly description of the path leading from Molecular Biology to the development of these modern technologies will be presented in Chapter 2, Section 2.

In order to understand complex biological systems starting from the knowledge of molecular components, the integration of mathematical models with experimental data has brought to a widespread approach called network analysis [15–17]. A biological network is a model representing the bio-molecular relationships in terms of nodes and edges. The nodes represent the components of the system, such as genes and proteins, or other molecules, while edges represent their relation, such as chemical transformations, regulatory relationships or functional associations. These biological networks are commonly investigated for extracting relevant knowledge regarding cellular functions and disease mechanisms [17, 18].

In the genomics field, the great availability of quantitative data deriving from high through-put methods for gene expression profiling [19, 20], has led to the development of techniques of network modelling. These methods are based on the inference of interactions directly from the gene expression level obtained by processing organic samples. These models, derived from the investigation of pairwise gene-expression profile associations, are commonly called

co-expression networks.

On the other hand, in the proteomics field, the spread of different techniques to directly investigate molecular interactions, such as yeast two-hybrid systems [21], X-ray crystallography [22] and protein chip technologies [23], has led to the diffusion of the so called Protein-Protein Interaction (PPI) network. These PPI networks can be obtained retrieving information from many public databases which collect protein interactions obtained from both experimental techniques and various computational approaches. The latter include techniques based on genes, the amino-acid sequence, the molecular structure and machine learning [24]. In Section 3 of Chapter 2, PPI networks will be presented, dealing with the different nature of the interactions and the state-of-art bioinformatics tools to manage them.

In this scenario, PPI networks are among the most important and widely studied networks [25, 26]. The widespread of PPI networks is justified by their versatility, promoting applications for: experimental data integration [27], protein function discovery [12], molecular mechanism comprehension [13], and drug discovery [14].

In recent years, as well as in the genomic field, the spread of MS-based techniques provides availability of large scale proteomics data. However, the potential of these high-throughput technologies have yet to be fully exploited [28, 29]; these data may improve the effectiveness of network-based approaches, thus providing new insight of the investigated systems. Therefore, the availability of these proteomics data represents a challenge for the bioinformatics field, which should provide new methodologies and computational techniques to infer context-specific protein network models.

1.1.2 PPI network analysis

The structure of biological networks is closely related to the biological functions performed by a system (cell or tissue) under a given condition. Starting from this point, many studies aim to face biological questions by investigating network models in terms of topology [30] and modular properties [31]. Given the network sizes, typically involving thousands of elements, the analysis often requires automated methods [32, 33]. For this reason, the bioinformatics field has received a boost towards the development of graph-theory based tools to manage and analyse network models, in order to extract biological insight to elucidate the bio-molecular mechanisms which are at the basis of physiological/pathological states of interest.

The term *topology* refers to the arrangement of graph elements, nodes and edges, in the network; therefore topological analysis means the investigation of the PPI network structure according to the graph theory principles at the basis of these models. A key point of topological studies is the definition of mathematical models and metrics to describe the network's properties in order to select the most relevant nodes and substructures that may be of biological significance. This theoretical area of research has led to the definition of (I) several metrics, generally called *centralities*, to investigate the property of nodes, edges, or whole networks and of (II) mathematical models to describe biological networks, such as Erdös–Rényi random graphs [34] and scale-free model proposed by *Barabasi et al.* [35]. Section 5 of Chapter 2 is dedicated to (I) introduce the main state-of-art *centralities* and (II) discuss the main state-of-art mathematical models used to describe biological network features. In this context, a common approach in analysing PPI networks is performed through the identification of sub-networks, or *modules*, showing specific topological and/or functional characteristics [5, 17, 36–38]. In biological networks, the term module has acquired three meanings: topological, functional, and pathological/disease. The analysis of the network structure allows to detect the topological modules defined as a highly interconnected group of nodes [31]. These nodes are often related to well defined molecular functions. Therefore, their detection in PPI networks can help identifying functional modules [39], defined as a group of functionally linked proteins/genes, such as connected by genetic/physical interactions, by co-expression relations, as well as members of the same molecular complex or biological pathway [5]. The comparison between pathological and physiological conditions has finally led to the definition of disease modules, i.e. a set of nodes with a putative key role concerning mechanisms impaired due to disease [14, 17]. Topological, functional, and disease modules are generally not fully overlapped and often, a single topological module can be linked to different functional or disease modules or vice-versa (Fig. 1.1).

Due to the complex connectivity of the biological networks, the identification of modules is a challenging task. Various methods have been proposed and most of them are exclusively based on network topology. Section 6 in Chapter 2 presents the module identification problem in PPI Networks in detail, in particular several unresolved issues linked with the biological nature of the context will be faced.

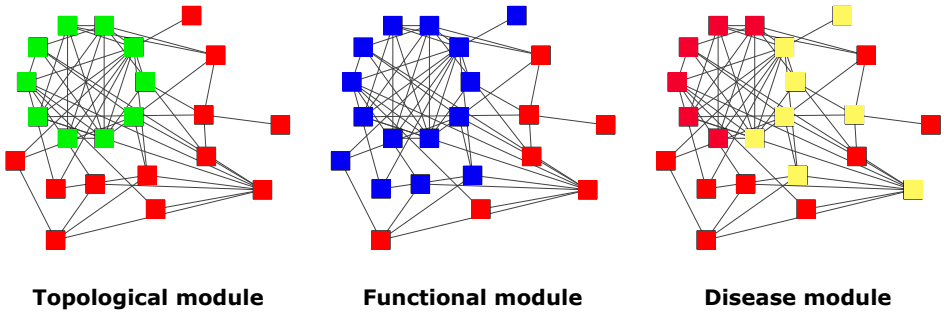


FIGURE 1.1: Example of topological, functional and disease modules not fully overlapped. The green nodes indicate a topological module, the blue nodes indicate a functional module, while the yellow nodes indicate a disease module.

1.2 Motivation and aim

1.2.1 Protein Co-expression Networks

A PPI interaction network can be used to elucidate a specific biological context since it is able to represent the protein interactions from which that context gives rise, i.e. the bio-molecular mechanism making possible the cellular functions linked to the context. Given that the presence/absence of the protein interactions is strictly dependent on the biological conditions, in order to pursue this objective it is necessary that network interactions derive directly from the same context which is the object of study. Nonetheless, in general, PPI networks are built by retrieving the protein interaction from public repositories, therefore deriving from experiments carried out in different biological contexts/conditions than the specific studied system. This aspect represents a weak point of the use of PPI networks in clinical/biological studies; a new procedure able to build protein interaction networks, exploiting the data produced by a specific biological context, should be searched. Regarding graph inference, a great boost consists in using information on gene/protein expression levels in order to predict network structures. Microarray techniques made available quantitative data regarding the expression level of thousands of genes in specific experimental conditions. These data led to building the so called co-expression networks, where the edges represent gene relations detected from the corresponding expression profiles. These models are based on the assumption that, in time or across several experimental conditions, genes with statistically similar (highly correlated) expression profiles are linked from a biological perspective. In the proteomics field, the opportunity to obtain large-scale data regarding protein profile expression has been explored more recently, especially

thanks to the enormous progress achieved in MS-based proteomics. Thus, the use of protein expression level data to infer co-expression networks has been explored through few studies [40–46]. Similar to PPI and gene co-expression networks, these networks have been evaluated at a topological level, in terms of edge rearrangement, as well as of modules, i.e. node groups associated with common cellular functions. Although different aspects, including data collection and network reconstruction, need to be improved, the preliminary results are proving this approach is promising. Protein co-expression networks are an alternative to evaluate large-scale proteomic data, in order to provide new hypotheses concerning key molecules acting in patho-physiological states [18, 43, 47–49]. These procedures may have important effects in clinical applications, opening the way towards the discovery of multiple biomarkers (key molecules indicative of disease presence) and their relationships. They could represent a first step in developing advanced diagnosis and prognosis methods.

In this thesis, one of the main scopes is the investigation of the co-expression network approach in order to evaluate large-scale proteomics data. Section 4 in Chapter 2 introduces this approach, discussing the state-of-the-art techniques to implement it as well as the issues linked to its application on proteomics data. On the other hand, Chapter 3 is dedicated to some practical examples; firstly, several state-of-the-art applications will be discussed, secondly, a pipeline of analysis specific for the amyloidosis disease study will be presented. The final aim is to propose a building procedure of condition-specific co-expression networks in order to provide a model able to give the possibility of generating biological insight and hypotheses on the influence of the various proteins on the pathology.

1.2.2 Module identification in PPI networks

With the growing amount of PPI data recorded, PPI network analysis has become a common practice in proteomics, genomics, and computational biology, with direct translational impacts in pharmaceutical and medical applications (such as drug repositioning or drug target discovery) [50]. With hundreds to thousands of nodes, and even more edges, PPI networks are impossible to manually analyse in detail. For this reason, more and more algorithms have been proposed so as to automatically identify functional parts of these networks, called modules.

These methods provide a set of topological modules as output, commonly also known as clusters, grouping together nodes sharing topological properties, such as characterized by a high density of connections. These techniques face

the problem only from a topological/graph-based perspective, often forgetting the biological nature of the input model. A discussion on the limits of this techniques will be presented in Chapter 2 Section 6.1.

Moreover, traditional approaches provide solely a set of clusters as a result. As a consequence, the output is far from providing a PPI network evaluation from a biological perspective; therefore, a second step of analysis is needed to investigate the biological role of clusters. To supply this lack, after module detection, biologists routinely use an *enrichment analysis* to assess the biological relevance of the identified clusters [51, 52] (Fig. 1.2). *Enrichment analysis* is a technique based on statistical tests to evaluate if a set of genes/proteins are enriched in common functional properties [53]. This analysis is generally based on Gene Ontology (GO) [54], a collection of terms aimed at describing all the functions at the basis of living systems (see Section 1 in Chapter 2 for details).

As a result, two typical steps of PPI network analysis can thus be outlined: *module identification*, which consider the problem from a topological/graph-based perspective, and *enrichment analysis*, which consider the problem from a biological perspective.

Generally, the standard methods for enrichment analysis treat each gene/protein as isolated objects. However, in the last few years, several network-based enrichment approaches have emerged, taking into consideration also interactions among bio-molecules [55–57]. On the other hand, to improve the accuracy of module identification, the integration of functional information is increasingly used in several algorithms [58–61]. These methods often exploit the a-priori knowledge provided by GO; for example, it can be used to compute a similarity score which measures the edge weights and drives the module detection [62, 63]. These recent methods will be presented in Chapter 2, Section 6.2.

As a result, the two typical steps of PPI network analysis are becoming ever closer. This is due to the fact that *module identification* algorithms seek to integrate a-priori functional information (via GO knowledge) and *enrichment analysis* attempts to involve graph theory principles.

The aim of this PhD has been to conceive a novel algorithm of module identification specific for PPI networks, pursuing (I) the idea to integrate topological analysis and biological analysis in a single step, and (II) the objective to improve the detection of hard-to-identify modules, such as small and sparse, thanks to the use of GO knowledge. This algorithm is known as MTGO, Module detection via Topological information and Gene Ontology knowledge. In other words, it simultaneously takes into account limits and properties typical of both the topological and biological nature of the network model treated. In fact, MTGO relies on both a graph-theory-based measure, Modularity [64], and GO knowledge

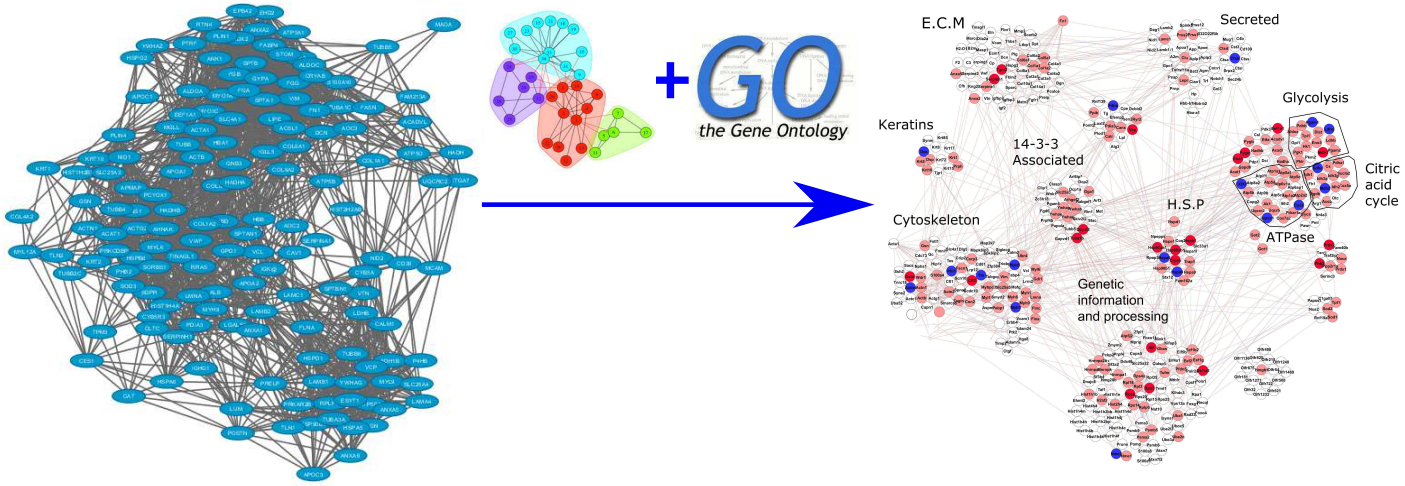


FIGURE 1.2: Procedure used to identify/predict modules in biological networks. The network structure is used to identify groups of highly connected nodes by graph clustering algorithm, while the GO annotations are used to improve the accuracy of the cluster prediction. The final result are clusters of nodes highly connected and related to functions/processes significantly enriched, thus acting at the basis of the emergent phenotypes.

regarding proteins involved in the network. Moreover, MTGO output consists both in a set of clusters (i.e. topological modules) and a set of GO categories, representing the functional modules. In this way, each cluster is tagged with a specific GO term, describing its biological nature.

The algorithm will be presented in detail in Chapter 4, and further insight will be provided in the Appendix Section. MTGO has been implemented in a software version; its use and case study applications will be shown in Chapter 5. To validate MTGO algorithm, the comparison with seven state-of-the-art algorithms (including recent GO-based ones) have been performed. Three different gold-standard protein complex sets have been used as target sets to compare the predicted modules to each algorithm. The algorithms were tested on four different PPI networks, from *Saccharomyces Cerevisiae* and *Human* organisms. Several metrics were used to compare the target sets with the predicted modules, including *Accuracy*, *Maximum Matching Ratio*, *F-Measure* and *Composite Score*. The biological quality of the predicted modules was measured through GO Term Finder, a software used to perform GO enrichment analysis [65]. Moreover, a statistical test was used in order to compute a p-value for each GO term provided by MTGO, so as to verify that a GO term is statistically significant compared to the protein set of the cluster to which it is associated with. Furthermore, since a weak point in state-of-the-art algorithms is the detection of small or sparse modules, MTGO ability to detect these specific sets of modules was evaluated. All these results will be presented in Chapter 6, including a short short discussion about run time.

As MTGO rely on random components, stability of the results across different runs is a critical aspect of the algorithm. In addition, when evaluating an algorithm specific for PPI Networks, an important aspect is the stability in presence of false positive and false negative edges. For these reasons, two different stability analyses were executed in order to evaluate MTGO performance. The first was used to evaluate the stability of the result over many runs starting from a same input and in order to consider the range of variability introduced by the random components of the algorithm; the second was used so as to evaluate the robustness of the output clusters when input is affected by noise and uncertainty. These analyses will be discussed in Chapter 7.

Chapter 2

Protein Interaction Network Background

In this chapter, the basic concepts will be introduced for a proper understanding of this dissertation. Starting from the key concepts of the Molecular Biology field, protein interaction networks will be discussed, with insight regarding protein co-expression networks and state-of-the-art analysis methods, specific for these models. Several parts of the chapter are based on the scientific paper "From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data" by *Vella, D., Zoppis, I., Mauri, G., Mauri, P., & Di Silvestre,D.*, published on *EURASIP Journal on Bioinformatics and Systems Biology*.

2.1 Organisms and cells

All living organisms are composed of small cells. A cell can also exist as independent organisms to the effect that they can grow, reproduce, convert energy, control their internal working, respond to their environment, and so on [66]. More evolved organisms, including humans, are communities of cells, performing specialized functions and coordinated by complex systems of communication. Cells vary widely in size, shape and functions, as well as in chemical requirements and activities. Some appear to be specialized factories for the production of particular substances; some are engines, like muscles, burning fuel to do mechanical work, while others are electricity generators. Nevertheless, cells resemble one another to a surprising degree in the details of their chemistry, sharing the same machinery for the most basic of functions. Cells consist of molecules participating in chemical reactions. There are four basic types of molecules implied in life: (I) small molecules, (II) proteins, (III) DNA and (IV) RNA. Proteins, DNA and RNA are collectively known as macromolecules.

Small molecules can be the building blocks of macromolecules or can have independent roles, such as signal transmissions or a source of energy or material for cells. In addition to water, other relevant examples are sugar, fatty acids, amino-acids and nucleotides. The rest of living material consists of macromolecules [67].

2.1.1 Proteins

Amino-acids molecules are the main constituents of proteins. Short chains of amino-acid constitute peptides and numerous poly-peptides arranged in a biologically functional way constitute a protein. There are 20 different amino-acid molecules which have peculiar properties. When combined in different sequences and polypetides, they give rise to thousands of different proteins. Proteins are the workhorses of the cell and are the most abundant and functionally versatile of the cellular macromolecules. The most important classes are:

- transmembrane proteins, which play a key role in the maintenance of the cellular environment and communication between cells;
- structural proteins, which are the main basic building blocks of organisms;
- enzymes, which increase the rate of chemical reactions (catalyse) such as altering, joining together or dividing other molecules. Together, the reactions and pathways they make is known as metabolism. Generally, enzymes are very specific and catalyse only a single type of reaction, however, the same enzyme can play a role in more than one pathway.

  Protein interactions underlie the assembly of macromolecular machines, mediate signalling pathways in cellular networks and control cell-to-cell communication. Nearly 650,000 interactions regulate human life [68]. Proteins are the core ingredient of this thesis. In fact, the long term target of the scientific line of this thesis is the investigation of protein functionality in a specific biological context through the exploration of their mutual interactions. In order to approach this aim, the relations between proteins will be later investigated through a knowledge extraction from condition-specific proteomics data and their integration with a-priori knowledge.

2.1.2 DNA, RNA and Genes

Deoxyribonucleic acid (DNA) is an informational molecule constituted mainly by sequences of nucleotides. Its main role consists in containing the information required to build all proteins of an organism, hence the cells and tissues of that

organism. Due to this feature of information storage, it is often compared to a set of blueprints or recipe, or even a code. The DNA segments that carry this genetic information are known as genes [69]. The gene definition given by Lodish et al. in [66] is: "A Gene is the entire nucleic sequence that is necessary for the synthesis of a functional polypeptide." Biologists once believed in the paradigm "one gene - one protein". Today, we know that this is no longer true, as due to biological phenomenon such as alternative splicing and post-translational modifications, one gene can produce a variety of proteins. Although the product of a gene in general is a different entity from a protein, for practical reasons, the term gene product is often used as synonymous with protein. The synthesis process, starting from DNA and ending with proteins, is mediated by another key molecule, namely Ribonucleic acid (RNA). RNA was discovered after DNA; it is constructed from nucleotides, as DNA, but differs in several important details. RNA functions are necessary for protein synthesis. Its main role consists in reading the genetic information contained in DNA molecules (transcription process) and allowing for protein synthesis (translation process). For this reason, RNA molecules are often referred to the term *transcripts*.

2.2 From Molecular Biology to modern technologies

A phenotype is the ensemble of observable traits/characteristics in a living organism; for example, a human phenotype clearly includes eye and skin colour. Proteins dictate virtually every reaction in the cell and are thus directly responsible for observable characteristics. For instance, a slight variation in the activity of an enzyme for pigment synthesis in a plant may result in white rather than orange flowers. In contrast, a genotype is the genetic endowment of the individual. Through the mediation of proteins, a phenotype is the result of a genotype and its interaction with several environmental factors [29]. Much of life science research has been focused on understanding the complex relationship between genotype and phenotype, in other words, how the information encoded in the genome is expressed and modulated by external/internal factors in order to generate a specific phenotype. This interest is likewise motivated by the fact that understanding how genes and proteins act, in order to produce a specific phenotype, allows to better understand specific pathological conditions and diseases. Over the past decades, the “one gene-one protein-one function” paradigm proposed by Beadle and Tatum [70] has dominated the ideas of biologists and guided the research into the Molecular Biology field. This paradigm requires a direct link between gene and protein function, implying that knowledge of all

genes and their translation products can explain biological functions and cellular functioning. The series of processes in which the information encoded in a gene is used to produce a functional protein is commonly known as *Gene expression*. In detail, only a fraction of the genes in a cell are expressed at any one time, and only these expressed genes may determine the protein production of that cell in a given time. Therefore, the study of gene expression processes (which include the transcription and the translation) plays a critical role in determining what proteins are present in a cell and in what amount. As protein dictates cell functions, the study of expressed genes has gained great attention in recent years. In the genomic age, powerful technologies have been developed to support research at a global scale within the Molecular Biology paradigm. These technologies attempt to measure the level of the gene expression from cell and tissue samples, characterized by specific biological conditions/contexts. In particular, these techniques include methods to identify (I) which types of bio-molecular components are present in samples (qualitative analysis), and (II) their respective abundances, commonly known as expression profile (quantitative analysis). The most recent techniques include genome sequencing in order to identify all protein-coding genes of a genome (RNA/transcript molecules), such as Microarray [71] and proteomic methods to identify and quantify the proteins in a biological sample, such as Mass Spectrometry coupled with liquid chromatography (LC-MS) [7].Moreover, a great interest has been addressed in the development of techniques in order to detect protein interactions, such as yeast two-hybrid (Y2H) method [21]. These technologies, alongside many others, have led to the availability of a big quantity of data, commonly referred to as -omics data. These recent advances in the Molecular Biology field have led to the development of new disciplines, with the main being genomics and proteomics. Genomics is a multidisciplinary approach to studying genes, their products, and the interactions among them in order to mediate physiological responses; proteomics is a branch of research aimed at assessing protein activities via their modifications, their localization, and their interactions, in order to improve our understanding of system level cellular behaviour [72]. In this context, clinical research has shown an increasing interest towards these disciplines, and in particular towards proteomics, hoping to benefit both by the identification of new drug targets as well as the development of new diagnostic markers. New development of these disciplines has been mediated by the success obtained in the computer science field, which in turn has received a boost towards the creation of new systems for data storage and analysis. These have led to the birth of many registries, databases and public repositories so as to manage the large

amount of data to facilitate and support research for biological and clinical applications. Thanks to these developments, we are today in the post-genomic era in which several bio-techniques have been developed to utilize -omics information in order to explain the biology and physiology of cells or tissues, as well as to explore the pathogenic mechanisms of various diseases.

Gene Ontology

At this stage, it is worth mentioning the Gene ontology (GO) project, the major bio-informatics initiative to unify the representation of gene and gene product/protein attributes across all species [73]. The objectives pursued by the Gene Ontology Consortium is manifold. It aims at maintaining and developing a controlled vocabulary of gene and gene product attributes, providing tools for easy access to data in order to enable the functional interpretation of experimental data through the use of GO. For this purpose, three independent ontologies accessible on the World-Wide Web are being constructed: biological process, molecular function and cellular component [54]. Thanks to annotation data, each GO term is assigned to a specific set of gene products/proteins, i.e. the set of bio-molecules involved in the specific function represented by the term.

2.3 Protein interaction networks

Systems Biology is a discipline which collects several approaches sharing a common aim: the understanding of biological processes performed by a system in terms of all its components. These approaches are based on the assumption that all phenomena can be viewed as a web of relationships among elements [74]. This concept has risen thanks to recent advances in technologies which have made available a large quantity of data regarding bio-molecules acting at the basis of biological systems. According to this holistic view, *graph theory*, that is to say network science, has been receiving even more increasing interest. The birth of graph theory dates back to the year 1736 in Euler's description of a map of seven bridges of the river in Konigsberg (a city in Prussia). Its description lays the foundations of *graph theory*, in fact, Euler presents the map in abstract terms, eliminating all features except the list of land masses and the bridges connecting them. Since then, graph theory has developed into an extensive branch of mathematics. Today, it is a powerful abstracting machinery which allows for the modelling of several types of systems, both natural and human-made, ranging from Biology to the science of Sociology [75].]. The main concept at the basis of

network science is the *graph*. A graph, also known as a network, provides a system representation in terms of relationships among the elements that compose it; a set of nodes V stands for the elements of the system, while a set of edges E stands for their relations. Mathematically, we refer to a graph as G = (V, E).

The development of systems biology approaches based on graph theory [76–78] is receiving a great boost due to the improvement of proteomics technologies [79, 80], in particular those increasing knowledge regarding protein-protein interactions (PPI) [21]. As a result, the use of network models, which combine information from PPI and protein/gene expression levels, is widespread today in the study of biological systems [27, 81, 82].

With respect to biological networks, nodes may be associated to attributes representing characteristics of interest, such as expression levels or GO terms. In the same way, edges may possess attributes describing the relation between nodes, for example indicating the strength of the interaction or its reliability; edges may also be directed or undirected. In this dissertation, we shall mainly deal with undirected edges. Using the framework described in Fig.2.1, a protein interaction network is defined as a complex graph, where nodes are proteins and edges represent their relation, generally physical or functional, as proposed by Vidal *et al.* [18].

2.3.1 PPI: physical and functional protein interactions

A protein interaction network usually refers to physical PPIs [83], but several meanings have been attributed to this term. In fact, a group of proteins working together to perform a biological function not necessarily are in direct contact, but their relation may be of regulation or influence, for example, making use of intermediary molecules. For this reason, the term PPI has not only been exclusively used to indicate a physical contact between proteins, but also proteins connected by functional links. It is important to bear in mind that proteins participate to physical-chemical connection depending on the biological context where they are [84]. Thus, the interactions measured in a specific experiment could not occur in any cell or at any time. However for simplicity, if two interacting proteins are experimentally identified in a given sample, it is generally assumed they also interact in a specific studied context, thus their relation is reported in the reconstructed PPI network representing the specific context.

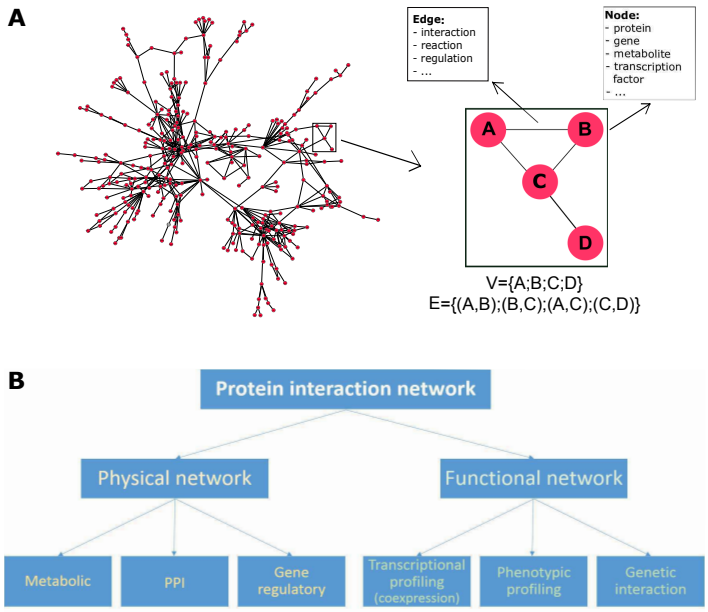


FIGURE 2.1: Biological networks. **A** Nodes may represent several types of biological elements, while the edges describe the nature of their relationship. If A and B are two nodes connected by an edge, (A,B) $\in$ E, B is a neighbor of A or A and B are adjacent. **B** Protein network classification proposed by Vidal *et al.* [18]

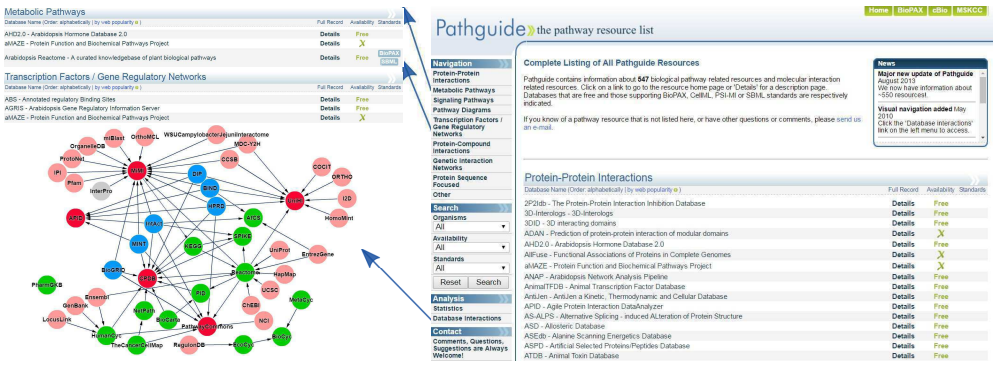


FIGURE 2.2: Pathguide website [94]. A repository containing information about 547 resources of molecular interactions and pathways

2.3.2 PPI: detection, storage, and analysis tools

The main approaches to demonstrate physical interaction between proteins are the yeast two-hybrid (Y2H) method and the tandem affinity purification coupled with mass-spectrometry (TAP-MS) [21]. To reduce the identification of false interactions, these experimental data are complemented with computational methods of prediction [85–87]. Other methods are used to identify functional relationships, and most of them rely on protein expression data [46], analysis of gene co-expression patterns [88], and analysis of sequences or phylogenetic properties, as Rosetta Stone or Sequence co-evolution methods [89]. Both physical and functional PPIs are stored in public repositories. The most popular include MINT [90], IntAct [91], STRING [92], and HPRD [93]. The latter specifically collects interactions related to Homo Sapiens, while other databases like STRING collect different kinds of interactions (from experiments/biochemistry, annotated pathways, gene neighborhood, gene fusion, gene co-occurrence, gene co-expression, and text-mining) and different organisms. A useful list of repositories presented by De Las Rivas *et al.* [83] provides a classification in categories (primary, meta, and prediction database) according to method used to detect interactions. Moreover, an exhaustive collection of more than 500 databases is available in the Pathguide website (Fig. 2.2) [94].

The development of computational tools to retrieve, visualize, and analyze biological networks is a key aspect of the systems biology studies, like the production of accurate -omics data and the collection of reliable molecular interactions. The most broadly adopted software include Cytoscape and its plugin [95], VisANT [96], atBioNet [97], PINA [98], and Ingenuity [99] which represents a commercial solution. On the contrary, Cytoscape is a software developed

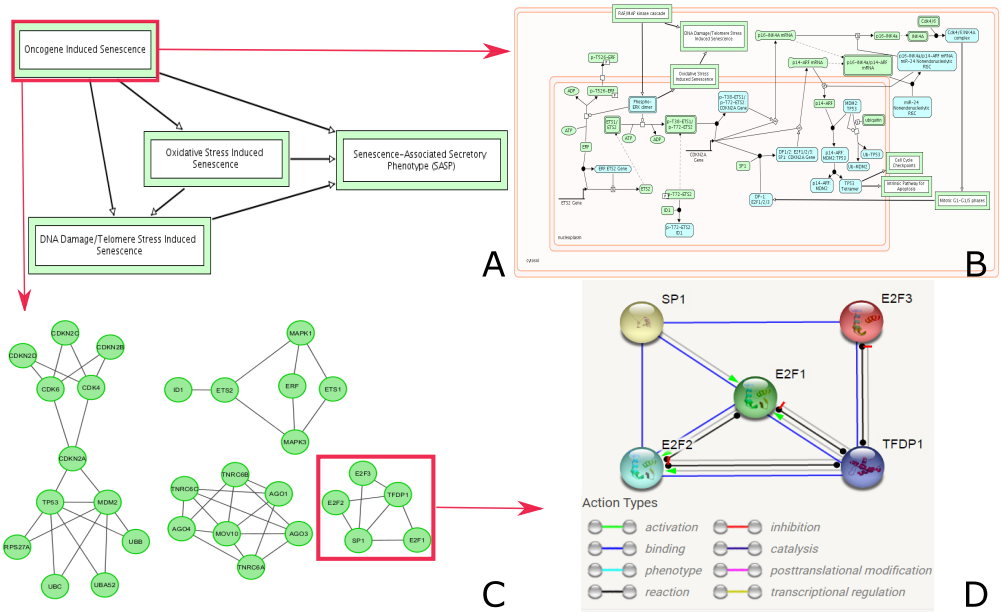


FIGURE 2.3: ReactomeFIViz: from disease pathway to PPI network. Main steps to obtain a protein functional and a physical protein network, starting from a specific pathway (oncogene induced senescence). Using ReactomeFIViz, pathways can be visualized in relation with others (a), can be detailed as a diagram showing all inter-molecular relationships (b), and as a protein functional interaction network (c) showing just the relation among proteins that cooperate to perform a given molecular function. Finally, starting from a group of protein of interest, it is possible to obtain a network of protein-protein interactions by STRING; in the reported example, the interactions shown are limited to physical type, in particular binding, activation and inhibition (d)

by an international consortium of open-source developers. Figure 2.3 shows a possible use of the ReactomeFIViz Cytoscape's plugin to obtain networks (both functional and physical) associated with a given biological function. ReactomeFIViz is focused to pathways and patterns related to cancer and other pathologies [100]. The reconstruction and the analysis of networks are of importance in the context of biomedical research, and detailed reviews about network models to investigate complex diseases have been published by Cho *et al.* [101] and by Vidal *et al.* [18]. Both works show how functional and physical links can be used to investigate disease mechanisms, and PPI networks emerge as effective model to evaluate different biomolecules acting in complex biological systems, thus providing an insight on phenomenons involved in a given physio-pathological context.

2.4 Co-expression networks

The large amount of data produced by microarray and RNA-seq technologies has driven the need for methods to objectively extract meaningful information. A widely adopted approach to evaluate transcript levels is based on statistics which measure the dependence between variables [102]. In this context, the use of co-expression scores is of great interest for clinical/biological research (see Paragraph 2.2 for an explanation of expression process). These scores, through the processing of transcript profiles, attempt to define a relation between pairs of transcripts. These approaches are based on the concept that transcript profiles of time series, or result of specific perturbations, may be indicative of the acting mechanism between transcripts. For example, several studies have shown that functionally related genes sharing Gene Ontology (GO) terms usually present higher co-expression scores [103]. The co-expression score computation represents the first step of inference in order to build a *co-expression network*, defined as an undirected graph, where nodes correspond to genes and edges indicate significant co-expression relationships, but not causality. The aspect of causality is faced in the context of transcriptional regulatory networks [104], where pairs of genes are considered in a systemic perspective of cooperation, including co-regulation, activation/suppression, and indirect control through the action of RNA, proteins, metabolites, and epigenetic mechanisms. This complexity makes the inference of transcriptional regulatory networks difficult by using exclusively transcriptional profiles. In fact, in this type of model, in addition to co-expression, the following levels of inference require more information and different modelling techniques, including Boolean networks, Bayesian networks, or differential equations (ODEs) [104]. Many are the network models used to interpret and represent bio-molecular mechanisms; that discussed in this dissertation is the co-expression based model. Regarding these models, the most widespread are gene co-expression networks, as more progress has been made in obtaining information on genetic molecules (DNA, RNA). Conversely, global analysis of proteins was exceedingly difficult in the past. However, the availability of information on proteins is crucial, as proteins can interact with all of the other classes of bio-molecular components, influencing all cellular activities at various levels. In some studies, several attempts to infer protein abundance from RNA molecules have been made in order to supply this lack of information, since RNA is directly responsible for protein synthesis. However, changes in protein abundance cannot be simply inferred from RNA data; many studies show how abundance of RNA poorly correlates with protein abundance [105].

Over the last 10 years, the improvement of Mass-Spectrometry (MS) based

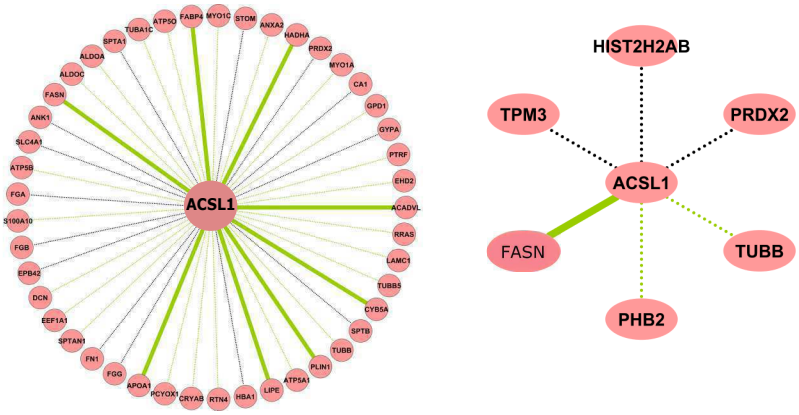


FIGURE 2.4: The figure shows the ACSL1 protein and its neighbors in two co-expression networks obtained by processing the protein expression profiles of a control group and a group of patients affected by amyloidosis disease. In the considered groups of samples, ACSL1 shows a different interaction number. It suggests that this protein may have a key role in the emergent phenotypes. Green edges represent a positive correlation between the expression profiles, while black edges indicate negative correlations. The thick edges indicate known interactions present in public repositories as PPI.

proteomics has given a great boost to large-scale proteomics analysis, making available the expression profiles of thousands of proteins per sample [55]. Therefore, MS offers a new range of opportunities to improve existing models of how phenotypes emerge [29]. Due to the similarity of the output between genomics and proteomics technologies, the use of proteomic data to infer protein co-expression networks has been recently explored in order to investigate the role of proteins in specific physio-pathological contexts. For example, variations of the co-expression score are evaluated so as to select relevant proteins whose number of bio-molecular interactions changes under specific conditions or perturbations [44] (Fig. 2.4).

Although different aspects need to be improved, this approach takes into account condition-specific protein relationships. Thus, with respect to conventional methods (i.e. *via* PPI public repositories, as discussed in Paragraph 1.3), it represents an alternative approach to gain a deeper insight of the proteins characterizing a given system.

2.4.1 Aspects of construction

To build a co-expression network, an important aspect concerns the computation of a co-expression score, which weighs the correlation of two genes/proteins in

TABLE 2.1: Measures of dependence between two variables

Co-expression measures	What measures?	Input/Output	Features
Pearson's correlation (PC)	Tendency to respond in opposite/same direction across different samples	Input: Gene expressions value Output: • [0,1] both genes increase • [-1,0] one increase and other decrease	• Sensitivity to outliers • bad array of expression level can determine positive PC value • Measure linear relations
Spearman's correlation (SC)	Tendency to respond in opposite/same direction across different samples	Input: Ranking values from expression levels in samples Output: • [0,1] both genes increase • [-1,0] one increase and other decrease	• Robust to outliers • Detect non-linear associations
Mutual information	Reduction of uncertainty of a gene given the knowledge about other gene	Input: Gene expression values Output: • 0 there isn't interdependence • > 0 there is interdependence	• Measure complex non-linear type relations (Rarely present in biological data) • More samples are needed than PC,SC • Time consuming computation
Kendall	Correspondence/compatibility among two rankings	Input: Gene expression value Output: • 1 perfect correspondence • -1 rankings exactly inverted	• Similar to SC • Robust to outliers • assumes fewer values than SC in the range [-1,1]

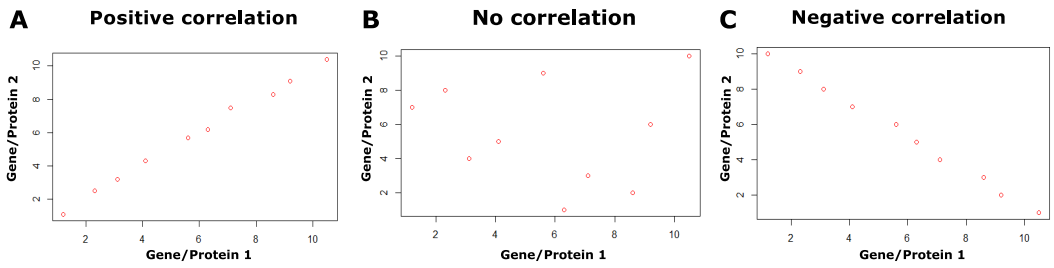


FIGURE 2.5: Possible cases of correlation between two variables.
a Positive correlation. b No correlation. c Negative correlation

response to the considered conditions (Fig. 2.5).

To address this issue, metrics to measure gene/protein co-expression have to be considered (Table 2.1); the most used metrics include Pearson correlation (PC), Spearman correlation, Kendall correlation, and mutual information [102, 106]. Various methods have been also proposed to define proper thresholds to select significant relations. Some of them are based on statistical analysis [107] and on network properties [108], while other interesting approaches aim to minimize the false positive links [109]. Finally, not less important is the selection of appropriate experimental samples/conditions to be processed. A condition-independent analysis is used to find relations of co-expression actual in different biological contexts; on the contrary, a condition-dependent analysis aims to find relations associated with specific phenotypes. The co-expression score computation may be faced by using any statistical or computational tool that allows to evaluate the dependence between variables. Some tools have been specifically designed to construct, visualize, and analyze co-expression networks. For example, the ExpressionCorrelation Cytoscape's plugin allows to process microarray data and provides a similarity matrix computed by PC [110]. In addition to being user-friendly, the main advantage of this tool is that the reconstructed networks are directly imported in Cytoscape (Cytoscape software is presented in Paragraph 1.2.2), where it may be evaluated by other plugins.

In this context, Weighted Gene Co-expression Network Analysis (WGCNA) is one of the most used approaches to build and to analyze gene co-expression networks [111], and it has been recently adapted for proteomics use also [40, 41, 44–46, 100, 112]. It provides a weighted network model by converting a co-expression measure to a connection weight. The network is fully specified by an adjacency matrix, where the component a_{ij} defines the strength of connection between nodes i and j. The value of a_{ij} is computed through the coexpression similarity s_{ij} (2.1), defined as the absolute value of correlation among the profiles of nodes i and j. It can be defined in two ways: to obtain an unweighted

network, the s_{ij} is filtered by a threshold τ such that a_{ij} takes on value [0,1] (hard-thresholding) (2.2), while to obtain a weighted network a_{ij} is defined by a power adjacency function (soft-thresholding) (2.3):

$$\mathrm{s_{ij}} = |\mathrm{cor}(i,j)| \tag{2.1}$$

$$a_{\mathrm{ij}} = \begin{cases} 1 & s_{\mathrm{ij}} \geq \tau \\ 0 & s_{\mathrm{ij}} < \tau \end{cases} \tag{2.2}$$

$$\mathrm{a_{ij}} = {s_{\mathrm{ij}}}^{\beta} \tag{2.3}$$

WGCNA pipeline has been implemented as a package of R software, which is a free software environment for statistical computing and graphics. Thanks to its modeling capabilities and its flexibility, R is one of the most used software in bioinformatics field. The R WGCNA package provides the possibility to use different types of metrics, including Spearman, Pearson, Kendall correlation (function *cor*, see Table 2.1), and the biweight midcorrelation (function *bicor*) [113]. Spearman's correlation is a non-parametric measure of correlation (see Table 2.1). Pearson's correlation can be used when data are normally distributed, but it is quite susceptible to the presence of outliers (see Table 2.1). Biweight midcorrelation is a similarity measure based on the median rather than the mean, for this peculiar characteristic it results less sensitive to outliers. Data from gene samples are often characterized by a high presence of outliers, thus the recommended correlation measure is the biweight midcorrelation, since it is more robust to outliers. The package allows to compute both the correlation and the Student p-value for multiple correlations in case of missing data, thanks to the functions *corAndPvalue* and *bicorAndPvalue*; while the function *qvalue* computes the q-value to measure the significance of each feature in terms of false discovery rate rather than false positive rate [114]. The unweighted network displays sensitivity to the choice of the correlation values cut-off, thus, it is important to use a proper criterion to select the edges to include in the network. It is important to take into account the correlations are computed among each pairs of genes/proteins leading to a high rate of false positive values. Thus, to build an unweighted network and to reduce the inclusion of not significant correlations, it is recommended to set a cut-off also for p and q values. Concerning the weighted networks, the choice of the β parameter is based on the scale-free topology criterion [115]. This method represents an improvement over unweighted networks based on dichotomizing the correlation matrix; the continuous nature of the gene co-expression information is preserved, and the results of weighted network analyses are highly robust with respect to the choice of the

parameter β (soft-thresholding power). However, this thresholding method is based on the assumption that the network follows a scale-free topology, a hypothesis weak in some cases, as discussed in Paragraph 2.5.1.

WGCNA package also provides a analysis procedure for the obtained network, consisting of a hierarchical clustering algorithm based on a distance matrix calculated by similarity measure between gene/protein pairs [111]. According which each node is assigned to modules, and then an aggregate module signature, called eigenvector, is computed; it can be considered as an object representing the expression profiles of the molecules belonging to the module, thus, it simplifies the comparison of different modules [116].

2.4.2 WGCNA and proteomic issues

When the WGCNA is applied to proteomic or to metabolomic data, the choice of the optimal cutting parameters should be evaluated in relation to the nature of the data analyzed. In fact, due to the low coverage of the current analytical technologies, the produced dataset are often incomplete, and the methods need to be properly modified [117]. A major concern is the high rate of missing values that introduce loss of information and significant bias. To address this issue, several approaches including K nearest neighbor, least square methods, or local least square methods have been proposed for proteomic and metabolomic datasets too [117]. In other cases, a very simple approach has been adopted, such as the removal of all species with a number of missing data bigger than a given threshold [118]. However, to implement a more accurate analysis, it is recommended to process data by using an imputation method taking into account the nature of missing data. Three types of missing value have been identified: MCAR (missing completely at random), i.e., due to stochastic fluctuations in a proteomic dataset, MAR (missing at a random), i.e., due to multiple minor errors, and MNAR (missing not at a random), i.e., due to limits of abundance of peptides/proteins that instruments are able to detect. In general, methods work fine when a low percentage of missing value ($\leq 10\%$) is present, but this threshold could be different in relation to the missingness mechanisms and imputation approach used [117, 119]. In addition to missing value, another important step of proteomic data preprocessing concerns their normalization [120]. Batch effects may occur in datasets run in different days or by different technicians. This phenomenon may increase by using isotope reagents which allow the quantitation of a limited number of samples, thus, preventing a simultaneous analysis of multiple samples which could reduce data heterogeneity. For these reasons, an appropriate data transformation is a prerequisite to capture

true correlations. Also in the case of protein co-expression, valid correlations have to be selected by applying proper thresholds. To date, the most applications of WGCNA method on proteomic datasets used the soft-thresholding, which defines the β value according the scale-free criterion [40, 112, 118]. However, since the application of WGCNA to proteomic dataset is a recent issue and literature reports few examples, the future evaluation of hard-thresholding approach might be useful.

2.5 Network topological analysis

The literature underlines how the topological structure of a network model is closely related with biological relevant features of real-world systems. To fully understand disease mechanisms, the knowledge of both the *topological* and the biological aspects of these models is needed [75]. In this dissertation, the term *topology* is used to refer to all the properties and characteristics of a network, related to the structure of the graph, i.e. related to the way in which nodes and edges are arranged to give rise to the network. An important aspect of topological studies is the construction of mathematical/theoretical models in order to describe features observed in the networks. The capacity to abstract from experimental data to mathematical models could provide a better understanding of biological systems. In this Paragraph, the main approaches found in literature on topology analysis will be presented. The first part will describe the main metrics used to investigate which are the key proteins in a network. The second part will discuss the main mathematical models affirmed in describing the topology of biological interaction networks. The discussed models, namely random graphs, small-world, scale-free and geometric random graphs, will not be described in detail as it is not the objective of this dissertation to illustrate their mathematical basis. Rather, the aim of this paragraph is to introduce the reader to the main state-of-the-art techniques used to investigate and explore the topology of these network models. In order to show that the topological properties are directly linked to specific biological meanings, thus their investigation represents a way to infer useful knowledge about the biological phenomenon represented by the network.

2.5.1 Centrality measures

Table 2.2 lists the main basic centralities used in the network topological analysis [121]. In the context of network organization, these centralities facilitate the

answer to question about which proteins are most important and why. For example, to give an idea of such analysis, a vertex (i.e., a protein) is important (or central) if it is close to many other vertexes. There are many different centrality measures that have been proposed in literature but probably the most applied, and simple, is called vertex degree. The degree $d(v)$ of a vertex v, in a network $G = (V, E)$, counts the number of edges in E incident upon v. Given G and a constant value D, $f(D)$ is defined as the fraction of vertexes $v \in V$ with degree $d(v) = D$. For different $D_1, D_2, ..., D_n$, the collection $f(D_1), f(D_2), ..., f(D_n)$ is called the degree distribution of G. A useful generalization of degree is the notion of vertex strength, which is obtained simply by summing up the weights of edges incident to a given vertex. The distribution of vertex strength is sometimes called the weighted degree distributions defined in analogy to the ordinary degree distribution. Another centrality measure widely used is known as betweenness [122]. It can be defined as follows: this measure summarizes the extent to which a vertex is located "between" other pairs of vertexes. In this case, centrality is based upon the perspective that importance relates to where a vertex is located with respect to the paths in the network graph. In other terms, betweenness centrality is based on communication flow. Nodes with a high betweenness centrality are interesting because they lie on cellular communication paths and control information flow. Also called hubs/bottlenecks [123], they can represent important proteins in signaling pathways and can form targets for drug discovery. Formally, betweenness can be defined as

$$B(v) = \frac{\sigma(s, t|v)}{\sum_{s \neq t \neq v \in V} \sigma(s, t)} \tag{2.4}$$

where $\sigma(s, t|v)$ is the total number of shortest paths between s and t that pass through v, and $\sigma(s, t)$ is the total number of shortest paths between s and t (regardless of whether or not they pass through v). Other centralities used to globally evaluate the structure of a network include:

- Degree distribution: a function describing the proportion of nodes related to each observed degree
- Modularity: evaluates the presence of modules, such as a group of nodes characterized by the tendency to form more connections within the group than outside [124]
- Cluster coefficient: the ratio of the number of edges among a node and its neighbors and the maximum possible number of edges among all of them [125]

- Motif/graphlet frequency: evaluates the presence of small subgraphs with a specific pattern that appear in a real-world network more frequently than in the relative random network [126]
- Edge clustering coefficient: the ratio between the number of triangles (three nodes connected by three edges) including an edge, and the maximum number of possible triangles may include the edge [127]
- Maximal Clique centrality: a property of a node taking into account the cliques (i.e, a subgraph in which each pair of nodes is connected) including the node [128]

The simplest way to perform a network topological analysis by evaluating these properties is through Cytoscape's plugins, such as CentiScaPe [121] and NetworkAnalyzer [129], that provide the main basic methods to compute the topological properties of nodes, edges, and networks, both directed and undirected. Moreover, new plugins implementing recent developed topological centralities are CytoNCA [130] and CytoHubba [128].

2.5.2 Theoretical mathematical models

As regards biological networks many studies have been carried out to compare experimental networks with mathematical models with the aim to found the best fitting model for using it to evaluate experimental results [75]. As for theoretical mathematical models proposed to describe the biological networks, the most claimed are Erdös–Rényi random graphs [34] and scale-free [35] (see Fig. 2.6).

Other models, such as the geometric random graph (GEO) [132] and the small-world [125], have recently been proposed. In the context of biology, the random graph, proposed in 1950, has been overtaken by the scale-free model; in fact, the degree distribution of the scale-free model is a power-law curve that fits better than Poisson curve (typical of random graphs and small-world) the degree distribution of the experimental networks [35] (Fig. 2.7).

In networks with a power law distribution, most nodes have a degree value far from the mean value; specifically, most nodes have a low number of interactions while few nodes have a high number of interactions. These features lead to a network structure little vulnerable respect to the random removal of a node and make the related system biologically robust [133]. Of note, the degree distribution may reflect the different role of proteins/genes, and those with a highest number of connections, so-called hubs, have a higher probability to be more biologically relevant than others. In other words, removal or modification of hubs

TABLE 2.2: CentiScaPe's centralities and their topological and biological meaning. The ~indicates network's properties. The $*$ indicates node's properties. The $*]$ indicates edge's property.

Centrality	Description	Biological Meaning
Diameter ~	Defines the longest shortest path in the network	
Average Distance ~	Defines the mean length of all the shortest paths in the network	
Degree $*$	Describes the number of neighbors a node has	Highlights the number of nodes that are regulated/regulate the node v
Eccentricity $*$	Describes the longest shortest paths a node develop, giving us a proximity information	Highlights the easiness of a protein to reach/to be reached by all the other proteins in the network
Closeness $*$	Describes, for the node v, the minimal sum of all the distances in the network	Highlights the probability of a protein to be functionally relevant for several proteins, but irrelevant for a few other
Radiality $*$	Describes the integration of a node into the network	Highlights the ability of a protein to be functionally relevant for several proteins, but irrelevant for a few other
Centroid $*$	Describes the neighborhood of nodes by highlighting nodes that have the highest number of neighbors separated by the minimal shortest path	Highlights a protein that tends to be functionally capable of organizing discrete protein clusters or modules
Stress $*$	Describes the number of shortest paths that pass through a node	Highlights the relevance of a protein as functionally capable of holding together communicating nodes
Betweenness $*$	Describes, for each couple of nodes, the number of shortest paths that pass through a specific node	Highlights the relevance of a protein as functionally capable of holding together communicating nodes
Bridging $*$	Describes the neighborhood of nodes by highlighting nodes with a high number of high-degree neighbors	Highlights a protein possibly bringing in communication sets of regulatory protein
Eigen Vector $*$	Describes a sort of weighted degree, where not only the number of the neighbors is important but also the Eigen-Vector of the neighbors itself	Highlights a protein interacting with several important proteins, suggesting a central super-regulatory role or a critical target of a regulatory pathways
Edge Betweenness $**$	Describes, for each couple of nodes, the shortest paths that pass through a specific edge	Highlights the relevance of the interaction as capable of organize regulatory process

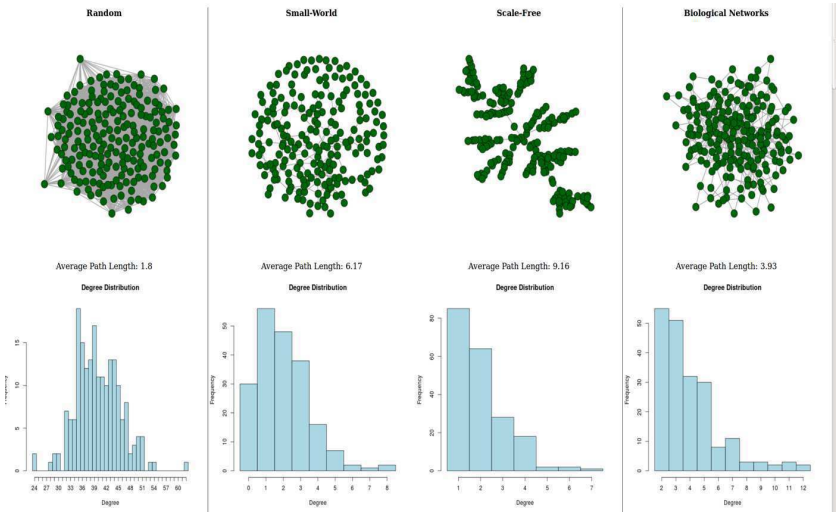


FIGURE 2.6: Shape and degree distribution of random, small-world, and scale-free model with respect to a biological network. Models were calculated by ELIXIR web tool [131]

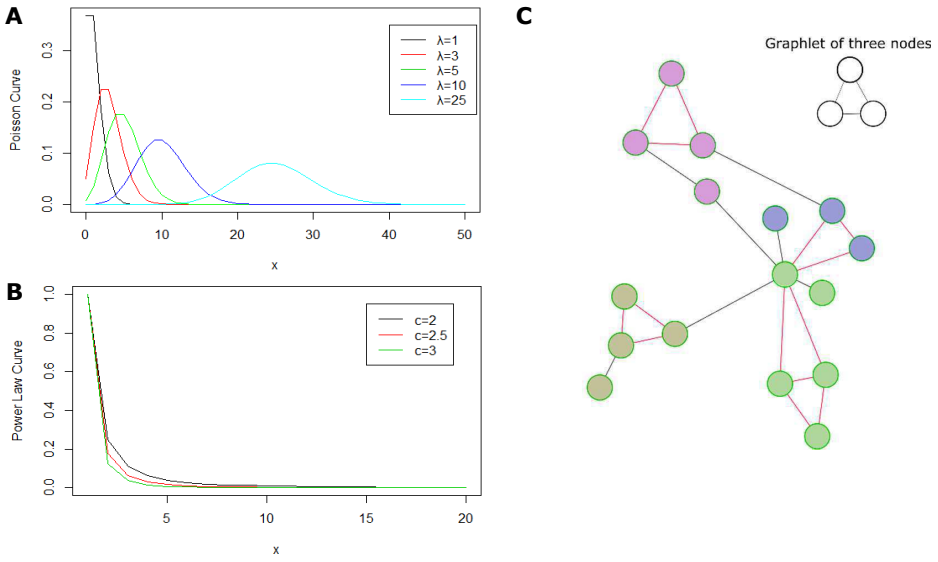


FIGURE 2.7: Functions used to describe the degree distribution of biological networks. Poisson curve a and power-law b shown for different parameters. c Example of graphlet of three nodes with frequency equal to 5

may induce stronger alteration of the system equilibrium rather than removal or modification of nodes with low degree [134]. Although some topological properties are well described by a theoretical model, it may not be enough to affirm that the model represents well the real-world network considered [135]. For example, a study on PPI network of Drosophila Melanogaster and Saccharomyces Cerevisiae showed that the degree distribution was in agreement with scale-free model, but diameter, cluster coefficient, and graphlet frequency were closer to GEO [136]. Of note, based on graphlet frequency, the comparison among scale-free, random graph, and GEO models has shown a higher agreement of GEO with PPI network from eukaryotic organism [137, 138]. A possible reason of these findings is that the scale-free model fits networks that emerged from a stochastic growth, not subjected to an optimization process; while, PPI networks emerge from stochastic processes, and their structure is influenced by the evolutionary optimization that living systems have gone through [136]. Another model used to describe the PPI networks is the small-world, characterized by low average path length, an high clustering coefficient, and (like the random graph model) by a Poisson curve. In a study focused on the investigation of proteins regulating the fat storage, the corresponding PPI network had a degree-distribution close to a Poisson curve rather than a power-law [139]. Moreover, the investigation of its topology indicate a network organized into communities, i.e. with a modular structure. This modular arrangement of the nodes has been often observed in PPI networks [140–142]. Modules are associated to groups of proteins that work together to achieve some specific biological functions. The presence of these communities allows to split a complex process, as the fat storage function, in sub-tasks each performed by a specific cluster of nodes. The small-world model preserves the modular structure, and it is not characterized by hub nodes, these properties make the network more robust in the case of removal or modification of any node [133].

The proposed models were born from the need to describe the various specific features observed in real-world networks and are thus not able to summarize the complexity of the whole system investigated. Moreover, certain model properties do not agree with experimental data. The real-world networks used to infer these general theoretical models are not standard and often represent only an incomplete part of the interactions of the studied system. Another important aspect to consider is the error introduced by technologies used for detecting interactions; for example the observed structural modularity may be an artifact because the current PPI data include interactions detected through technological approaches that create modules [142]. Moreover, several edges/interactions may occur in the experimental procedure but not in vivo,

due to the fact that protein pairs may not be expressed under different time-space conditions [137]. This evidence leads to the incapacity to completely prove or rule out a model. The actual high-throughput technologies for data acquisition are far from providing complete and reliable data regarding bio-molecule interactions [75]; however, as more data becomes available, more insights could be gained and as consequence, the definition and evaluation of mathematical models can improve.

2.6 Module identification

Studying the modularity of a network not only provides structural information on the network, but may also reveal the underlying mechanisms that determine the network structure. Regardless of the approaches used in obtaining a biological network, the detection of protein/gene modules is of great interest, because they represent the functional units at the base of the mechanisms responsible for cellular life [143, 144].

A PPI module represents a group of proteins taking part in specific, separable functions such as protein complexes, metabolic pathways or signal transduction systems. A module is identified on the basis of its double role (i) as an isolated entity, being responsible of specific steps of cellular processes; and (ii) as part of a connection pattern, in which one process influences another in order to perform higher-level cellular functions [5]. For example, the Generic Transcription pathway (R-HSA-212436) [100] achieves its functions through its sub-processes, such as the nuclear Receptor Transcription pathway, the NotchHLH Transcription pathway, etc. (see Figure 2.8). Moreover, each sub-process can be described as a module made of proteins and other molecules working together to perform a specific step of a bigger pattern. In network biology and graph theory, it is possible to define *topological* and *functional* modules [145]. The first refers to a group of nodes having many more connections with the nodes of the group rather than with the ones outside of it. The second refers to a group of nodes sharing a common biological function. Note that a group of nodes potentially representing a module might possess *both* topological and functional properties. Ideally, topological and functional modules should coincide; in practice, they constitute two different entities, though typically, they largely overlap [17]. As a consequence, both the network topology and the functional information contribute to the overall understanding of the biological mechanisms underlying the PPI network.

Topological properties are measured with specific metrics such as modularity, betweenness, degree distribution, density and closeness [36]. On the other

hand, functional properties are widely described by the three Gene Ontology categories (GO), Biological Process, Molecular Function and Cellular Component [54].

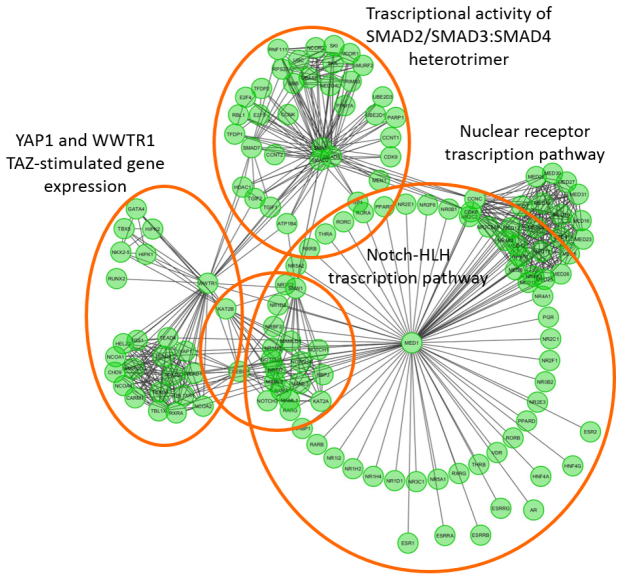


FIGURE 2.8: The figure represents the processes at the base of the Generic Transcription pathway (R-HSA-212436). Each process consists of a group of proteins with intra-modular and inter-modular connections. The image has been obtained with ReactomeFVIZ software [100].

Due to the complex connectivity of the networks, the identification of modules is a challenging task. Various methods have been proposed, and most of them are exclusively based on network topology (see Paragraph 2.5.1 for a description of the metrics used to measure topological properties). Some representative examples include the betweenness based method [64], the modularity optimization method [146], the spectral partitioning method [147], the core-attachment based method [148], and the graph-theoretic approach relying on cliques [149] and other topological properties [150]. A number of scientific reviews describe many of these tools for module detection [118, 151, 152]. Among the most used there are ClusterOne [153], MCODE [150], Markov Cluster Algorithm (MCL) [154] and CFinder [149].

ClusterOne consists of three major steps. Firstly, starting from a single seed, i.e. a vertex, some vertexes are added or removed so as to find groups on the basis of a new introduced measure, namely cohesiveness. This measure takes

two aspects into account: the module should contain many interactions, and it should be well-separated from the rest of the network. This process is repeated from different seeds, thus forming multiple and overlapping groups. Overlapping means that the groups share one or more nodes. In the second step, groups with an overlapping score above a specified threshold are merged. In the third step, the groups that contain less than three proteins, or whose density is below a given threshold, are discarded [153].

The MCODE algorithm consists of three phases. Firstly, a score is assigned to each vertex; the score is the cluster coefficient relative to the neighbourhood (directly connected vertexes) of the vertex. Secondly, protein modules are constructed adding vertexes to those with the highest score. The growth process of a module is limited by a parameter. Moreover, another parameter executes a control on the scores of the vertexes within the same module. Thirdly, the modules found are post-processed in two ways. Firstly, by removing the vertexes that are connected by only a single edge to the rest of the module and secondly, by trying to expand the module with other vertexes, if they are connected to a large number of vertexes of the same module [150].

At the heart of the MCL algorithm lies the idea to simulate flow within a graph; MCL promotes flow where the current is strong and demotes flow where the current is weak, in order to detect natural groups present in the graph. The MCL algorithm has a single parameter known as inflation. Larger inflation values result in smaller clusters, while smaller inflation values generate only a few large clusters [154].

CFinder was one of the first overlapping clustering methods published in the literature. CFinder finds all k-cliques of the original network, i.e. the cliques made of k nodes, (where k is a parameter of the algorithm). These cliques are used to construct a graph, where the nodes are represented by the k-cliques and the edges connect k-cliques which share k-1 vertexes. The connected components of this graph are then used to derive the overlapping topological modules [149].

Among the most popular algorithms for module detection, it is important to mention those based on *Modularity* optimization, such as fast-greedy, walktrap, label propagation, spinglass and multi-level community [155]. *Modularity* is a metric specific for the evaluation of the modular structure of a network [124]; as discussed in Paragraph 2.5.2, this structure is typical of PPI Networks.

Although many methods are available in detecting topological modules, none of them provide a biological interpretation of the protein groups found. For this reason, other approaches are used in order to annotate the identified modules with known cellular functions. For example, GO term enrichment analysis is

routinely used to assess the biological relevance of the identified modules [51, 52]. Making use of statistical tests, these approaches evaluate if genes/proteins of a module are enriched in common functional properties represented by GO terms (see Paragraph 2.2 for a description of the GO).

2.6.1 Issues linked with module identification algorithms

Several graph-based algorithms have been developed to tackle PPI module identification. Most of these approaches infer modules relying solely on their topological properties. These methods exploit community detection algorithms developed for generic graphs, readjusting them to the context of biological networks [155, 156]. Module identification algorithms are also most commonly known as clustering algorithms. Nevertheless, clustering algorithm is a general term including also other types of algorithms, which ignore the graph structure typical of input models specific to module identification algorithms. While the topological approach is sound in network theory, it is sub-optimal in the case of PPI networks, due to their biological nature.

The scarce sensitivity of PPI discovery techniques (such as yeast two hybrid method and tandem affinity purification coupled with mass-spectrometry) leads to the presence of noise, in the form of falsely detected edges [85]. As a consequence, the modules obtained with algorithms based solely on network topology are strongly influenced by the presence of noise. Algorithms should consider this noise during module detection; for example, further information may be introduced to model the uncertainty associated with interactions [145, 157]. To overcome the issues of noisy edges, several recent algorithms pre-process the network with a-priori knowledge, such as co-expression relations and/or functional associations. In practice, they filter out the low reliability edges, and/or enrich the network with edge weights [62, 63, 144, 158, 159]. Despite the integration of a-priori information, module identification in these algorithms is strictly topological.

Since PPI network analysis aims to clarify the molecular mechanisms involved in the physiological/pathological context investigated, a main requirement of module identification algorithms is the ability to provide *full coverage* of the network. Indeed, to obtain a bird's eye-view of the phenomenon, all the elements of the system should be considered, given that high-level functions emerge from the combined work of many modules/processes [13, 145]. Topological approaches are often focused on the identification of high quality clusters, thus not considering many of the network hard-to-label nodes. Moreover, module identification algorithms focus mainly on the detection of densely

connected subgraphs, ignoring functional modules that are often sparsely connected [145, 160], and/or very small, i.e. composed of only two or three proteins [59]. Ignoring these modules means excluding key proteins influencing/driving the inspected biological process. Alongside these aspects, there are two relevant requirements which these algorithms should assure: overlapping [157] and the inclusion of known protein annotations [145]. As a protein can have several different roles in cellular functioning, it can therefore be involved in different modules; for this reason, a key characteristic of these algorithms is the ability to provide overlapping modules i.e. modules sharing proteins. Although some strategies have been proposed over recent years to assure overlapping, this problem is still a challenging issue [157].

In recent years, thanks to knowledge obtained through omics technologies, biological literature on the roles of proteins has been growing steadily. In detail, proteins are annotated with attributes (such as GO) to encode information such as functions, localization, and biological processes that they are involved in. The wealth provided by this knowledge opens the door to new opportunities of creating superior quality modules, although the inclusion of this information is challenging, due to the high dimensionality of protein annotations [145].

2.6.2 Module identification: the inclusion of GO annotations

To ensure that the identified modules are biologically meaningful, these algorithms should take into account not only topological features but also known functional information [144]. For this purpose, several algorithms have been proposed attempting to integrate this information into the network model. For example, SWEMODE is an approach in identifying dense subgraphs using network measures which combine functional information with topological properties. Each edge is weighted taking into account two measures: clustering coefficient and functional similarity, obtained according to GO annotations. Finally, SWEMODE creates modules starting from certain nodes and adding more in a similar way as MCODE does [158]. Another approach has been proposed by *Wang et al.* [161], which uses GO annotations to estimate the reliability of interactions in PPI networks. Interaction pairs with low GO similarity are removed from the network, as they are considered unreliable, and a clustering algorithm is used to detect modules. Moreover, a similar approach is followed by *Zhang et al.* [162], where the authors apply a clustering algorithm to an augmented network, constructed to integrate GO annotations in the PPI Network.

All these approaches integrate GO information into the network model, later applying a topological-based approach to detect modules; in this way, GO is not

included in the process of module assembling but is rather used in a preliminary phase of network model pre-processing.

Recently, several novel methods have been proposed, they try to include GO annotations directly in the process of module identification, such as DCAFP and GMFTP. In the DCAFP method, a preference vector is introduced for each protein to indicate the functional categories (GO) to which it belongs. Then, the problem is formulated as a constrained optimization problem based on a likelihood matrix which considers graph topology as well as the preference vectors of proteins [61]. In the GMFTP method, the functional profile (given by GO annotations) is coupled with the network topology to detect overlapping protein complexes. The process of module assembling is dominated by two variables: one represents the degree of a protein related to a considered module, while the other represents the protein functional profile. The module detection problem is later converted into a parameter estimation problem [60].

Although new approaches have been proposed, the module detection problem remains a challenging task. The possibility to include GO knowledge has at present been little explored. Moreover, all these methods do not provide any biological interpretation concerning the modules found. For this reason, further analysis is needed in order to investigate the role of modules found in the biological mechanism represented by the PPI network.

2.6.3 MTGO: a novel algorithm for module detection

The main research scope of this PhD work has been the development of a new algorithm for module detection. This has been done in order to overcome the issues linked with the biological nature of the problem and to supply the lack of methods able to provide a PPI Network evaluation, both from a topological and biological perspective. The algorithm proposed is MTGO (Module detection via Topological information and Gene Ontology knowledge). It combines information from network topology and knowledge on the biological role of proteins. In order to identify relevant modules, MTGO executes iterative steps in order to obtain repeated network partitions. A partition is a subdivision of the graph into sub-graphs, covering all the network nodes. For each step, starting from a network partition, moving nodes among sub-graphs, a new partition is computed. In this way, at each step the modules are reshaped on the basis of both the GO annotations and the graph Modularity. Therefore, the module detection problem is faced through a process of optimization, taking into account the network structure as well as its biological nature.

Moreover, MTGO differs from previous works as it integrates GO terms directly into the construction of functional modules. As a result, it provides both a set of clusters/communities (topological modules) and a list of functional modules, represented by GO terms. In other words, MTGO facilitates and simplifies PPI Network analysis, coupling both the clustering analysis and the biological/GO analysis in a single step (as it provides both the clusters and their biological meaning, through the set of GO terms describing the clusters). In this way, MTGO simultaneously obtains both full coverage of the network and overlapping functional modules, the ideal characteristics of functional module identification algorithms. Furthermore, MTGO is tailored to search for small or sparse modules, which typically elude other approaches.

As opposed to previous approaches based on GO, such as DCAFP [61] and GMFTP [60], MTGO provides a unique GO term that best describes the biological nature of each identified module. This supports a better explanation of the results obtained, highlighting the main processes involved in the biological system represented by PPI network models. Due to its unique way of GO exploitation (directly when modules are assembled), MTGO differs from state-of-the-art algorithms, where GO does not directly guide module assembling.

Chapter 3

Protein Co-expression network analysis

3.1 Introduction

Accumulating findings show that the phenotypic impact of a disease is not a consequence of the abnormal action of a single gene, but reflects the perturbations on those molecules that are involved and interconnected with the damaged protein or gene [17]. In agreement with this concept an approach to study diseases consists in investigating the different way in which bio-molecules are connected among them in the two different conditions: healthy people and affected people [44, 112]. This is possible through the study of the rewiring of the condition-specific protein interaction networks related to the studied pathology. In order to implement this strategy the network models related to each condition should be built. This means that the information about the bio-molecules acting in a specific context should be used to directly infer their relation-ships, that are the edges of the network model.

Given a list of proteins of interest, a common way to build a protein interaction network consists in retrieving information about their interactions from public repositories [43, 52]. In fact, there are many databases (MINT, DIP, STRING, IntAct, HPRD) collecting both physical and functional protein interactions, obtained from experiments and computational methods of prediction (see Chapter 2 Paragraph 2.3.2 for further details). Since the proteins participate to physical-chemical connections depending by the biological context where they act, the interactions collected in these databases are peculiar of the system conditions during the experiments. As consequence, the use of these interactions to describe another system, likely characterized by a very different biological context, could limit the analysis and lead to misunderstanding of the phenomenon studied.

To overcome these issues, some original works have used graph models rising from data obtained combining the mass-spectrometry and liquid chromatography (LC-MS) [40, 41, 44]. These techniques allow to investigate all the proteins in a given cell/tissue/sample in qualitative and quantitative terms and provide a big amount of proteomics data, providing a new range of opportunities to improve existing models and to build pathology-specific networks. Moreover, as PPIs often do not cover all the interactions of an organism, the use of proteomics data from LC-MS gives the opportunity to evaluate those organisms that lack information on PPIs. The set of proteins and related information extracted form a sample of a specific phenotype is commonly called *protein profile*. The big quantity of protein expression data produced by these technologies is a source to build network model for investigating *disease signature*. Where, *disease signature* is an expression commonly used, which stands for a set of genes/proteins whose levels of expression can be used to predict a biological state (for example, in the case of cancer, gene signatures have been developed both to distinguish cancerous from non-cancerous conditions) [163]. Currently, robust methods for network construction starting from protein expression data is still unrealized and the most used network are built from PPI database [43].

These proteomics datasets show a similar structure as gene expression data, usually coming from Microarray technologies. The big diffusion of tools and statistics for analyzing genomics data and the similarity of these outputs lead to various attempts to adapt these procedures for proteomics data. This is the case of the protein co-expression network analysis, rising from the application of gene-expression network analysis to proteomics data. To date there are few examples of application, mainly because there are some issues to adapt this approach to the protein field, in fact some data properties are specifically related to the analytic technology used to generate data (Chapter 2, Pararaph 2.4 presents a discussion about this topic). In this chapter, firstly, some state-of-art applications of protein co-expression network analysis for disease studying will be discussed; secondly, an implementation of the protein co-expression network approach to evaluate large-scale proteomics data coming from a study on amyloidosis disease will be showed. The Figure 3.1 sums-up the main steps of the analysis pipeline proposed.

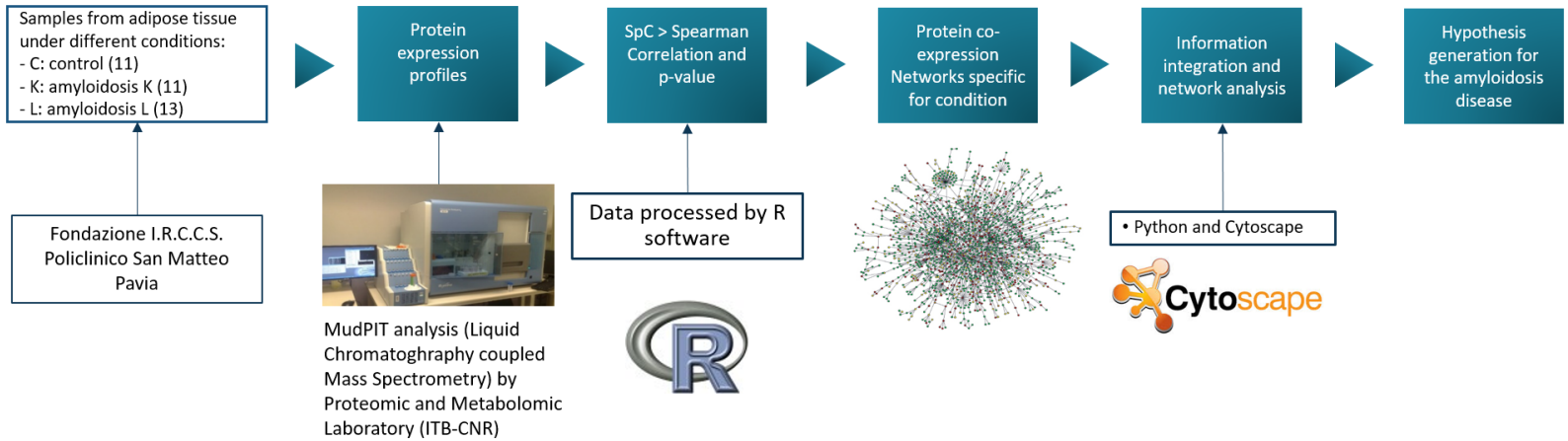


FIGURE 3.1: Building pipeline of condition-specific protein co-expression networks

3.2 Form protein expression profile to protein co-expression networks

The combination of PPI networks with protein expression profiles is demonstrating useful for rapidly identify biological pathways and sub-networks affected by disease mechanisms [164]. An interesting approach came from a study by Brambilla *et al.* [52], where the information on protein profiles is integrated in a network model. They used Multidimensional Protein Identification Technology (MudPIT) to obtain the protein expression profiles from two different groups of subjects, amyloidosis patients and relative controls. The expression profiles have been used for the identification of differentially represented proteins, i.e. the protein expressed with different abundance in two samples under different conditions. Then, these results have been combined with protein interactions retrieved from major public online repositories. In this way, they built a network where the edges represent PPI and nodes correspond to the differential proteins. The integration of different types of knowledge from expression data and literature data in a same model lead to new findings, which provide both a relevant contribute for clarification of the mechanisms of human amyloidosis and a basis for further experimental/clinical studies [52].

The protein profiles has been used in an original way by Wu *et al.* [43], for understanding renal dysfunction mechanism and finding phenotype-related biomarkers. At first, they followed a standard approach. The differentially expressed proteins, among different status of kidney transplant patients, have been used to build two networks, retrieving PPI from STRING database [92]. Then, they have introduced a new idea. In fact, the information on protein profiles has been exploited to validate the found PPI networks. They compute Pearson Correlation (PC) between each pair of PPI, reserving in the networks just those interactions with a high value of correlation, assuming those with a low PC value would not occur in corresponding sample. Finally, they extract from the two networks the differential interactions, i.e the subgraph made of the edges exclusive of each network, i.e. the uncommon edges between the two networks. From this subgraph they selected twelve relevant proteins involved in the pathology. This strategy links PPI networks with protein profiles, permitting to obtain condition-specific models related to studied phenotypes. The use of protein expression profiles to study differential interactions rather than differential proteins represents a new approach respect to conventional methods.

The investigation of quantitative proteomic data by WGCNA approach (this technique is discussed in Chapter 2 Paragraph 2.4) has been first addressed by Gibbs *et al.* to infer the protein abundance and to overcome issues linked to

peptide-protein mapping [41]. Starting from experimental datasets obtained by LC-MS, the authors proposed a protein co-expression network approach (ProCoNa) where the nodes are peptides (segments of proteins) and the edges are calculated by processing peptide intensity, a measure indicative of their abundance in analyzed samples. The modules computed by co-expression analysis were strictly correlated with the investigated phenotypes and showed a significant enrichment of some GO terms. Following these findings, the authors explored the relationship between co-expression networks reconstructed from transcriptomic and proteomic data [40]. In this study, concerning SARS-CoV infection, they used a graph analysis to evaluate phenotype associations and module correlation between protein and transcript networks. In this way, providing a foundation of a true multi-omics disease signatures. The idea to use the WGCNA method on proteomic data was followed also by MacDonald *et al.* [44] to clarify the role of the glutamate signaling in schizophrenia (SCZ). The topological evaluation of the co-expression networks from SCZ affected subjects and healthy controls led to observe in SCZ affected group a lower average node degree (see Chapter 2 Pragraph 2.5.1 for a node degree explanation). This result was probably due to the loss of coordination of the biological functions, as well as disease heterogeneity. However, in SCZ network, it was found the exclusive presence of a module enriched in GO terms related to glutamate signaling pathway and whose proteins had a significant increased degree. This analysis has brought to the conclusion that this module is linked with spine loss symptom in schizophrenia patients. The application of the WGCNA on protein expression profiles was also faced by Chang Guo *et al.* to characterize the role of different protein isoforms in E. Coli resistance to serum killing [112]. Like in other cases, the authors evaluated the topological variations of the co-expression networks between control- and serum-treated groups. By considering the connectivity of modules identified in both networks, a protein, IleS, was found with a differential number of connections in control and treated groups. Of note, its involvement in the response to serum killing was confirmed by independent functional test based on a gene-deletion mutant, thus, confirming the utility to use protein co-expression networks also to identify putative drug targets. Likewise, Yu *et al.* investigated the molecular mechanisms underlying the glioblastoma multiforme (GBM) [46]. They analyzed samples of macaque rhesus brain by both iTRAQ (a technique for protein profiling) and RNA-seq approaches. The proteins identified were combined with STRING database and, for each experimentally validated PPI, the PC score was calculated using both protein and transcript levels. Since the PC score from proteomic data resulted significantly

higher than score calculated using transcript levels, the authors focused on protein co-expression network to identify protein modules involved in the disease. Finally, a more detailed evaluation of these modules allowed the selection of eight genes of interest, and two of them were already known drug targets of GBM.

These findings confirm that the topological differences of network models building on protein expression data related to condition-specific groups can be directly linked to phenotypic differences among the groups. Thus, this type of models represent a starting point to investigate protein alterations linked to tissue-specific abnormalities due to pathological states.

3.3 An analysis pipeline based on protein co-expression networks

Amyloidosis diseases include rare as well as common dysfunctions, such as Alzheimer's disease, and are overall responsible for about 1 of 1000 deaths in developed countries. The common clinical manifestations are protein misfolding and tissue deposition as interstitial amyloid fibrils; the disease could lead to severe dysfunction of vital organs and ultimately to the patient's death [165]. Much is still unknown about the molecular mechanism of cell/tissue reaction to misfolded proteins and amyloid fibrils leading to the devastating clinical manifestations [52, 166]. The use of proteomics analysis of amyloid affected tissues has recently gained attention. In particular, in the scientific work *Reliable typing of systemic amyloidosis through proteomic analysis of subcutaneous adipose tissue*, *Brambilla et al.* proposed a method based on two-dimensional liquid chromatography coupled to tandem mass spectrometry (2DC-MS/MS), also referred to as MudPIT, for high-throughput proteome analysis of fresh abdominal fat tissue obtained from patients with various types of amyloidosis [167].

To develop a pipeline of analysis specific for amyloidosis disease based on protein co-expression networks, the starting point was the analysis of a sample set of adipose tissue from amyloidosis patients. The analysis procedure is the same of that proposed by *Brambilla et al.* [167] and was performed by the Proteomic and Metabolomic Laboratory of the Institute of Biomedical Technology of the Italian Research Council (ITB-CNR). These data have been published in *Journal of Proteome Research* by *Brambilla et al.* [52].

The sample set is made of three groups: the **control**, made of 11 healthy people, the group **K**, made of 11 patients affected by amyloidosis disease of type K and the group **L**, made of 13 patients affected by amyloidosis disease of type

L. The amyloidosis type, **k** or **L**, refers to which protein is the main constituent of the fibrils. All the samples are provided by *Fondazione I.R.C.C.S. Policlinico San Matteo Pavia*.

The output of the analysis is presented as a matrix in which the columns represent the samples, coming from the analyzed subjects, and the rows the proteins found in the corresponding sample. The matrix contains a total of 958 proteins, identified in the three groups. Each cell contains a Spectral Count (SpC) value, which measures the abundance of each protein (row) found in the corresponding sample (column). SpC corresponds to the total number of spectra taken from a given protein, provided by the MudPIT analysis. This value is linearly correlated with the protein abundance and allows to define the expression profile of a protein across different samples. Moreover, the SpC value is corrected by a correction factor that normalizes the spectral counts of a protein to its molecular weight [168].

The matrix presents the problem of missing data (discussed in Chapter 2 Paragraph 2.4.2), in fact the percentage of null SpC value was 74% in **control** group, 70% in **L** group and 81% in **k** group. To overcome this issue, those proteins with a number of missing data bigger than a given threshold has been removed. This simple approach has been previously adopted to process proteomics dataset [118]. The value of threshold used was 4. After this correction, the number of selected proteins were 153, 150 and 110 for **control** group, **L** group and **k** group, respectively. This matrix has been used to extract the relations among the proteins. The data has been processed using the function *CorAndPvalue* of the WGCNA package of R software. This function allows to compute the value of correlation between two proteins using the expression profile, i.e. the quantitative measures represented by the SpC values of a protein in a group of samples. In this analysis, the Spearman Correlation, a non-parametric measure of statistical dependence between two variables (see Chapter 2 Paragraph 2.4.1), has been computed using the SpC values. Moreover the *CorAndPvalue* function provides for each computed correlation a p-value to verify the significance. Moreover, in this analysis the function *p.adjust* has been used to apply the False Discovery Rate correction on the p-values computed, as this is the case of multiple comparisons. All the protein pairs showing a correlation value overtaking the threshold ± 0.8 and with a p-value less than 0.05 have been translated in a edge list for creating the network model. As result, three co-expression network models have been obtained, each one specific of a condition; the detail of node and edge number for each network is reported in Table 3.1.

Subsequently, the graphical aspect of the network models has been used to integrate other information, to enrich the model making it as informative as

TABLE 3.1: Network characteristics

	Nodes	Edges
Control	134	487
L	117	287
K	89	242

possible, in order to facilitate the comprehension of the biological context represented. In particular, the edge colour has been used to represent the sign of the correlation relation (green for positive sign and black for negative sign), while node colour and edge thickness have been used to integrate information from other analysis. Node colour represents differential expression analysis. This analysis allows to compare the expression of each protein both in control and patient group. In particular differential average (DAve) formula 3.1 has been used between the expression profiles of each protein in the two groups of comparison, **control** against **K** and **control** against **L**.

$$DAve = \frac{x - y}{2 * (x + y)} \tag{3.1}$$

In DAve function, x is the normalized SpC of the protein in the first sample, while y represents the normalized SpC of the same protein but in the second sample. In detail, DAve average $\geq$ + 0.2 correspond to proteins down-represented in patients, while DAve average $\leq$ - 0.2 correspond to proteins up-represented in patients [168]. The colour red represents proteins up-represented in **control** group, while light blue represents proteins up-represented in **K**/**L** group. To include a-priori information about the proteins involved in the analyzed biological system, PSICQUIC [169] has been used to retrieve information about known Protein-Protein Interactions (PPI) between the protein pairs linked by edges in the networks. PSICQUIC is a query interface for computational access to molecular-interaction data resources, collecting more than 16 million interactions from major molecular interaction databases. The PSICQUIC project site [170] offers open-source libraries for programmatic access to the PSICQUIC registry. Thus, a Python script has been implemented for registry interrogation. The interactions found have been represented in the network models in form of edge thickness; in particular, thick edges represent PPI linking two proteins, while thin edges represent just a sample-specific correlation relation between protein expression profiles. In this way, the integration of protein co-expression network with the commonly used PPI has been realized, to enrich the models and facilitate the comprehension of the biological mechanism represented. Moreover, the overlapping between the correlation relations and the

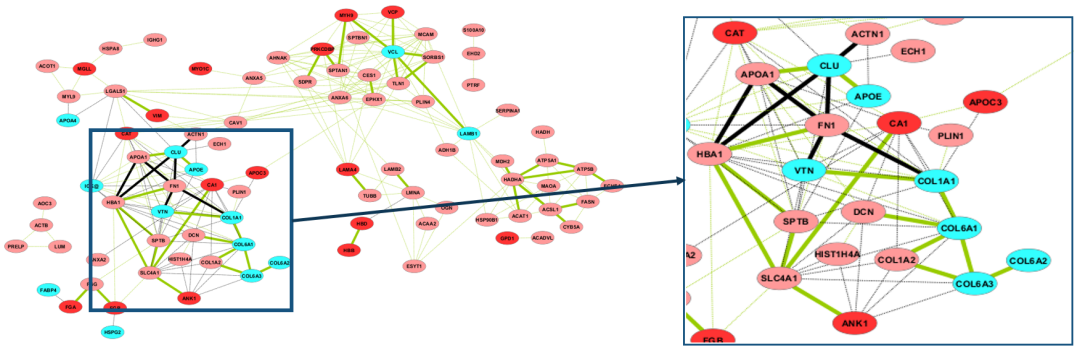


FIGURE 3.2: Node color represents differential analysis information: red for up-representation in **Control** group and light blue for up-representation in **K**/**L** group. Edge color represents the correlation sign: blac for negative correlation and green for positive correlation. Edge thickness represents presence in public repositories: thick means present in public repositories and thin means not present.

known interactions have been evaluated: the percentages of overlapping are $12\%, 22\%, 19\%$ respectively for **control**, **L** and **K** networks.

Following these steps four graphical network models have been obtained: one network model for amyloidosis **L** (Figure 3.3), one network model for amyloidosis **K** (Figure 3.4) and two networks for the **control** group. In these two networks from **control** group the node colour represents the differential analysis respect to **L** group (Figure 3.5) and respect to **K** group (Figure 3.6). A visual analysis of these networks led to identify some groups of highly interconnected proteins belonging to a same protein families (these groups are highlighted in Figures 3.3,3.4,3.6,3.5). The identification of these groups, such as Fibrinogen, Laminin, Collagen, suggests that the proposed approach is able to represent the bio-molecular mechanism acting at the basis of the investigated biological context. All the elaborations of information integration have been executed through Cytoscape software, an open source platform for visualizing molecular interaction networks [95]. The three networks, **control**, **L** and **K**, have been analyzed with CentiScape plug-in of Cytoscape [121], to identify the most topological relevant proteins. In details the three metrics, Betwenneess, Centroid and Stress, have been computed for all the network proteins (a description of these metrics can be found in Chapter 2 Paragraph 2.5.1). For each metric the mean value has been used as threshold to select the proteins more relevant, Table 3.2 shows the thresholds used for each network. For each network, just the proteins that exceeds the thresholds for all the three metrics has been selected. Finally,

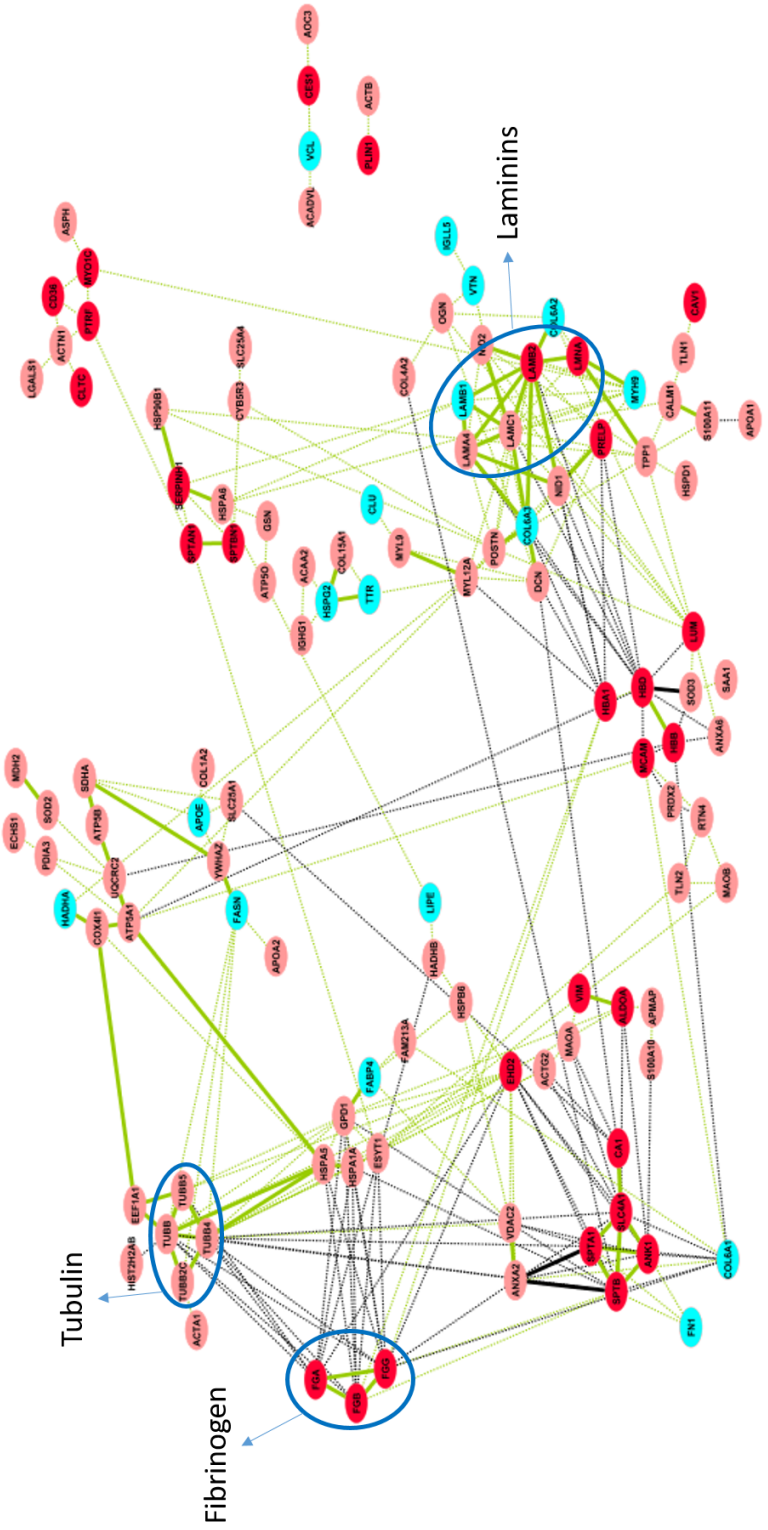


FIGURE 3.3: Network model for amyloidosis group L.

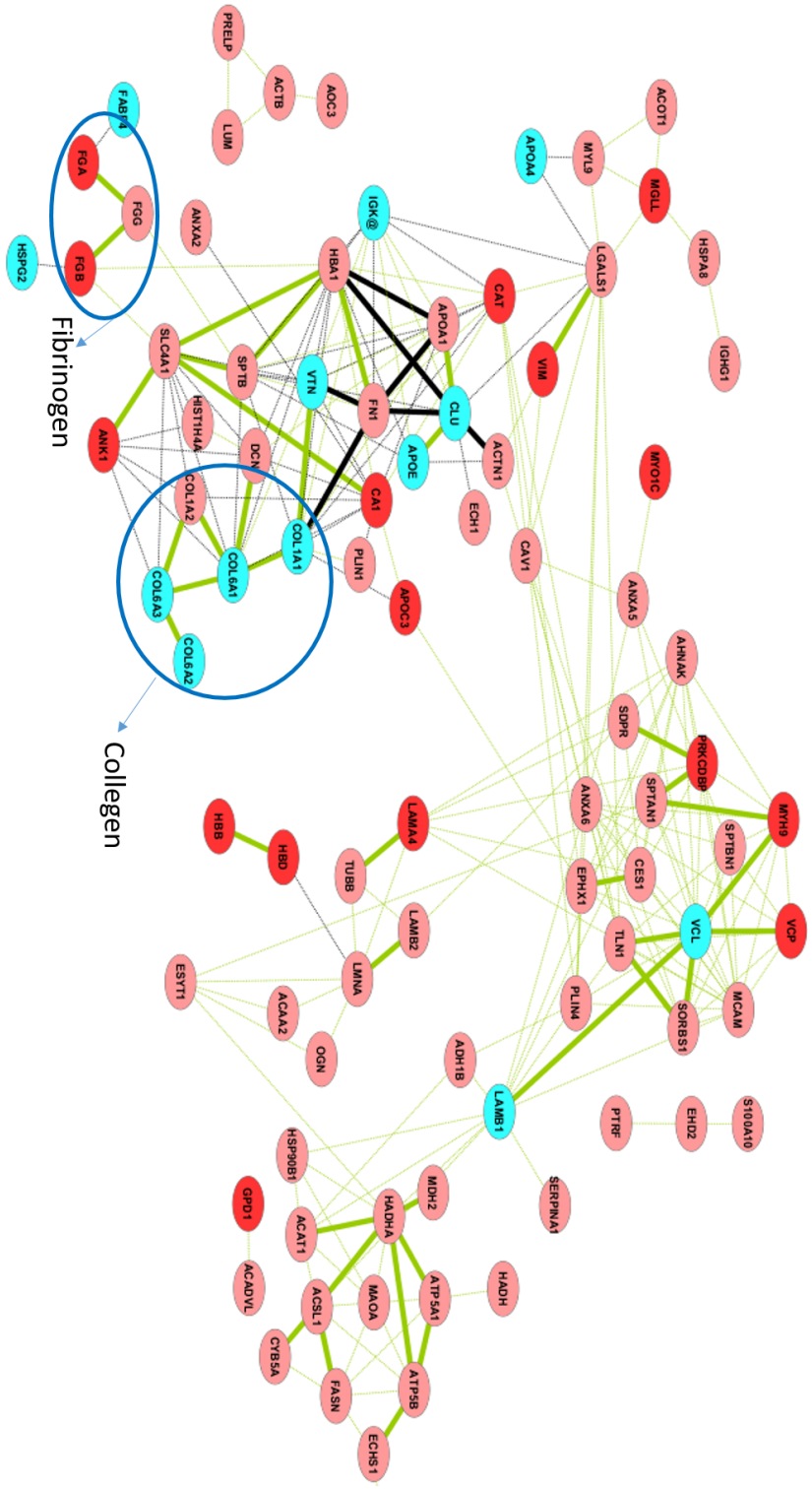


FIGURE 3.4: Network model for amyloidosis group **K**.

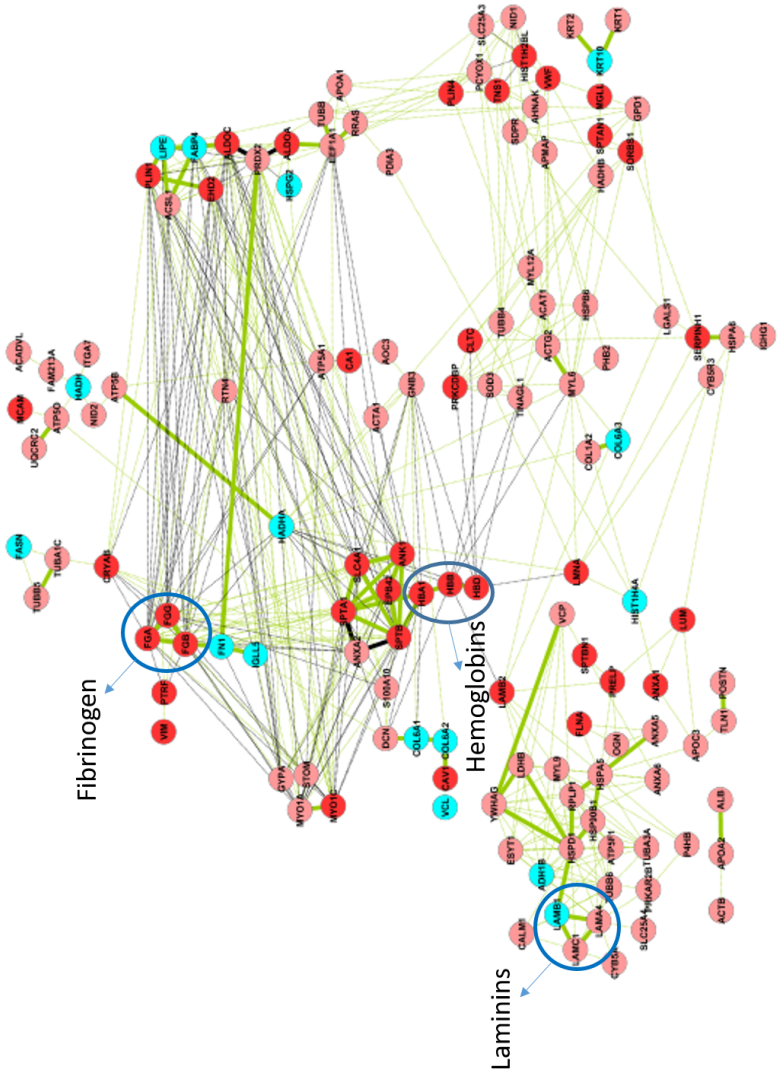


FIGURE 3.5: Network model for amyloidosis group **Control** with differential analysis of group **L**.

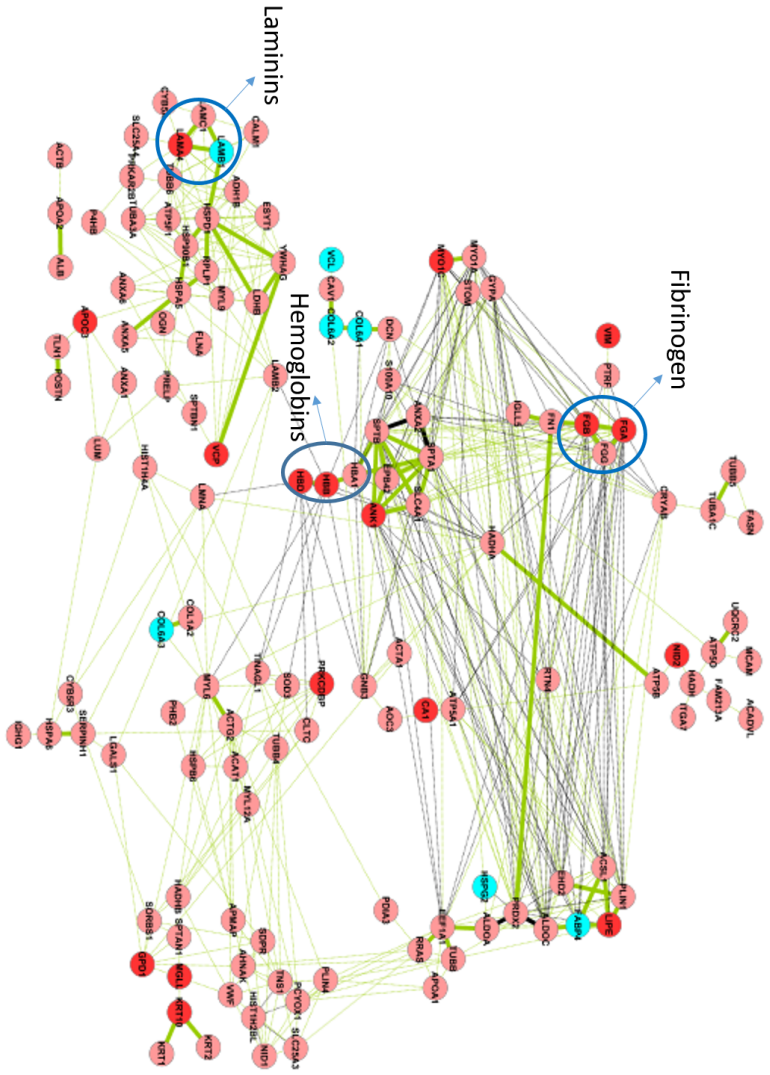


FIGURE 3.6: Network model for amyloidosis group**Control** with differential analysis of group **K**.

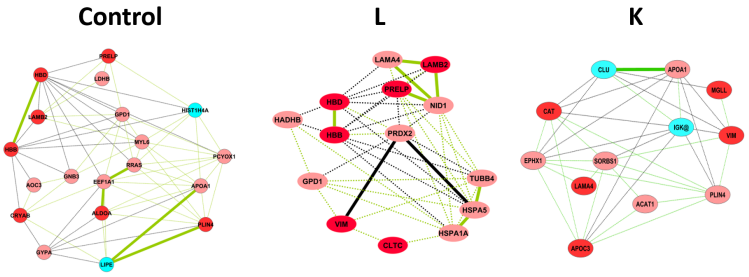


FIGURE 3.7: Three group-specific sub-networks made of nodes topologically relevant detected by CentiScape plug-in [121]

each found set of relevant proteins has been used to build a group-specific sub-network (Figure 3.7 shows the three sub-networks).

TABLE 3.2: Metric thresholds

	Betwenneess	Centroid	Stress
Control	177	-61	3454
L	163	-53	3114
K	133	-42	180

Moreover, the intersections among the three networks have been computed to investigate the relations in common but showing a different behaviour as regards the correlation sign. In this way some interesting proteins and relations have been selected; Figure 3.8 shows for each obtained intersection (**Control-K**,**Control-L** and **L-K**) just the common relations found with a discordant sign of correlation. These identified relations, thus the corresponding proteins, may represent key elements for the mechanism of action of the amyloidosis. These findings should be studied with deeper analysis or further biological experiments.

These results have been provided to the biologist group of the Proteomic and Metabolomic Laboratory of the Institute of Biomedical Technology of the Italian Research Council (ITB-CNR). The analysis of these models led to the identification of four relevant proteins, involved in the biological pathway of the lipid metabolism, showing a different behaviour in the three models both among them and with the other proteins in the networks (see Figure 3.9); for these reasons they could play a key role in the molecular mechanism of action of the amyloidosis disease. Finally, the first neighbors of these proteins have been selected to investigate the behavior of these key proteins respect to the other proteins in the three group-specific networks. The result is showed in Figure 3.10. The configurations obtained seem like coloured wheels, for this reason a funny

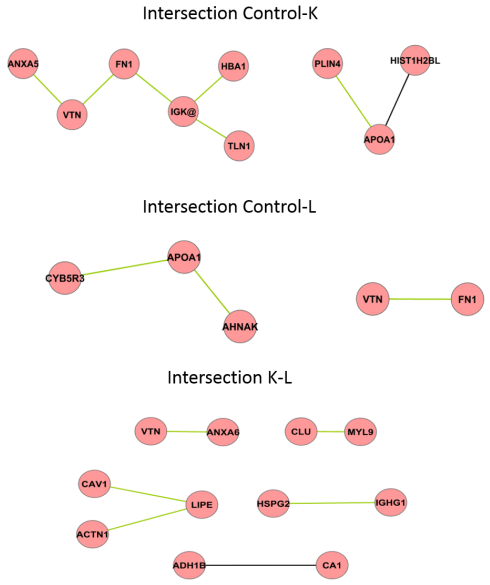


FIGURE 3.8: The Figure shows for each intersection computed the common relations with discordant correlation sign. In **Control-L** intersection the green color indicates positive sign in **control** group and negative in **L**. In **Control-K** intersection the green color indicates positive sign in **control** group and negative in **K**. Finally, in **K-L** intersection the green color indicates positive sign in **L** group and negative in **K**

.

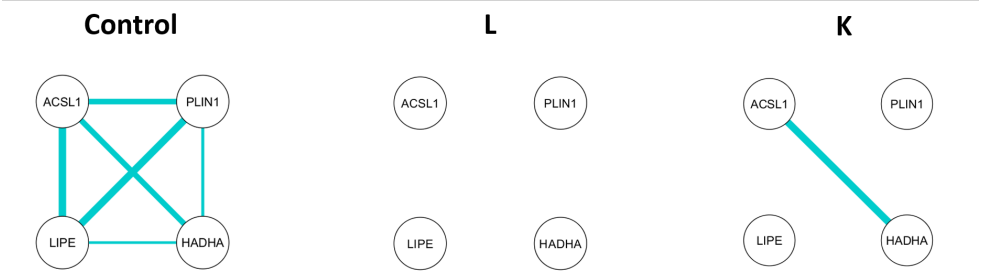


FIGURE 3.9: The Figure shows the Pearson correlation relations found for the four identified proteins in the three conditions. The edge thickness represents the correlation value found, ranging from 0.6 to 0.9

.

name was coined by biologist **"Proteomics eyes"**, because of similarity with the famous **London eye**. It is interesting to note that the selected proteins show in many cases a very different number of neighbors in the different conditions. For example, LIPE proteins has many connections in **control** group, but just three connections in **K** group; this finding may suggest that the biological processes in which this protein is involved are impaired because of the effect of amyloidosis. As consequence LIPE protein may have a key role for the investigated disease. These **"Proteomics eyes"** have been used to perform an enrichment analysis to find the Reactome Pathways [171] most represented in each sub-network, in order to find some impaired biological functions because of the amyloidosis disease. The analysis has been executed using ReactomeFIViz, a Cytoscape plug-in [100]. The results show that some biological functions in **Control** group are represented by a bigger set of proteins respect to the disease group, **K** and **L**. For example in Figure 3.11 the results of the analysis for ACSL1 protein are showed: almost the same pathways are identified in the two conditions and interesting, in the **Control** group the pathways are represented by a bigger set of proteins respect to **L** group. This situation may reflect that these functions are impaired in the disease group, as these proteins are involved in the mechanism of action of the pathology. Clearly, these findings are just hypothesis and should be tested by further focused experiments; however, we can conclude that the obtained models are a valuable base for improving the comprehension of the studied biological context and to infer information for generating new hypothesis. The described method represents an attempt to build context-specific protein networks starting from a data-set of MS proteomics data about amyloidosis disease. This elaboration has the final aim to provide an easy-to-understand model of the phenomenon studied, in order to facilitate the work of interpretation and understanding of the big quantity of proteomics data, in particular to support the work of biologist or clinician to extract from these rough data-sets useful information for biological/medical applications (the whole analysis process is described in Figure 3.12).

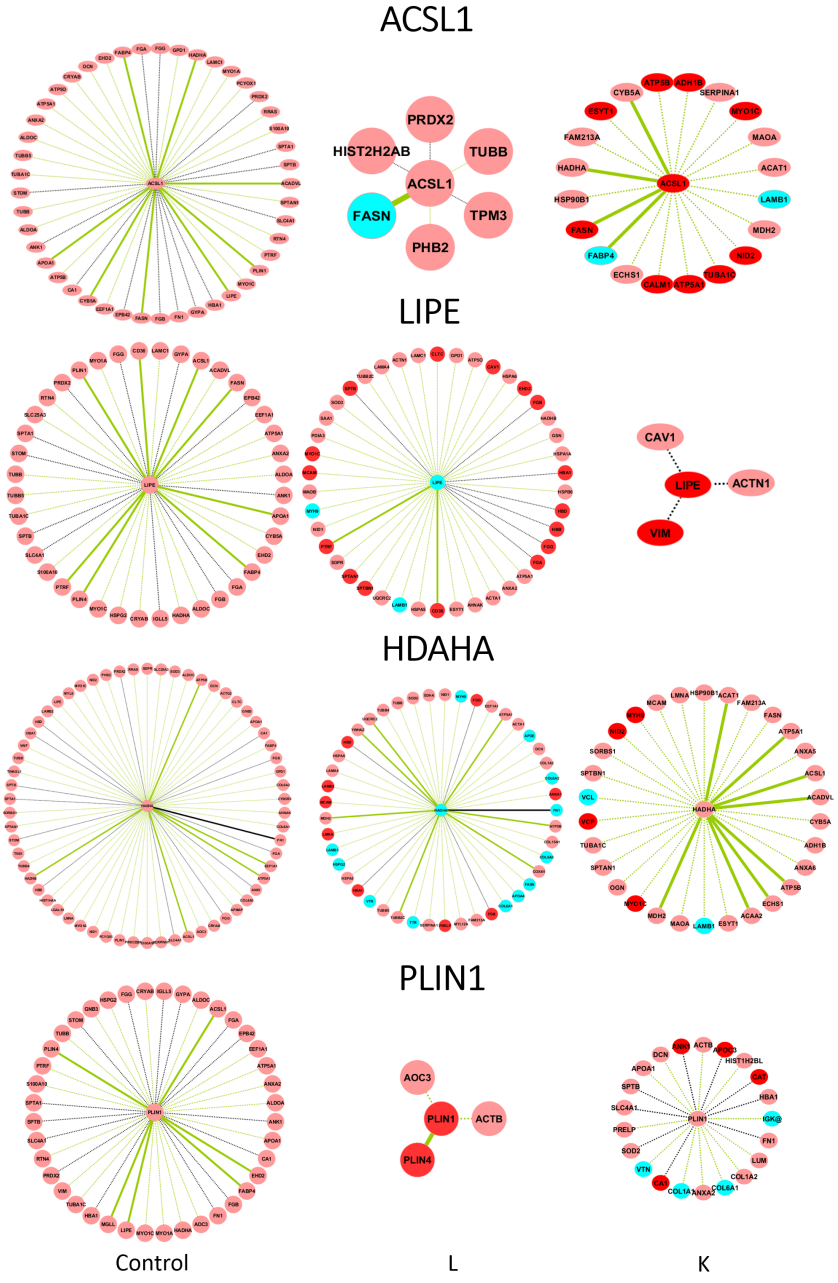


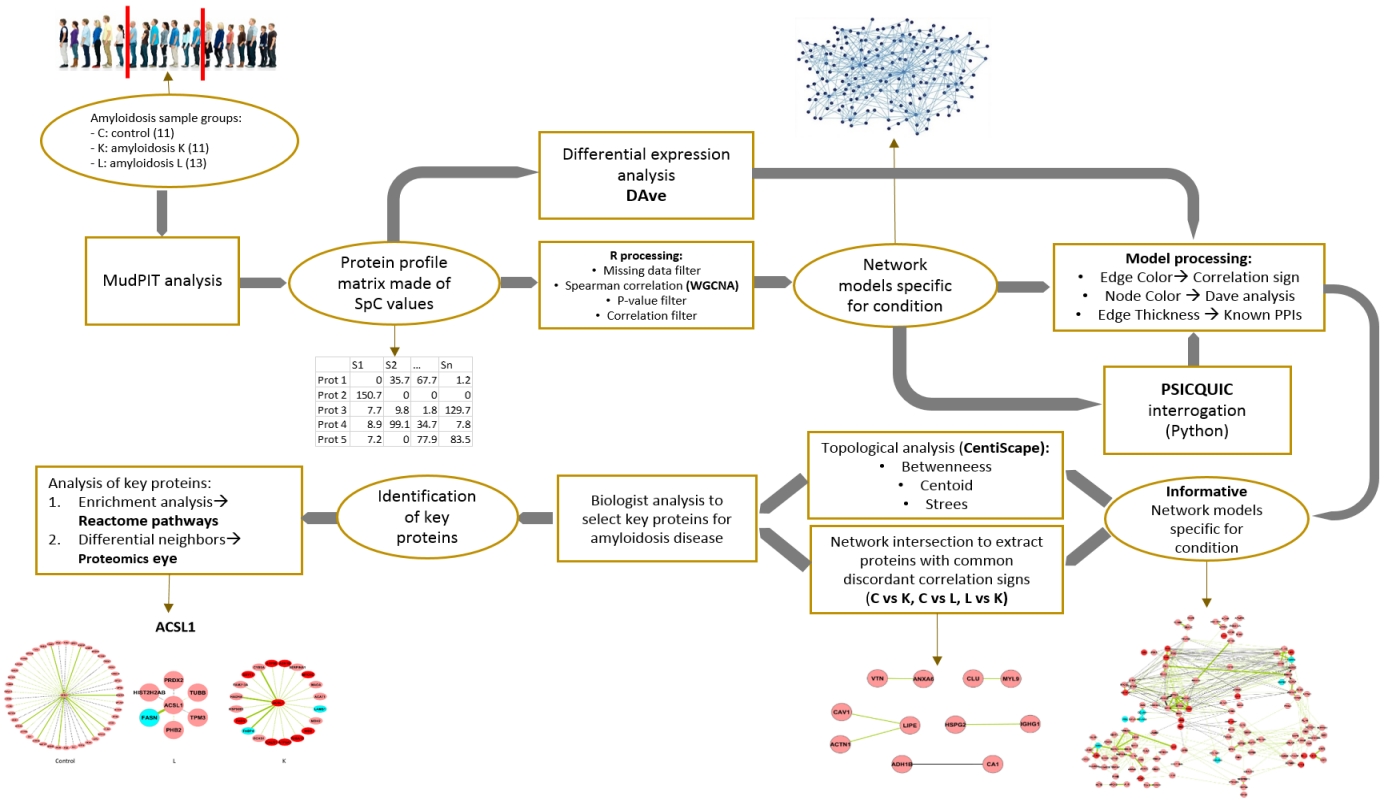
FIGURE 3.10: The Figure shows the first neighbors of the four identified proteins in the three group-specific network **Control** (left), **L** (center) and **K** (right)

.

Reactome Pathway in Control	P-value	Proteins
Fatty acid, triacylglycerol, and ketone body metabolism	0,0002	HADHA,APOA1,ACSL1,FASN,GPD1,ACADVL
Triglyceride Biosynthesis	0,0004	ACSL1,FASN,GPD1
Metabolism of lipids and lipoproteins	0,0009	HADHA,APOA1,ACSL1,PLIN1,FASN,GPD1,ACADVL,FABP4,LIPE
Fatty Acyl-CoA Biosynthesis	0,0025	ACSL1,FASN

Reactome Pathway in L	P-value	Proteins
Fatty Acyl-CoA Biosynthesis	0,0001	ACSL1,FASN
Triglyceride Biosynthesis	0,0002	ACSL1,FASN
Fatty acid, triacylglycerol, and ketone body metabolism	0,0080	ACSL1,FASN

FIGURE 3.11: Enriched Reactome pathways for the ACSL1 Proteomics eyes related to the groups **Control** and **L**.



	S1	S2	...	Sn
Prot 1	0	35.7	67.7	1.2
Prot 2	150.7	0	0	0
Prot 3	7.7	9.8	1.8	129.7
Prot 4	8.9	99.1	34.7	7.8
Prot 5	7.2	0	77.9	83.5

FIGURE 3.12: Analysis Pipeline.

Chapter 4

MTGO algorithm

4.1 Introduction

The increasing amount of -omics data leads to development of models to interpret and analyse them. A common approach consists in representing data as PPI Networks. These models can be very complex and informatics tools are needed to analyse them. In this chapter MTGO, an algorithm of module detection specific for PPI Network, will be presented. This algorithm exploits both the network topological information and the Gene Ontology (GO) knowledge about network proteins. MTGO output consists in a network partition, where each obtained cluster is labelled with a specific GO term describing its biological nature. In a single step, MTGO performs a double PPI network analysis; from a topological perspective, through the individuation of a meaningful network partition and, from a biological perspective, through the selection of significant GO terms describing the biological role of network proteins.

Some preliminary results on MTGO algorithm have been presented for the workshops NETTAB 2017 *Methods, tools & platforms for Personalized Medicine in the Big Data Era* and have been published in the abstract *MTopGO: a tool for module identification in PPI Networks* by following authors Danila Vella, Simone Marini, Francesca Vitali and Riccardo Bellazzi [172]. Currently, a scientific paper on MTGO algorithm, including some results of the validation analysis presented in Chapter 6, has been submitted to Scientific Reports Journal: *"MTGO: PPI network analysis via topological and functional module identification"* by Danila vella, Simone Marini, Francesca Vitali, Dario Di Silvestre, Giancarlo Mauri and Riccardo Bellazzi.

4.2 Input and output

A *PPI* network can be represented as $G = (V, E)$, where V and E are the nodes and edges of the network, respectively. V is the set of proteins and it is defined

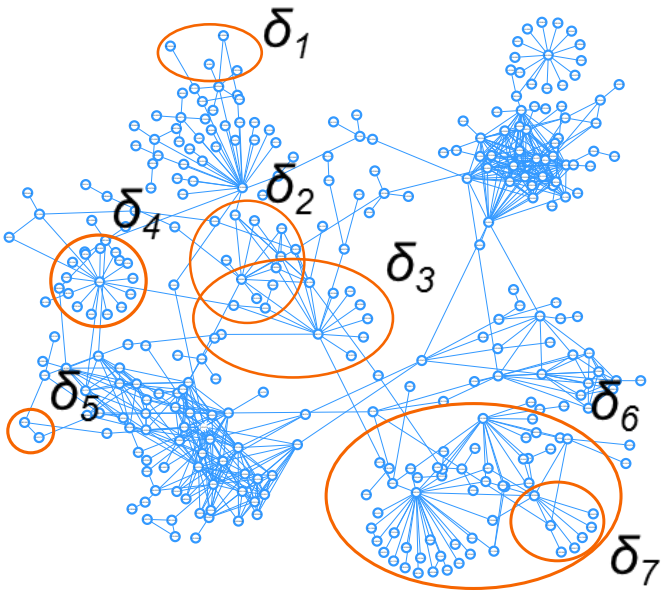


FIGURE 4.1: Example of δ elements represented in a network, they may share more nodes or be included into a bigger category.

as V $= \{v_1, v_2, v_3, ..., v_N\}$, with N is the total number of proteins/nodes. E represents the set of all the relationships between network nodes and it is defined as E $= \{e_{i,j} | 1 < i, j < N, i \neq j\}$. Therefore, G carries the *PPI* topological properties. In order to integrate biological function information in the PPI Network, we can assign GO terms to the network nodes. Given a user-provided list of GO terms (e.g. the entire GO or a sub-list), MTGO computes the set $T = (L, \Delta)$, where the $p - th$ element is $t_p = (l_p, \delta_p)$, l_p is the ontology term, while δ_p is the l_p-associated set of network proteins. Examples of the network δ_p elements and their structure are shown in Figure 4.1. Note that if a GO term of the input list is not associated with any network protein, MTGO automatically filters it out.

$I = (G, T)$ is the input of the system. The goal of MTGO is to process G to find groups of nodes sharing both the topological (V, E), and the functional (T) properties. The result of MTGO is the final output $R^F = (C^F, \Phi^F)$, where C^F is the set of the topological modules, Φ^F is the set of functional modules, and H is the total number of topological and functional modules, i.e. $|C| = |\Phi| = H$. The relation between the elements of C and Φ is 1:1. MTGO iteratively computes C and Φ, and the pair $R^F = (C^F, \Phi^F)$ is selected as final output. Note that modules are generally called *clusters* in literature. Since MTGO considers two different kinds of modules, here for clarity and simplicity we will not use the term cluster, but *topological* and *functional* modules. The model R is a global representation of the system in terms of modules, each one with a topological

(C^F) and a functional (Φ^F) representation. The set of the topological modules C is a partition of the network, defined as $\mathrm{C} = \{c_1, \ldots, c_\mathrm{h}, \ldots, c_\mathrm{H}\}$ such that:

$$c_1 \cap c_2 \ldots \cap c_\mathrm{h} \ldots \cap c_\mathrm{H} \equiv \emptyset; c_1 \cup c_2 \ldots \cup c_\mathrm{h} \ldots \cup c_\mathrm{H} \equiv \mathrm{V}; \tag{4.1}$$

Note that by definition, each node of a partition C is uniquely assigned to a single topological module. The set $\Phi = \{\varphi_1, \ldots, \varphi_\mathrm{h}, \ldots, \varphi_\mathrm{H}\}$, on the other hand, describes the functional modules involved in the network. Φ is defined as follows:

$$\varphi_1 \cap \varphi_2 \ldots \cap \varphi_\mathrm{h} \ldots \cap \varphi_\mathrm{H} \neq \emptyset; \varphi_1 \cup \varphi_2 \ldots \cup \varphi_\mathrm{h} \ldots \cup \varphi_\mathrm{H} \subseteq \mathrm{V} \tag{4.2}$$

Where $\Phi \subset T$, i.e. Φ is the subset of T selected by MTGO to describe the biological functions the PPI network represents.

Full coverage and *overlapping* are considered the ideal features of module identification algorithms [145]. MTGO grants both with its dual complementary output C and Φ, respectively. In particular, the C topological modules represent a network partition, thus granting full coverage by definition. On the other hand, the Φ functional modules *overlap*, allowing the assignment of a node to two or more modules. This feature is particularly important since it reflects the behavior of biological systems, where a protein may be involved in multiple functions.

4.3 MTGO functions

The MTGO algorithm is based on four main functions: two local metrics *Modularity Variation* MV, as regards topological aspects, and *Selection* γ, as regards biological aspects, applied *locally* on each single module during the building process; and their *global* counterparts *Modularty* (topological) and *QGO* (biological), which are applied to the whole network.

4.3.1 MTGO local metrics

Function γ is used to assign to a single topological module c_h a pair $t_B = (l_B, \delta_B)$, where l_B is the GO term best describing the biological meaning of c_h, and δ_B is (i) the list of network nodes associated to l_B, and (ii) the functional module linked to c_h. To explain the meaning of c_h and t_B, a practical example is presented below. Let's suppose that $c_h = C$, where $C = \{K7FU93, P20065, Q6S9C5\} \subset V$, then a possible example of t_B is $l_B = \{GO : 1905271\}$ and $\delta_B = D$; where D is the set of proteins/genes associated to $l_B = \{GO : 1905271\}$ and belonging to V,

for example $D = \{G3T9E9, K7FU93, P20065, P62327, P62328, P62329, Q6S9C5\} \subset V$. The nodes in the set $S = \{G3T9E9, P62327, P62328, P62329\}$, such that $S \subset D$ and $S \notin C$, are distributed to other topological modules in the network.

The rationale is that nodes are added/removed to let the topological module c_h fit its assigned functional module t_B, preserving the topological nature of the module (checked with the MV function).

Selection function

We define the γ function as

$$\gamma(\delta^k_{p,h}) = \frac{|\delta^k_{p,h}| - |\delta^k_{p,h} \cap c^k_h|}{|\delta^k_{p,h}| - 1} + \frac{|c^k_h| - |\delta^k_{p,h} \cap c^k_h|}{|c^k_h| - 1} \tag{S1}$$

where c^k_h is the *h-th* topological module at iteration k; $\delta^k_{p,h}$ is the *p-th* element of Δ^k_h and it represents a candidate functional module linked to c^k_h; and Δ^k_h is a subset of the set Δ (see Section 4.2 for Δ details), i.e. the GO terms associated to the network nodes belonging to c^k_h. The term $|\delta^k_{p,h} \cap c^k_h|$ represents the intersection between the topological module c^k_h and the candidate functional module. In the formula S1, the first addend has a low value (i.e. a good value) when the nodes belonging to $\delta^k_{p,h}$ and excluded from c^k_h are few respect to the total nodes contained in $\delta^k_{p,h}$. While, the second addend has a low value (i.e. a good value) when the nodes belonging to c^k_h and excluded from $\delta^k_{p,h}$ are few respect to the total nodes contained in c^k_h. The value 1 at the denominator allows that the result of each addend has as superior limit 1; the superior limit represents the worst case when the intersection between c^k_h and $\delta^k_{p,h}$ is just 1 (the minimum value of intersection possible).

The aim of *Selection* γ function is to choose a GO term (represented by $\delta^k_{B,h}$) as model to drive the building process of a topological module.

$$\delta^k_{B,h} = \underset{\delta^k_{p,h} \in \Delta^k_h}{\operatorname{argmin}} \gamma(\delta^k_{p,h}) \tag{S2}$$

The GO term assigned to a topological module should assure a good fitting, thus proving *both* a high overlapping and high specificity. In general, if a GO term has high degree of overlapping, it is very likely that the GO term is little specific for the topological module and vice-versa. In fact, the choice of a GO term involving all the topological module nodes could lead to select a low specificity GO term. A low specificity GO term contains many nodes not belonging to the topological module (i.e. included in other topological modules in the network). This problem is depicted in Figure 4.2. *Selection* γ is designed, therefore,

to find a trade off between overlapping and specificity. Figure 4.3 shows the behavior of the *Selection* function.

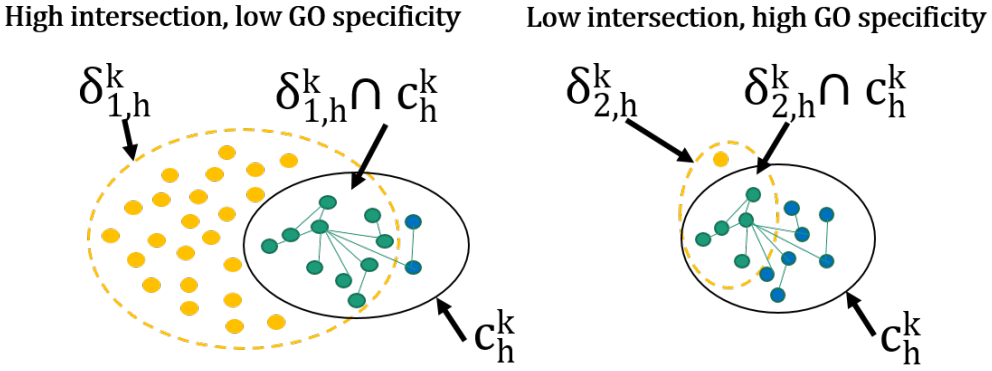


FIGURE 4.2: Two cases of fitting of δs to a topological module c_h^k. In the first case, $\delta_{1,h}^k$ has a good overlap with the topological module, involving almost all the nodes (green nodes), but it tags many other nodes (yellow nodes) outside of the topological module as well. Thus, $\delta_{1,h}^k$ is not very specific to this topological module. In the second case, $\delta_{2,h}^k$ is very specific for the topological module, in effect almost all $\delta_{2,h}^k$ nodes are included in it, but it has a low overlap (green nodes).

MV function

The main constraint of module modification is represented by the MV function. MV allows the topological nature of the module to be preserved. In fact, it is possible to add a node only if, by adding it, the module topology is ameliorated. We define the MV function as

$$MV(c_h^k, v_\mathrm{i}) = q(c_h^k + v_\mathrm{i}) - q(c_h^k - v_\mathrm{i}) \tag{S3}$$

Where q (S4) represents the contribute of a single topological module to global function Modularity Q (S5); $c_h^k + v_\mathrm{i}$ indicates the topological module including v_i, while $c_h^k - v_\mathrm{i}$ indicates the topological module without node v_i. Defining q as the modularity contribute of c_h^k, MV calculates the variation of q due to adding the node v_i.

$$q(c_h^k) = \frac{e_h^k}{|E|} - \left(\frac{d_h^k}{2 * |E|}\right)^2 \tag{S4}$$

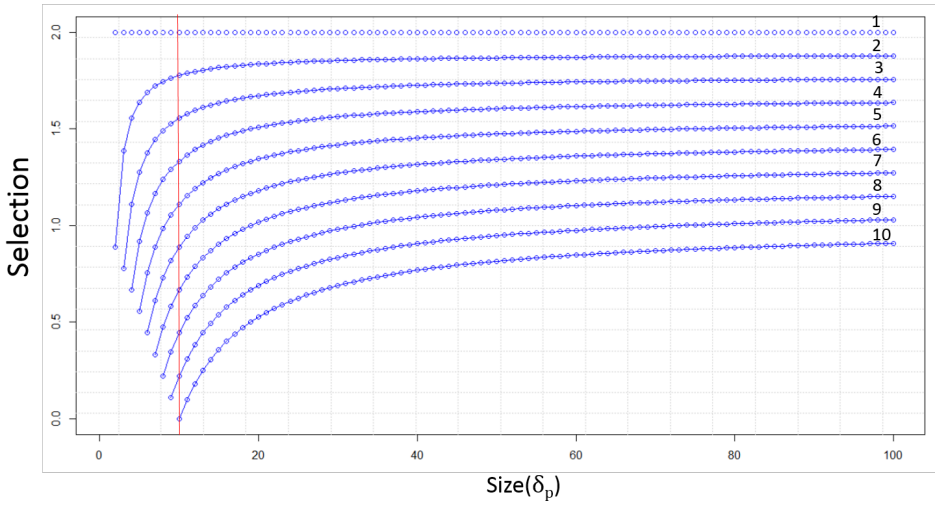


FIGURE 4.3: The *Selection* function operates a trade-off to associate a δ_p to a topological module, in particular mediating the δ_p nodes internal to the topological module vs. the external ones (Figure 4.2). This graphic shows the behavior of the *Selection* function when the size of the topological module is 10 nodes, and the size of the δ_ps ranges from 1 to 100. Each curve shows the function value for different values of intersection between topological module and δ_p, i.e. how many nodes are internal to module. The intersection values are showed in the right. The *Selection* value decreases according to both the δ_p size shrinking, and to the intersection value increasing.

In this function, e_h^k is the total number of edges in the *h-th* topological module; d_h^k is the sum of the node degrees of the *h-th* topological module; E is the number of total edges.

4.3.2 MTGO global functions

MTGO uses two different global functions to check if the convergence is reached: modularity (Q) [64] and Quality GO (QGO). Q evaluates the global quality of the partition C, while QGO evaluates the agreement between C and Φ. Ideally, C and Φ should overlap.

Modularity Q

Modularity Q is the most popular quality function for evaluating the graph partitions. The quality functions allow to assign a number to each possible partition, to assess the goodness and identifying a subset of meaningful ones for a graph. Its values range from -1 to 1, with positive values if there are more links within

topological modules than expected at random, and negative otherwise [156].

$$Q(C^k) = \sum_{1<h<H_k} \frac{e_h^k}{|E|} - \left(\frac{d_h^k}{2 * |E|}\right)^2 \tag{S5}$$

Here, C^k is the *k-th* partition; H^k is the number of topological modules; e_h^k is the total number of edges in the *h-th* topological module; d_h^k is the sum of the node degrees of the *h-th* topological module.

Modularity Q (S5) can also be written as the sum of $q(c_h^k)$s over the c_h^ks of a partition C^k.

Quality GO

$$QGO(C^k) = \frac{\sum_{1<h<H_k} |\delta_{B,h}^k \cap c_h^k|}{N_{GO}} \tag{S6}$$

Here $\delta_{B,h}^k$ is the functional module minimizing the *Selection* γ function for the topological module c_h^k (see *Iteration* Section, step 1); and N_{GO} is the total number of nodes with at least one δ_p assigned. QGO evaluates the degree of overlapping between C^k and Φ^k.

Moreover, *Modularity Q* (S5) can be written as the sum of $q(c_h^k)$s (S4) over the c_h^ks of a partition C^k.

4.4 MTGO algorithm.

In the following, we provide a description of MTGO. Given the input $I = (G, T)$, MTGO performs its tasks in three main phases: (i) initialization; (ii) iteration; and (iii) check for convergence. MTGO whole process is summed up in Figure 4.4.

4.4.1 Initialization.

In the initialization phase, V and E are used to create a random partition C^0 (Figure 4.5, Panel A), in which the number of topological modules is $\propto \sqrt{N}$. T is created from a GO term list provided by the user, according to the set V. Two user-defined parameters, *minSize* and *maxSize*, set the minimum and maximum size of T modules respectively, i.e. the minimum and maximum number of nodes in a δ_p.

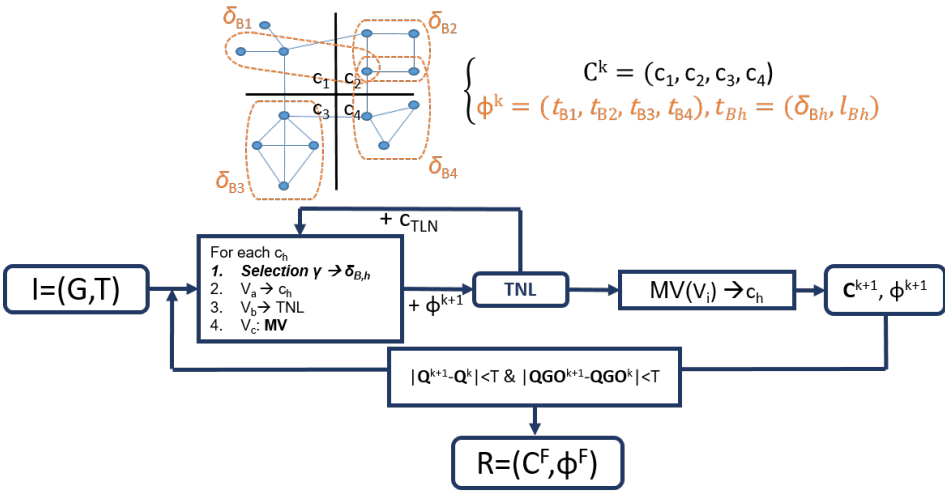


FIGURE 4.4: Workflow of MTGO. Iteratively, MTGO associates the functional module δ_{Bh} optimizing γ for each topological module c_h. Nodes of module c_h are redistributed according to the sets V_a, V_b and V_c. Hard-to-assign nodes are at first moved to the Temporary Node List (TNL). The TNL is emptied either moving its nodes to existing c_hs or to the newly created topological module c_{TLN}. At each iteration k, the output is a pair (C^{k+1}, Φ^{k+1}). MTGO checks threshold T for steady state. If reached, the pair C^F, Φ^F is the final output.

4.4.2 Iteration.

MTGO follows an iterative process. At each iteration, a pair (C, Φ) is computed: C by re-assigning the nodes of the previous partition, and Φ by selecting elements from T that best describe C. Each partition C is made of topological modules c_h with h representing the index of the single topological module and $1 \leq h \leq H$; (the total number of functional modules H varies at each iteration). Ideally, MTGO aims to assign nodes such that topological modules coincide with functional modules. In detail, the iteration phase is performed with two main sub-processes.

Step 1.

Topological modules are randomly processed at each iteration. Each c_h is processed as described in Figure 4.5. Firstly, $\delta_{B,h}$ is selected from the group of all the δs associated to c_h, i.e. the δs containing at least one node of c_h (Figure 4.5, Panels B and C). $\delta_{B,h}$ is the element minimizing the *Selection* function γ (S1), i.e. the one minimizing the number of not included nodes in $c_h \cap \delta_h$. The assignment of $\delta_{B,h}$ to c_h defines three node sets V_a, V_b and V_c. V_a is the set of nodes shared by $\delta_{B,h}$ and c_h; V_b is the set of nodes belonging to c_h but not to $\delta_{B,h}$; V_c is the

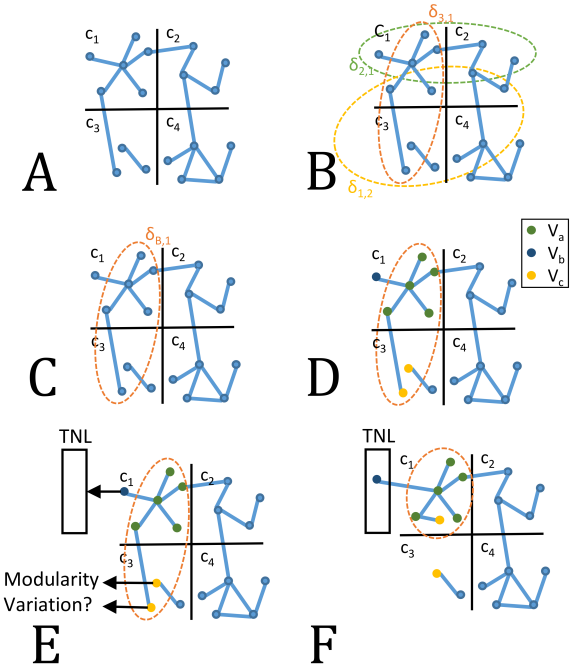


FIGURE 4.5: Iteration Phase of MTGO. Nodes are assigned to topological modules c_h (Panel A). Functional modules δ fit topological modules differently. For example, $\delta_{1,1}$, $\delta_{2,1}$, and $\delta_{3,1}$, overlap differently with c_1. The best functional module is $\delta_{3,1}$, since it minimizes the number of nodes out of the intersection between c_1 and itself. It is then selected as $\delta_{B,1}$ (Panels B and C). Once $\delta_{B,1}$ is selected, the nodes of $\delta_{B,1} \cup c_1$ are grouped into three sets: V_a, V_b, and V_c (Panel D). V_a are the nodes shared by $\delta_{B,1}$ and c_1; V_b are the nodes belonging to c_1 but not to $\delta_{B,1}$; V_c are the nodes belonging to $\delta_{B,1}$ but not to c_1. V_a nodes stay in c_1; V_b nodes are moved to the TNL; V_c nodes either remain in their topological module c_3, or are moved to c_1, according to the Modularity Variation function. Here, one V_c node is embedded in c_1, while the other stay within its original topological module c_3.

set of nodes belonging to $\delta_{B,h}$ but not to c_h. Note that V_c nodes belong to other topological modules of the partition (Figure 4.5, Panel D). From here, nodes in c_h are re-assigned as follows:

- V_a nodes remain in the topological module c_h.
- V_b nodes are moved to the *Temporary Node List* (TNL). The TNL is a temporary repository of nodes discarded from their original topological modules, and waiting to be re-assigned (Figure 4.5, Panel E).
- V_c nodes can either stay in their original topological module c_m ($m \neq h$) or be assigned to c_h, as they are biologically related to it, since they share $\delta_{B,h}$. A node $v_i \in V_c$ is moved to c_h if it increases the global Modularity [156] (S5 and further details in paragraph 4.3.1), according to a Modularity Variation (MV) function (S3), and in particular if $MV(c_h, v_i) > MV(c_m, v_i)$ (Figure 4.6).

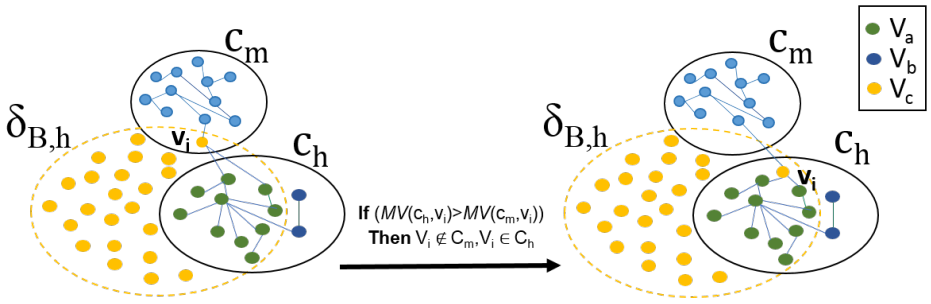


FIGURE 4.6: V_c node repositioning. The node v_i, belonging to $\delta_{B,h}$ and c_m moves to c_h topological module if $MV(c_h, v_i) > MV(c_m, v_i)$.

Step 2.

In this step the TNL nodes are re-assigned. All the TNL nodes with at least one associated δ, N_{GO}, are used to create a new topological module c_{TLN}. It is worthwhile to note that N_{GO} is a subset of the total nodes present in the PPI Network, in fact some nodes may not be covered by any GO term. While, each node v_i without any associated δ is assigned to the existing topological module optimizing the MV function (Figure 4.7). c_{TLN} is integrated into the network through the repetition of Step 1.

At the end of the Iteration phase, MTGO outputs the selected functional modules $\delta_{B,h}$s, along with their linked $l_{B,h}$s, grouped into Φ, and the newly computed topological modules c_hs, grouped into C.

Note that a detailed version of the MTGO Iteration phase is provided in the Appendix Section A1.

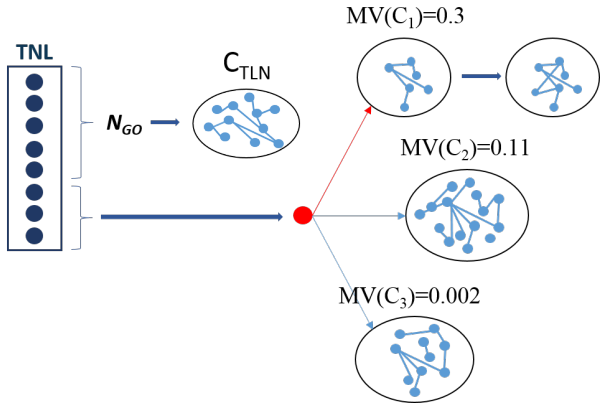


FIGURE 4.7: Step 2, the TNL is emptied. The nodes with at least one GO term (N_{GO}), the first TNL five nodes, are grouped to generate a new topological module c_{TLN}. Nodes without any GO term, the last three TNL nodes, are assigned to the topological module that maximizes the MV. In this example, the red node is assigned to the topological module c_1, showing the max value of MV.

4.4.3 Check for convergence.

Two different functions are used to check if the convergence is reached: Modularity (Q) (S5) [64] and Quality GO (QGO) (S6). Q evaluates the global quality of the partition C, while QGO evaluates the agreement between C and Φ. Set a threshold T, the steady state is reached when $|Q^{k+1} - Q^k| < T$ *and* $|QGO^k - QGO^{k-1}| < T$. When this condition is verified, the network partition configuration is stable, i.e. across next iterations the module composition aren't undergo to significant changes; both from a topological points of view (checked with Q function), and from a biological point of view (checked with QGO function). The solution $R = (C^F, \Phi^F)$ is taken as the one with maximum value of QGO. The set C^F is the partition maximizing QGO, while the set Φ^F is the set of all pairs $t^F_{B,h} = (\delta^F_{B,h}, l^F_{B,h})$ assigned for each c^F_h topological module. Note that in our experiments, we set $T = 10^{-4}$.

4.4.4 Parameters

The parameter *minSize* ranges from 2 to 15 and is used to limit the minimum number of nodes in modules. The parameter *maxSize* ranges from 30 to 300 and is used to limit the maximum number of nodes in modules. Tuning these

parameters is useful to adjust the final output in accordance to the needs of a specific study. For example, if the user is interested in identifying a specific process or small protein complexes, *minSize* and *maxSize* should be set to small values (e.g. 2 and 5). In order to identify the general/high level biological processes involved in the network, on the other hand, the user should set *minSize* and *maxSize* to high values (e.g. 10 and 200). Figure 4.8 shows an example of how different results can be obtained from the same network by changing the parameters.

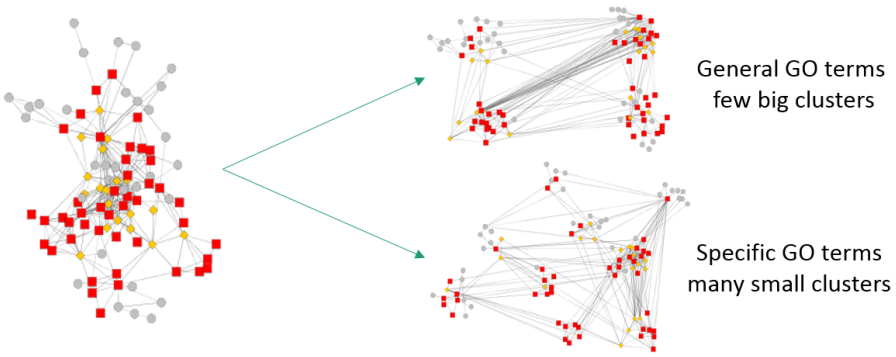


FIGURE 4.8: Different values of parameters *MaxSize* and *MinSize* lead to different results.

4.4.5 Density optimization result

MTGO provides also an alternative output R_D based on density optimization, to be used in case of poorly GO-enriched networks (see MTGO User Manual). For example, when the GO term list is very poor and the node percentage covered by the GO terms are under 50%. The solution $R_D = (C^D, \Phi^D)$ is taken as the one with maximum value of mean density over topological modules. For this solution, the set Φ^D is scarce in term of biological quality, while the set C^D preserves its quality in term of topological properties.

Chapter 5

MTGO: use and applications

5.1 Introduction

In the context of -omics approaches, Next Generation Sequencing instrument [173] and mass spectrometry-based proteomic technologies [174] are even more increasing amount and quality of data produced providing a good snapshot of the investigated systems. Due to the big mass of data produced per experiment, the challenge for biologists and bio-informaticians is to decipher the molecular mechanisms underlying the emergent phenotypes. For this purpose, both experimental transcript and protein levels are often combined with PPI models. Thus, computational tools dedicated to data processing and data mining are key to improve their functional evaluation and to facilitate their interpretation. MTGO tool belongs to this category; it provides the possibility to process a protein network model in order to extract useful information and elucidating the protein role respect to the biological system in which they are involved.

In this chapter, the main characteristics of the tool will be described in order to show how it can be a applied to a network model to obtain a graph clusterization and a biological description of the protein modules involved. Moreover, two practical examples of MTGO application will be presented.

5.2 MTGO software version

MTGO (Module detection via Topological information and Gene Ontology (GO) knowledge) is an algorithm of module detection in PPI networks. The module search is based on network topology and knowledge on the biological role of proteins. MTGO employs repeated partitions of the network; at each iteration, a new partition is created starting from the previous one, reshaping the modules on the basis of the GO annotations and the graph modularity. In this way the partition is learned through a process of optimization taking into account both the structure of the network and its biological nature. MTGO can be used to

analyze a PPI network to elucidate the biological processes involving the phenomenon represented by the network. In particular, the import of MTGO result in Cytoscape [175] allows to re-arrange a PPI network grouping nodes into functional modules. A software version of MTGO is available freely for non-commercial purposes at https://gitlab.com/d1vella/MTGO; the software has been implemented using the Java Programming Language. In this section the input and output files will be described in detail to show how to use in practical the MTGO software version.

5.2.1 Input files

Two are the input files of MTGO software version:

- *Edge*: this file contains the edges of the PPI Network. The file contains two columns of nodes, each row indicates an edge.
- *GO*: this file contains the Gene Ontology (GO) terms. The first column is a list of proteins and the second a list of Gene Ontology terms. Each row indicates an association between a protein and a GO term. The protein list contains the proteins of the PPI network. In the best case all the proteins of the PPI network are contained in the file; anyway MTGO is able to work even if some network proteins aren't associated to any GO term. The number of proteins associated to one or more GO terms is indicated in the output file *Single iteration properties* (described in the next paragraph) in the column *NGO*. If the protein list of the file *GO* contains proteins not included in the PPI network, they are ignored by MTGO.

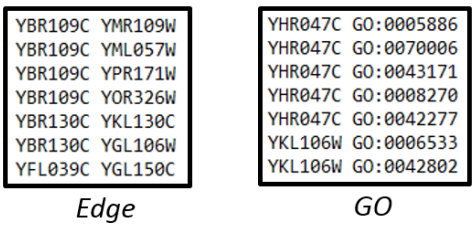


FIGURE 5.1: Example of input data: Edge and GO files.

The *GO* file can be build in two different ways:

- downloading the annotation file directly from the Gene Ontology Consortium website at link http://geneontology.org/page/download-annotations, selecting the corresponding organism of the input PPI network [54]

- using Bingo plug-in of Cytoscape [176] to select the list of Gene Ontology terms enriched in the network.

The first method consists in retrieving the whole list of GO terms linked to a given organism, this list contains a big quantity of terms. The second method allow to obtain a pre-selected list of GO terms, thus a smaller list respect to the first method. These terms selected by Bingo are specific for the PPI network; for this reason, this further step of analyses facilitates the MTGO research of the best-fitting GO term set for the network model, allowing to obtain final modules of better quality.

5.2.2 Output files

MTGO software version returns six output files, namely *Modules_Best_QGO*, *Nodes_Best_QGO*, *Modules_Best_Density*, *Nodes_Best_Density*, *Properties* and *Single_iteration_properties*. The two main files are:

- *Modules_Best_QGO*, where each row represents a module:
  - **"Modules phi: nodes"**: the nodes of the functional module δ_{h} of the set Φ^{F}
  - **"Cluster C: nodes"**: the nodes of the topological module c_{h} of the set C^{F} associated to the functional module δ_{h}
  - **"Gene Ontology Term"**: the GO term l_{h} (corresponding to the functional module δ_{h}) attached to the topological module c_h
- *Nodes_Best_QGO*, this file contains two columns, the first one is the list of all proteins in the PPI network, the second one indicates the GO terms (l_{h}, from the set Φ^{F}) attached to the topological modules of the set C^{F}. Each row contains a protein belonging to a topological module c_h and the GO term l_{h} associated to c_h.

For further explanations of the elements named in this files (δ_{h}, c_h, etc..) see Chapter 4. These two files allow to import the MTGO results into Cytoscape. *Nodes_Best_QGO* can be used to obtain the configuration showed in Figure 5.3, through the import of the GO terms as node attribute of the PPI Network and using the Cytoscape function **"Group Attribute Layout"**.

The user can use this configuration and the information contained in *Modules_Best_QGO* file to obtain a final result where each module is attached to its GO description. The C^{F} topological modules cover all the PPI network and can be used as a network partition, while the Φ^{F} functional modules don't cover all

A

Index	Modules Phi: nodes	Clusters C: nodes	Gene Ontology term
333	YBL026W,YBR152W,YCR020C-A,'	YBL026W,YBL074C,YBR055C,YBR	GO:0000398
318	YBR251W,YGR084C,YKL155C,YM	YBL090W,YBR251W,YCR089W,Y(	GO:0005763
106	YBL041W,YBL058W,YBR105C,YC(	YBL020W,YBL041W,YBL058W,YB	GO:0043161
325	YDR156W,YHR143W-A,YJL148W,	YDL042C,YDL150W,YDR156W,YH	GO:0001054
342	YDL130W,YDR087C,YDR496C,YE(	YBR150C,YDL130W,YDL168W,YD	GO:0030687
116	YBL093C,YER022W,YHR058C,YNL	YBL082C,YBL093C,YBR193C,YBR2	GO:0070847
144	YDL007W,YDR394W,YIL007C,YKL	YBR168W,YDL007W,YDL036C,YD	GO:0008540
93	YBR154C,YDL108W,YDR145W,YD	YBL059W,YBR154C,YBR259W,YD	GO:0006366
322	YBR268W,YCR071C,YDR116C,YD(	YBR268W,YCR003W,YCR071C,YD	GO:0005762
125	YAL021C,YDL084W,YFR037C,YGL	YCL044C,YDL179W,YFR037C,YGL	GO:0006368
42	YBR067C,YBR171W,YDR164C,YD(	YAL029C,YBR067C,YBR109C,YBR	GO:0071944
209	YGR095C,YIL079C,YNL232W,YOL	YDL111C,YGR095C,YGR158C,YHR	GO:0071038
12	YBL084C,YDL156W,YGR142W,YH	YBL084C,YDL156W,YDL235C,YER	GO:0034399

B

Nodes	Nodes Gene Ontology
YBL026W	GO:0000398
YBL074C	GO:0000398
YBR055C	GO:0000398
YBR119W	GO:0000398
YBR152W	GO:0000398
YCR020C-A	GO:0000398
YCR063W	GO:0000398
YDL030W	GO:0000398
YDL085C-A	GO:0000398
YDL098C	GO:0000398
YDR235W	GO:0000398
YDR240C	GO:0000398
YDR378C	GO:0000398

FIGURE 5.2: Example of output files: (A) Modules_Best_QGO and (B) Nodes_Best_QGO.

the network but identify the functional units acting at the base of the biological system. Furthermore, they can overlap unlike topological modules. Using this two files a final network model can be obtained to improve the interpretation of the biological phenomenon described by PPI network (see Figure 5.4 and Figure 5.7 in the next paragraph). Other two files are provided also, containing the final result computed at the iteration corresponding to the Mean Density maximum value of the topological modules: *Nodes_Best_Density* and *Modules_Best_Density*. This result can be used in substitution to the **"Best_QGO"** result when the GO terms, provided as input, have a low quality or cover a little part of network nodes (low *NGO*). In effect, this result relies on much more the topological properties of the network than the Gene Ontology (see Paragraph 2.3.5 in Chapter 4 for further details).

Finally, there are two other files:

- *Properties* contains some general properties of the two final results **"Best_QGO"** and **"Best_Density"**:
  - **"Modularity"**: modularity of the partition C^{F}
  - **"QGO"**: QGO computed on the set C^{F} and Φ^{F}
  - **"Density"**: mean Density of topological modules of the set C^{F}
- *Single iteration properties* shows the same three values of the file Properties (Modularity, QGO, Cluster Mean Density) and the value *NGO* (the number of network proteins associated at least to one GO term) computed at each iteration.

FIGURE 5.3: Import of the MTGO results in Cytoscape, through *Nodes_Best_QGO file*.

5.3 Two case studies

5.3.1 Myocardial infarction network model

To show an application of MTGO on real data, an undirected PPI network was used; this model was obtained by analyzing the proteomics of swine heart tissues affected by myocardial infarction (MI) and treated by human mesenchymal stem cells [177]. The network is made of 502 nodes and 4316 edges consisting in physical PPIs and it is represented in Figure 5.4, panel A. Although it may be considered a network of medium size, its structure is too complex to be manually interpreted and the use of tools like MTGO may be of great support to tease out the main hidden biological functions and processes. The parameter value used are *minSize*=5, *maxSize*=30. The input list consists in 1256 GO terms, belonging to the GO class Biological Processes. The obtained results clearly outline well known heart physiology processes, including ATP synthesis coupled to electron transport, muscle system process, regulation of cell adhesion or lipid oxidation, and glucose metabolic process, all in agreement with the investigated samples [177]. Moreover, many of these processes are associated also to well defined protein groups (ribosomal complex, heterogeneous nuclear ribonucleoprotein complex, myosin complex, ATP synthase complex, Proteasome complex, T-complex proteins, NADH dehydrogenase complex) showing the attitude of MTGO to correctly identify molecular complexes (Figures 5.4 panel B; Appendix Section, Table C). The overlapping of functional modules is achieved via GO terms attribution, i.e. some nodes belong to more than one functional module (Figure 5.6 depicts the network without PPIs, with nodes representing proteins and GO terms connected by *belongs to* edges). The GO terms found are quite general, if the user is interested in conducting a deeper analysis a lower value of *minSize*, for example 2, could help in retrieving more specific GO terms to describe the PPI Network. Following the application of MTGO algorithm, we used Cytoscape [175] to split the network nodes in well defined functional modules (Figure 5.4, panel B); this structure may be more easily interpreted by biologists and further improve identification of processes and functions modulated in the considered phenotypes [177].

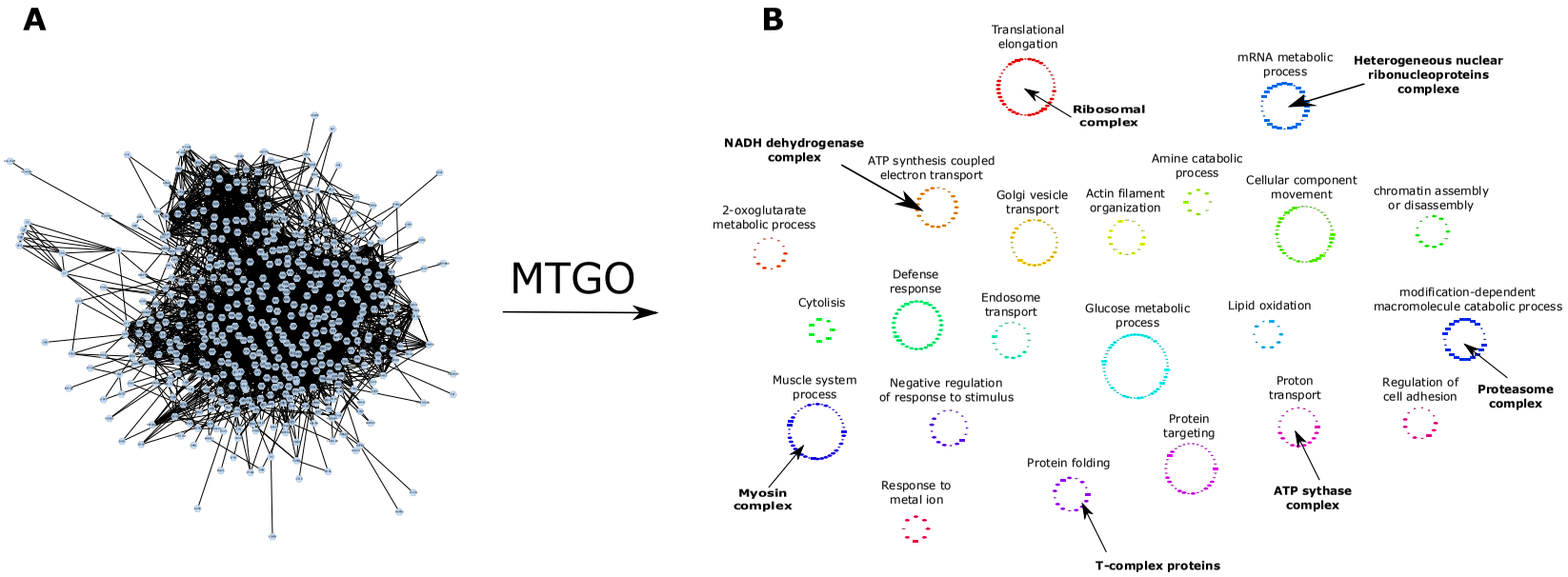


FIGURE 5.4: Application of MTGO algorithm to process an experimentally-derived PPI network. (A) Myocardial infarction PPI network consisting in 502 nodes and 4316 physical interactions. The network structure derives from Cytoscape following the application of the Organic layout. (B) myocardial infarction PPI network following MTGO algorithm and Cytoscape group attribute layout by using the MTGO output as attributes. Circular topological modules shown in figure correspond to topological modules obtained by MTGO (Appendix Section, Table A3). In the reported example we grouped our nodes using the MTGO C output to cover whole network. Finally, in both are indicated protein complexes associated with the assigned biological processes. Node details are explained in Figure 5.5.

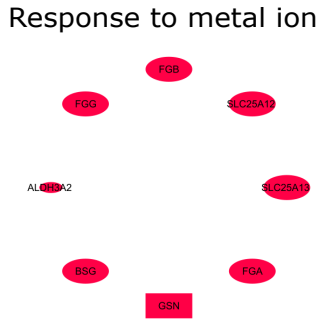


FIGURE 5.5: Detail of the network represented in Figure 5.4. Small nodes represent proteins excluded from the GO term assigned to the topological module. Big nodes (both circles and rectangles) indicate proteins associate with the same GO term assigned to topological module, while rectangles indicate nodes assigned to more than one functional modules.

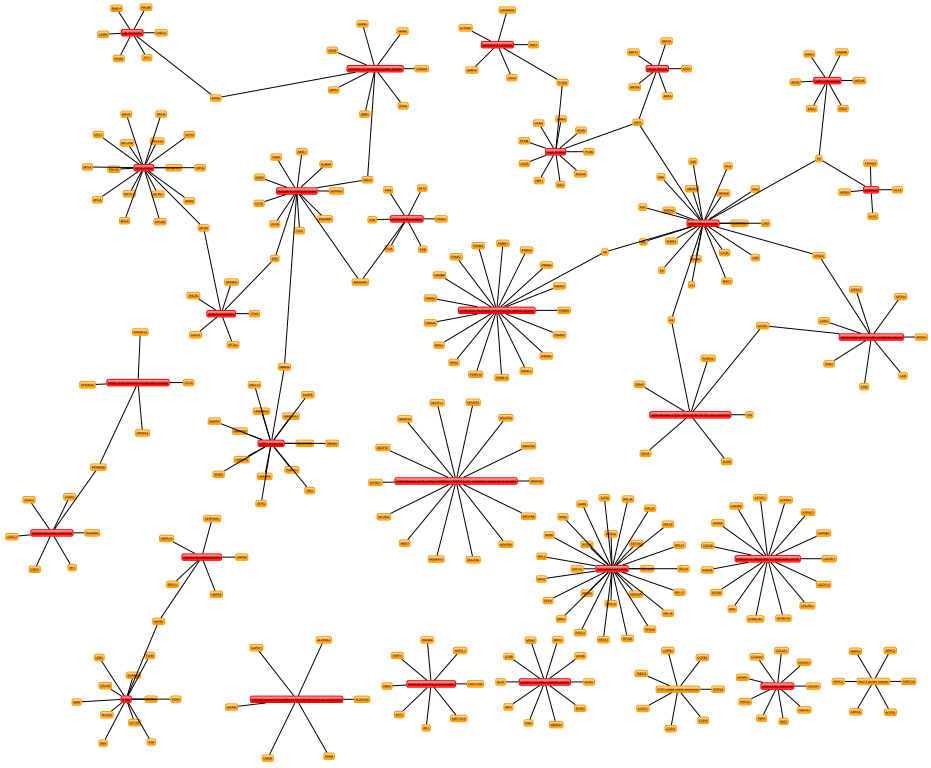


FIGURE 5.6: Functional modules identified by MTGO for the myocardial infarction PPI network. Here, links indicate protein membership to one or more functional modules, identified by the GO terms. PPI edges are not represented in this figure.

5.3.2 PPI Network model from String database

A second example of MTGO application regards a human PPI Network built with String database[92]. In this case, the network was built using as seed 54 genes mutated in acute myeloid leukemia [178] and retrieving PPIs from String database, both known interactions (curated databases and experimentally determined) and predicted interactions (gene fusion, gene neighborhood, gene co-occurrence). Some other interacting nodes were added to the seeds from String database to reach a final network of 78 nodes and 545 edges. The added nodes are retrieved automatically by String on-line application, among them interacting with the 54 seed nodes through known (curated databases and experimentally determined) and predicted (gene fusion, gene neighborhood, gene co-occurrence) relations. The list of GO terms includes those related to the Human organism and tagged with Experimental evidence and/or computational analysis evidence Score, for a total of 7909 terms [179]. The parameter value used are *minSize*=2, *maxSize*=30. The original PPI network and the MTGO output are showed in Figure 5.7. To evaluate the MTGO ability to detect a set of GO terms

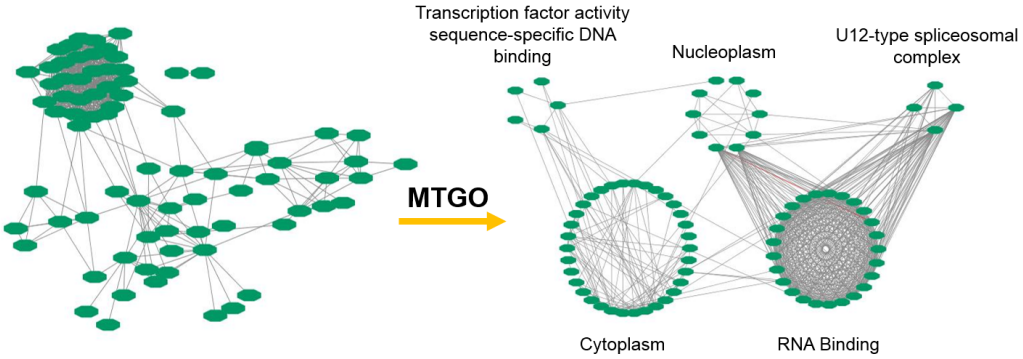


FIGURE 5.7: Example of MTGO applied on a Human PPI network. The algorithm produces a partition of 5 clusters, each one tagged with a specific GO term.

able to describe the network in terms of biological functions, the Fisher's exact test has been used to compute a p-value for each module and its corresponding GO term; the found p-values are all significant, under the 0.05 threshold (see Table 5.8). The details about the procedure of Fisher test application is reported in Paragraph 6.3.1. These results have been presented for the workshops NETTAB 2017 *Methods, tools & platforms for Personalized Medicine in the Big Data Era* and have been published in the abstract *MTopGO: a tool for module identification in*

	GO term	Description	P-value
1	GO:0005737	Cytoplasm	0.005033703
2	GO:0044822	RNA Binding	1.12E-12
3	GO:0005654	Nucleoplasm	0.000229789
4	GO:0003700	Transcription factor activity	0.003453057
5	GO:0005689	U12-type complex	6.17E-08

FIGURE 5.8: The p-values computed by Fisher's exact test of the GO terms attached by MTGO to each module.

PPI Networks by following authors Danila Vella, Simone Marini, Francesca Vitali and Riccardo Bellazzi [172].

Chapter 6

Validation

6.1 Introduction

To validate MTGO algorithm, a standard approach has been followed. In the same way as similar algorithms [60, 61, 153], MTGO has been compared with other state-of-the-art algorithms, including recent GO-based ones. Three different gold-standard protein complex sets have been used as target sets to compare the predicted modules to each algorithm. The algorithms were tested on four different PPI networks, from *Saccharomyces Cerevisiae* and *Human* organisms. Several metrics, presented in literature [153, 180, 181], were used to compare the target sets with the predicted modules, including *Accuracy*, *Maximum Matching Ratio*, *F-Measure* and *Composite Score*. The biological quality of the predicted modules was measured through GO Term Finder, a software used to perform GO enrichment analysis [65]; this approach has been used for the evaluation of similar algorithms [61, 158].

AS MTGO has the unique characteristic of providing a set of GO terms describing the PPI Network, a statistical approach has been designed to evaluate if this set of GO terms well explains the biological context represented by the network. For this reason, a statistical test was used in order to compute a p-value for each GO term provided by MTGO, so as to verify that each GO term is statistically significant compared to the protein set of the cluster to which it is associated with.

Furthermore, since a weak point in state-of-the-art algorithms is the detection of small or sparse modules, MTGO ability to detect these specific sets of modules was evaluated in detail. Finally, the MTGO run time has been evaluated respect to networks of different sizes. Most part of validation analysis has been carried out in collaboration with Ing. *Simone Marini* (PhD in Bioingeneering).

6.2 Data collection for seven scenarios

To evaluate the performance of MTGO, four real PPI networks have been selected, including Krogan PPI Network [182], Gavin PPI Network [183], Collins PPI Network [184] and DIP Hsapi Network [185]. The first three networks were built using yeast *Saccharomyces Cerevisiae* data, while DIP Hsapi network was built with Human data. Although the three networks of *Saccharomyces Cerevisiae* are in part overlapping, as they come from the same organism, it is important to test all of them because they are obtained with different experimental processes. In fact, the presence of false-positive edges and noise in a network is strictly dependent upon the experiment used to detect PPI, thus networks characterized by different noise sources should be used to test the robustness of module identification algorithms. Table 6.1 shows the main characteristics of each network, including the number of nodes covered by GO terms, used as input for MTGO.

TABLE 6.1: PPI network characteristics.

	Nodes	GO-covered nodes	Edges
Krogan	2709	2537	7123
Gavin	1856	1778	7669
Collins	1622	1596	907
Human	2734	2474	4058

This functional information has been retrieved downloading the annotation files submitted by GO Consortium members related to *Saccharomyces Cerevisiae* and *Homo Sapiens*. The GO terms used as input for MTGO include all the three categories of Cellular Component, Biological Process and Molecular Function (see Paragraph 2.2 in Chapter 2 for GO details). On the basis of reliability, only the GO terms tagged with an Experimental evidence and/or computational analysis evidence Score have been retrieved [54]. To evaluate the predicted modules with MTGO, gold standard protein complexes have been used as target sets. In fact, a protein complex is an aggregate of multiple proteins that interact with each other and perform certain biological activities [186]. This definition is conceptually very similar to the definition of a functional module. In detail, the protein complex sets used are: CYC2008 [187] and the union of MIPS [188] and SGD [189], for *Saccharomyces Cerevisiae* PPI networks; and CORUM [190] for Human PPI network. Protein complexes made of just one protein have been excluded. The number of curated complexes in CYC2008, MIPS+SGD and CORUM are 408, 509 and 1765, respectively. This led to *scenarios*, i.e. six for *Saccharomyces Cerevisiae* networks (Krogan, Gavin and Collins) against CYC2008

and MIPS+SGD target sets; and one for Human network against CORUM target set.

6.2.1 Q and QGO trends

Figure 6.1 shows the trend of the functions Q and QGO (see Paragraph 4.2.2 in Chapter 4 for function details). These two functions evaluate the global topological and functional properties of each partition. Each red line in Figure 6.1 depicts the value of Q computed with the fast greedy modularity optimization algorithm [191]. To compute this value, an implementation of the fast-greedy algorithm, provided by the package *iGraph* of the R Software, has been used. The red line can be considered the reference value of the maximum reachable modularity for each network. In the initial iterations the modularity shows a fast increment and it almost reaches its maximum value, while QGO shows a slower, steady increment. After reaching its peak, the modularity decreases, allowing a re-arrangement of the partition in order to improve the GO quality of the single topological modules (as supported by the slowly increment of global QGO). However, reaching the last iteration, when QGO reaches its maximum value, Q remains positive (i.e. GO quality increases at the expenses of the best modularity, but without cripple the topological properties of the partition). MTGO provides also an alternative model R_D based on density optimization, to be used in case of poorly GO-enriched networks (see Paragraph 4.3.5 in Chapter 4 for further details).

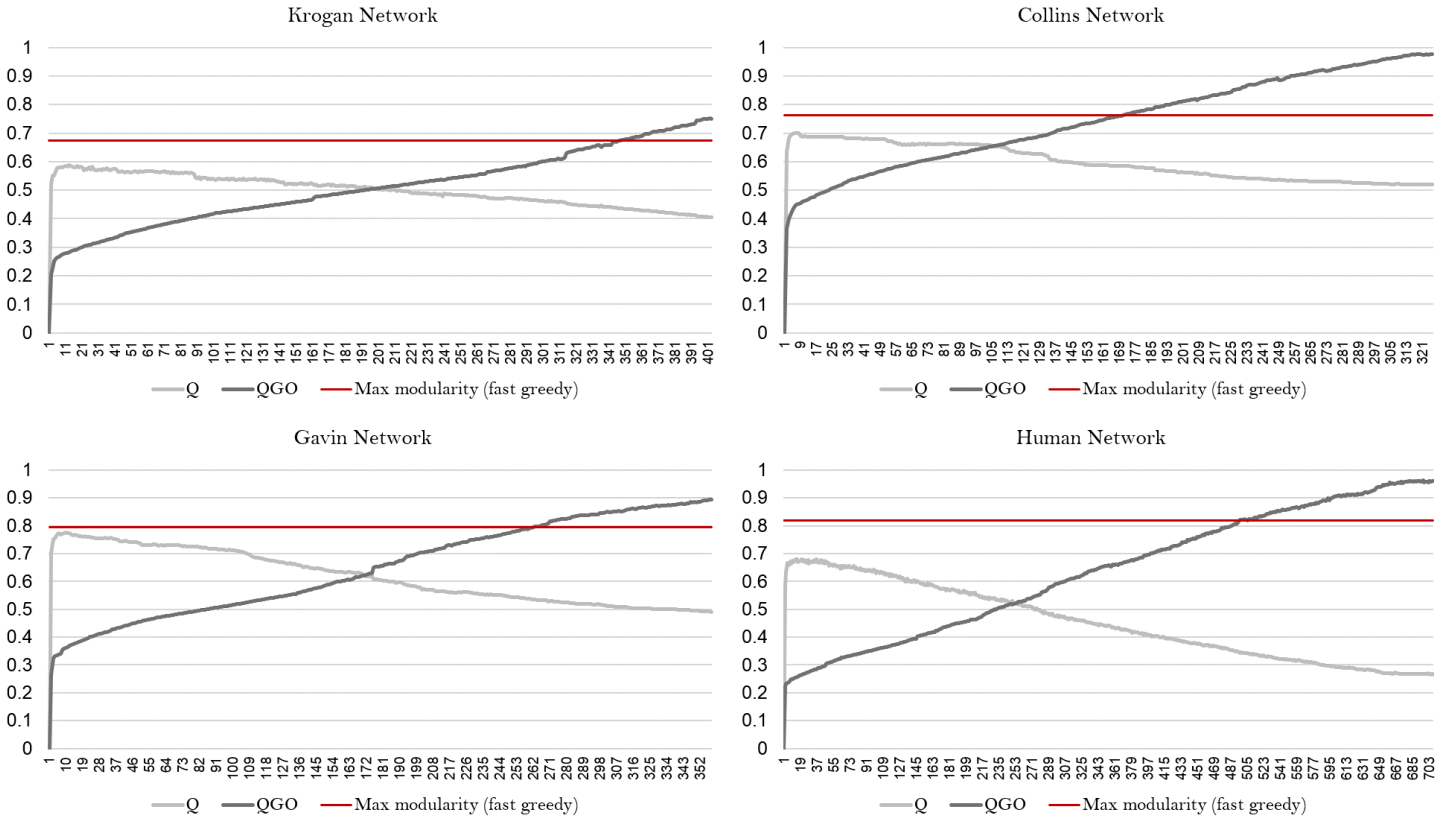


FIGURE 6.1: Q and QGO trends. Here the values of Q and QGO functions for the four networks, computed at each iteration, are showed. The x-axis indicates the value: Q, QGO, and mean density; while, the y-axis indicates the iteration number.

6.3 Comparison with other approaches

To evaluate the effectiveness of MTGO, results were compared with state-of-the-art algorithms. In particular, MTGO is compared with ClusterOne [153], MCODE [150], COACH [98], CFinder [149], Markov Cluster (MCL) [192] and DCAFP [61] and GMFTP [60]. While the first five algorithms are based only on topological properties, DCAFP and GMFTP, similarly to MTGO, exploit functional GO information as well. All the algorithms were run with default parameters, with the exception of the *k* parameter in CFinder, which has been chosen as the best among *k* = 4, 5 or 6 for each run. Note that this range is considered ideal for biological networks, as it is advised in literature [153]. MTGO parameters were set to default for Human network (*minSize*=2 and *maxSize*=100); for *Saccharomyces Cerevisiae*, on the other hand, *maxSize* was set to 80, according to the size of the biggest target complex [187] (see Paragraph 4.3.4 in Chapter 4 for parameters details).

Three independent measures were used to compare predicted complexes with the target sets: *Recall*, *Accuracy* [157] and *Maximum Matching Ratio* (MMR) [153]. To have a global vision another metric was used, the *Composite Score*, a comprehensive measure specifically introduced to assess module identification algorithms [153, 193]. The Composite Score is calculated as the sum of Recall, Accuracy and MMR. *Recall* is the fraction of true complexes matched by at least one predicted complex over the overlapping score (OS) [181], defined as

$$OS(TC, PC) = \frac{|TC \cap PC|^2}{|TC| * |PC|} \tag{S5}$$

where TC stands for Target Complex (module), and PC Predicted Complex (module). For this work, we set the OS threshold to 0.5. Recall is calculated as

$$Recall = \frac{N_{\text{OTC}}}{N_{\text{TC}}} \tag{S6}$$

where $N_{\text{OTC}} = \{C |\, C \in TC,\ \exists C \in PC,\ OS > 0.5\}$ is the set of TC with $OS > 0.5$.

Accuracy is defined as

$$Accuracy = \sqrt{Coverage * PPV} \tag{S7}$$

Where *Coverage* (also called Sensitivity) and *Positive Predictive Value* (PPV) are:

$$Coverage = \frac{\sum_i \max_j |TC \cap PC|}{\sum_i |TC|} \tag{S8}$$

$$PPV = \frac{\sum_j \max_i |TC \cap PC|}{\sum_j \sum_i |TC \cap PC|} \tag{S9}$$

where i indicates the i-th TC and j the j-th PC [181]. The Figure 6.3 shows a comparison of *Coverage* and *PPV* in all seven scenarios.

MMR has been introduced as a specific measure for module identification algorithms [153]. It is based on the maximal one-to-one mapping between PC and TC. It was proposed to overcome an *Accuracy*-related issue in the specific case of module identification algorithms, i.e. the misleading role of *Positive Predictive Value* if some proteins in a TC are present in either more than one PC or in none. The MMR is based on maximal matching in a bipartite graph, in which the two sets of nodes represent the reference and predicted complexes, respectively, and an edge connecting a reference complex with a predicted one is weighted by the overlapping score (OS) between the two. To obtain the maximum weighted bipartite graph, a subset of edges were chosen, such that the sum of the weights of such edges was maximal. The MMR between the predicted and the reference complex set is then given by the total weight of the selected edges, divided by the number of reference complexes. Moreover, *Recall* has been used to compute another metric, the *F-measure*, defined as follow:

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{S10}$$

where *Precision* is defined as:

$$Precision = \frac{N_{\text{OPC}}}{N_{\text{PC}}} \tag{S11}$$

where $N_{\text{OPC}} = \{C | C \in PC, \exists C \in TC, OS > 0.5\}$ is the set of PC with $OS > 0.5$. The Figure 6.4 shows a comparison of *Precision* and *Recall* in all seven scenarios. While, the Figure 6.5 shows a comparison of *Accuracy* and *F-measure* in all seven scenarios. The overall performance of MTGO and its competing algorithms on the seven scenarios is depicted in Figure 6.2. These results, along with other computed measures, including *F-measure*, *Precision*, *Coverage*, *Positive Predictive Value*, N_{OPC}, $|PC|$, N_{OTC} and $|TC|$ are reported in Appendix A Table A. Note that the performance of GMFTP on the Human network (Fig. 6.2, Panel B) is not recorded since the algorithm did not converge after multiple attempts.

MTGO showed the best overall performance in six out of seven scenarios (best *Composite Score*, *Recall* and *MMR*). The value of *Recall* is particularly good, for example in the Human scenario, where *Recall* is doubled compared to the second best algorithm (MTGO 0.12, MCL 0.06; MTGO and MCL unveil 203 vs

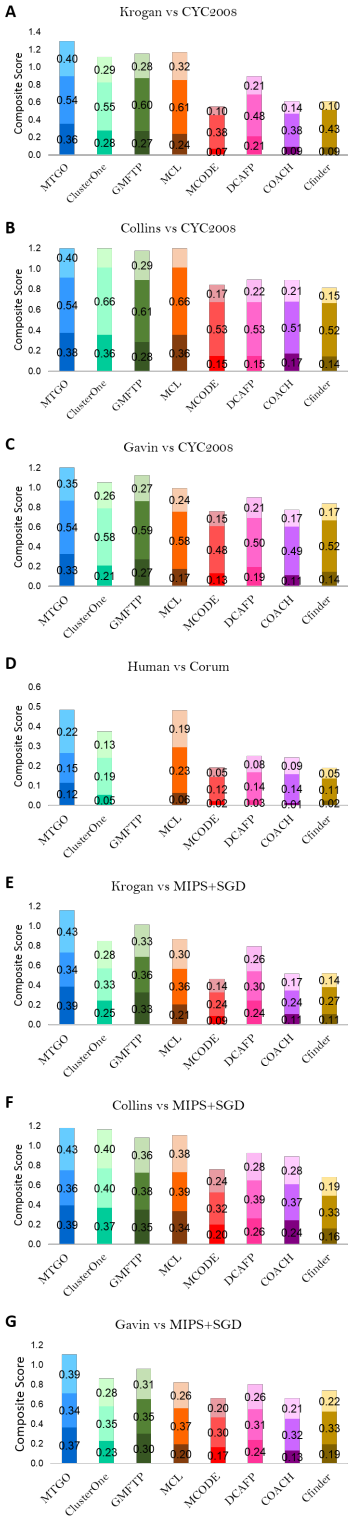


FIGURE 6.2: Composite Score of the methods over the different scenarios: MMR (light shade), Accuracy (neutral shade), and Recall (dark shade). GMFTP did not converge on the Human network.

111 modules respectively). Reaching a high *Recall* is one of the major challenges for module identification algorithms [59]. The worst performance of MTGO is on the Collins vs. CYC2008 scenario, where nonetheless it reaches the third best *Composite Score* (MTGO 1.31 vs ClusterONE 1.42). Interestingly, in the close scenario Collins vs. MIPS+SGD, where protein complexes are different, MTGO shows the best *Composite Score* (MTGO 1.18 vs ClusterONE 1.16).

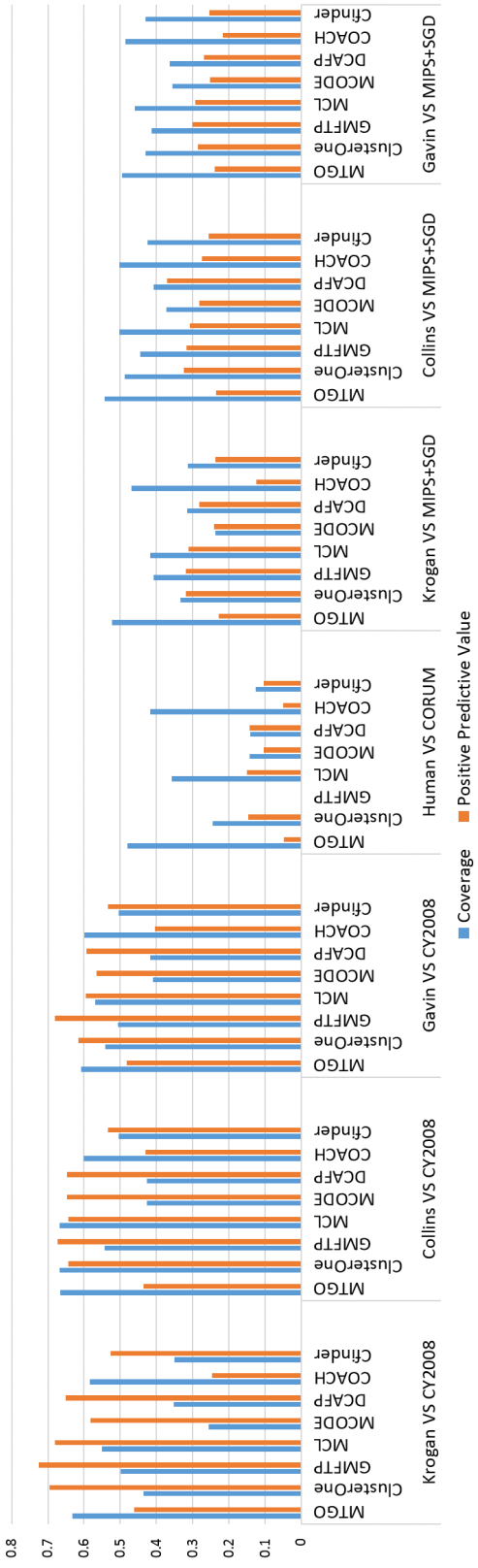


FIGURE 6.3: Coverage and PPV in all seven scenarios.

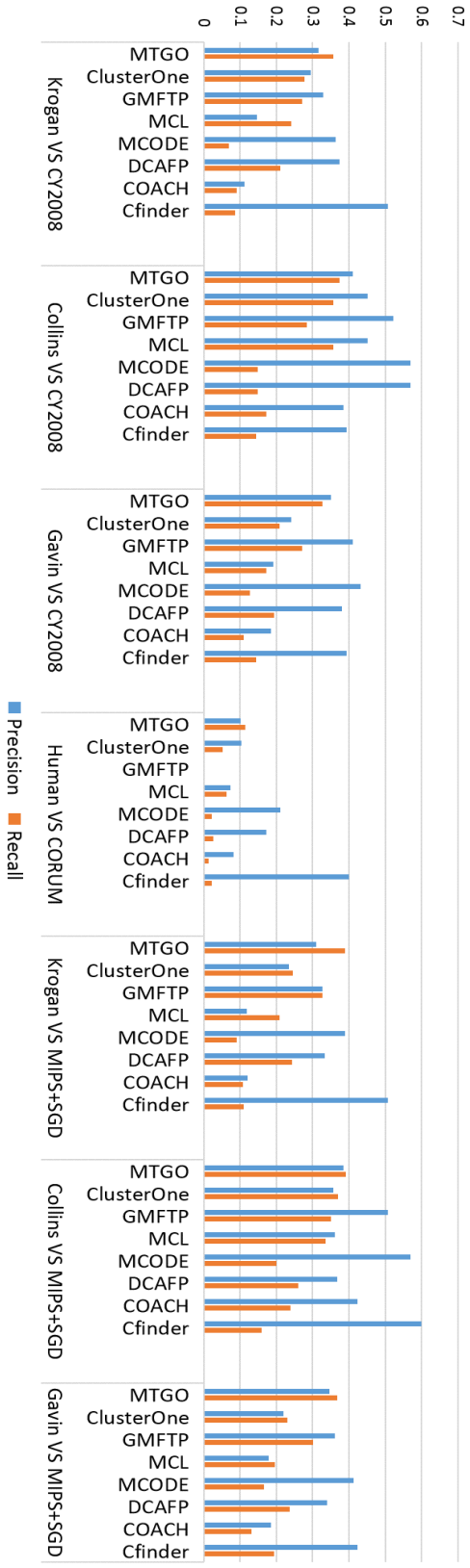


FIGURE 6.4: Precision and Recall in all seven scenarios.

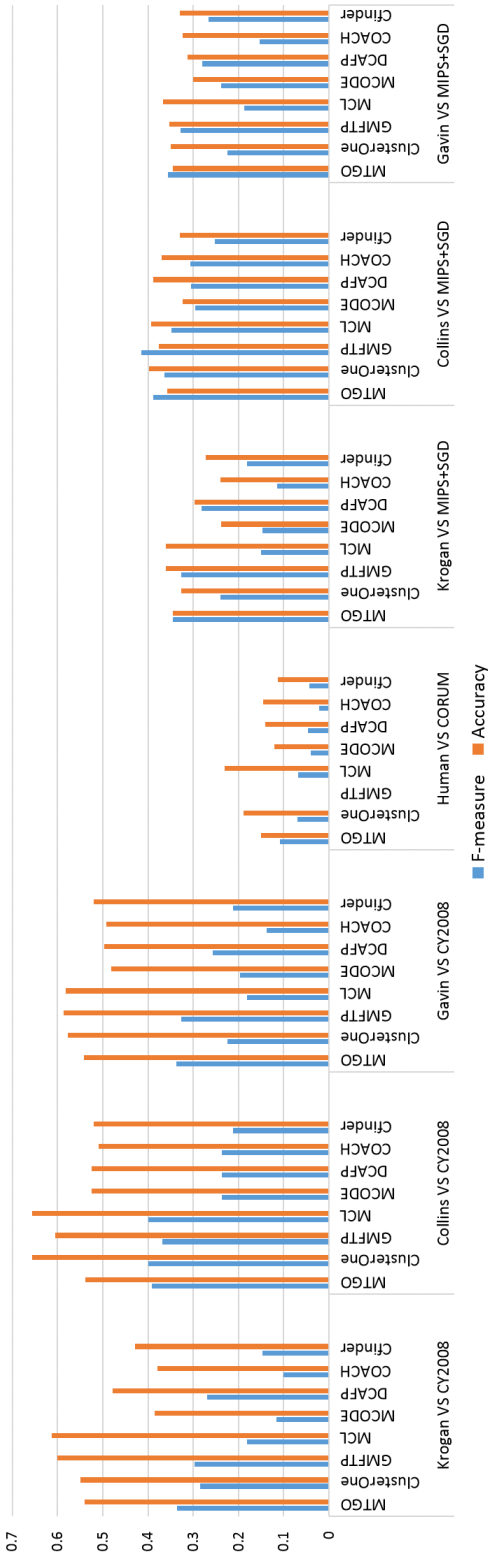


FIGURE 6.5: Accuracy and F-measure in all seven scenarios.

6.3.1 GO term analysis

In the literature, given a chosen p-value as threshold, a predicted module is defined as functionally significant if at least one GO term is significantly enriched (i.e. associated with a p-value lower than the threshold) in the module proteins [61]. For the protein complexes predicted in each network, we used GOTermFinder [65] to perform the function enrichment test with 10^{-3} and 10^{-10} p-value thresholds. We compared our results with DCAFP and GMFTP, both GO-based as MTGO. The results are reported in Figure 6.6 and in Appendix A Table B. Overall, by considering the sum of the enriched terms in all the three GO ontologies, MTGO outperforms DCAFP and GMFTP in all the networks but Collins (where DCAFP gets the best performance, consistently with the previously discussed Composite Score results). Nonetheless, MTGO outperforms DCAFP and GMFTP on the Biological Process related GOs in all the four networks (Supplementary Table S2 and Figure 6.6). Furthermore, the superiority of MTGO is clear in the Human network, where MTGO is able to retrieve a particularly high percentage of modules with at least one significant GO term. Compared to DCAFP for p-values of 10^{-3} and 10^{-10} respectively, MTGO retrieves 91% (vs 62%) and 55% (vs 42%) for Biological Process-related GO terms; 65% (vs 57%) and 27% (vs 15%) for Cellular Component-related GO terms; 81% (vs 43%) and 28% (vs 8%) for Molecular Function-related GO terms. Note that GMFTP results are not shown for the Human network as the algorithm failed to provide a viable result after multiple attempts.

To further validate our results, we measured the p-values (Fisher's exact test) of the GO terms related to each topological module of each partition. In detail, the Fisher test has been used to verify if the GO term assigned by MTGO to a cluster, i.e. as set of genes/proteins, is over-represented in the set. In order to perform a Fisher test, we need the following information:

- *Class*: the GO term we want to consider for the enrichment analysis; in this case the class is the GO term assigned to a cluster by MTGO
- *Background*: list of proteins; in this case the background is the full list of proteins present in the PPI network
- *Test set*: set of proteins on which to perform enrichment analysis; in this case the test set is the proteins set of a cluster identified by MTGO

On the base of these sets the contingency table and the corresponding p-value is computed. For p-value computation, the function *fisher.test* of R software has been used. To clarify, an example is described below. Let's suppose MTGO predicts a cluster of 3 proteins and the assigned term is GO:0000733; the *Class* is

the GO term GO:0000733; the *Background* is the set of 1622 proteins contained in PPI network, and the *test set* are the 3 proteins of the cluster. An example of contingency table is presented in table 6.2 The p-value computed is $1.183264e-$

TABLE 6.2: Contingency table

	GO:0000733	not GO:0000733
Into cluster	3	0
Not into cluster	6	1613

07; this value means that the probability of randomly selecting from the *Background* a set of 1622 proteins, containing one or more proteins associated with the GO:0000733 term is lower than $1.183264e-07$. Since the analysis involves multiple testings at the same time, in this case the number of clusters, thus multiple statistical comparisons need p-value adjustments. For this reason, the Bonferroni correction has been applied. Bonferroni correction consists in multiplying the p-value by the number of tested hypothesis, in this case the number of clusters. This correction has the limit to be very conservative, thus there is an high risk of introducing false negatives. Let's suppose the total clusters are 353, then the final p-value is $4.176e-05$. Following this procedure for all clusters, we found the great majority of GO terms (81% to 96% in all four networks) to be significant (< 0.001) and about a half (39% to 59%) to be highly significant (10^{-10}). These results are reported in Table 6.3.

TABLE 6.3: Percentage of significant attached GO terms

	10^{-3}	10^{-10}
Krogan	96%	49%
Gavin	89%	44%
Collins	81%	39%
Human	94%	59%

FIGURE 6.6: GO term enrichment. The histogram shows the percentage of predicted clusters, from each algorithm for the networks Krogan, Gavin, Collins, and Human, enriched with GO terms with a p-value under the threshold of 10^{-3} and 10^{-10}. The P, C, F labels indicate respectively the three classes of GO: Biological processes, Cellular Component, and Molecular Function.

6.3.2 Small and Sparse complexes

An open problem in module identification algorithm is the detection of small and sparse complexes. While small complexes are defined has having three nodes or less [160], there is no clear consensus about how to define sparse ones [59, 145, 160]. We defined four additional scenarios (one per network) to assess both small and sparse module detection. As regards sparse complexes, four different target sets have been created for each network, Krogan, Collins, Gavin and Human. In fact, the same target complex shows different density values according the network considered. Each target set has been created selecting from the whole target set (CYC2008 for Krogan, Collins, Gavin; and CORUM for Human) the subset of complexes with density lower than 0.5 with respect to the network considered. (For example, for the Krogan network the target set of sparse complexes is made of the CYC2008 complex subset showing a density of less than 0.5 with respect to the krogan network). As regards small complexes, two target sets were assembled by considering complexes made of three nodes or less were considered from CYC2008 and CORUM sets. Predicted complexes were compared to target sets using the overlapping score (see formula S5). Figure 6.7 shows results for small and sparse complex detection, while Table 6.4 shows the number of complexes in each target set built specifically for

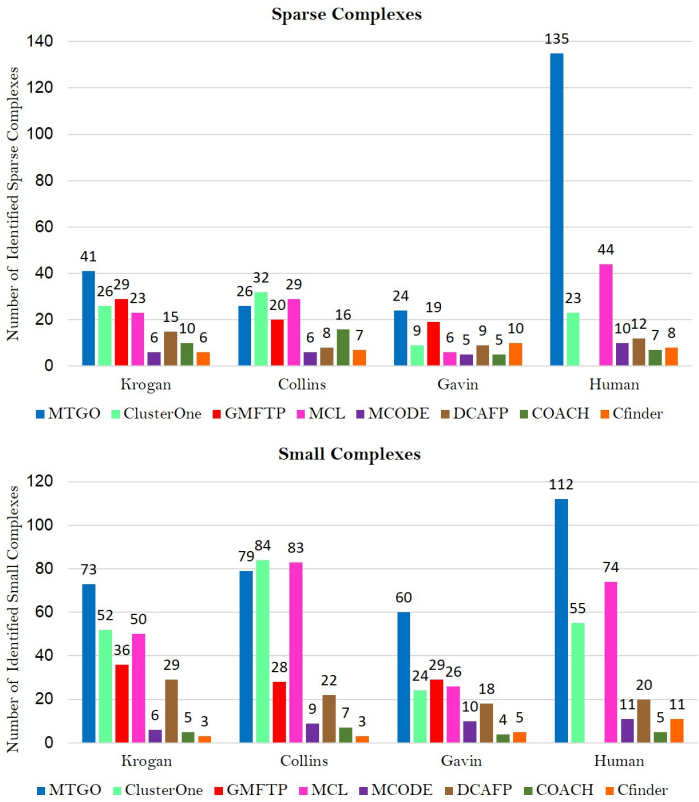


FIGURE 6.7: Small and Sparse complexes comparison. GMFTP did not converge on the Human network.

the small/sparse complexes.

TABLE 6.4: Complex number inside each target set.

	Small complexes	Sparse complexes
Krogan vs CYC2008	259	233
Gavin vs CYC2008	259	244
Collins vs CYC2008	259	217
Human vs CORUM	1046	1516

By optimizing a trade-off between GO terms and topology, MTGO is extremely accurate in unveiling small and/or sparse functional modules, often missed by other algorithms. We found MTGO performance to be consistent with the previous measure, i.e. MTGO outperforms all other algorithms in all scenarios, except in the Collins network. The performances on Human scenarios are remarkably high, especially in detecting sparse modules, MTGO correctly identifies 135 modules, while the second best MCL only 44, less than one third (Figure 6.7).

6.3.3 Run time evaluation

As regards the time, MTGO has been designed for networks of little/middle size, about five hundreds nodes. As MTGO aim is that to identify the biological pathways or cellular functions acting in a specific context (cells or tissues under pathological/physiological conditions), the target PPI networks for MTGO involve just the proteins identified in specific samples. This type of network generally collects about a few hundred of nodes; for example the PPI network obtained by *Brambilla et al.* is made of 584 nodes and 5246 edges [52], that used by *MacDonald et al.* contains 155 nodes [44] and that used by *Wu et al.* contains 37 nodes and 322 interactions [43].

For MTGO validation, in order to follow standard approaches [60, 61, 153], the networks used (Gavin, Krogan, Collins and Human network described in previous paragraphs) are very big. In fact, these networks are *interactome*, i.e. networks collecting all the interactions of an entire biological system, such as an organism. For this reason, MTGO in the four analyzed networks results very slow. For example, Table 6.5 shows the run time of all considered algorithms for the Gavin network, that is made of 1856 nodes and 7669 edges. It is possible to observe that the algorithms are very fast (more or less some seconds), except for those including GO information. Evidently, the inclusion of further information lengthens the time of processing of the input data. Moreover, GMFTP and MTGO are the slowest, this can be due to the fact that the two algorithms are both based on iterative processes. In details, for Gavin network GMFTP executed 100 iterations (according default parameters), while MTGO executed 355 iterations.

TABLE 6.5: Time for Gavin Network

	Gavin Network (sec)
MTGO	10237
ClusterOne	<1
GMFTP	51371
MCL	<1
MCODE	3
DCAFP	233
COACH	4
CFinder	2

Table 6.6 shows MTGO computation time for networks with different number of nodes and edges. In Leukemia and Myocardal networks (described in Chapter 4), MTGO shows a very short processing time. In Test Network, MTGO

TABLE 6.6: MTGO run time

	Nodes	Edges	Time (sec)
Leukemia Network	78	545	1
Test Network	446	24567	120
Myocardial Network	502	4316	30
Krogan Network	2709	7123	54000

shows a relatively short time, considering that the network has a very high number of edges (24567). Finally, in Krogan network MTGO is very slow. These results underline MTGO run time is robust against the increase of the edges, but not to the increase of nodes.

As the main part of this PhD research has been devoted to development and validation of the novel algorithm, and since MTGO shows a good run time for its PPI target networks (middle size), the time aspect has been relatively little considered. A deeper analysis about time should be executed, in order to evaluate the limits of the algorithm and to improve the processing time where possible. For example, a possible way to reduce the run time is using as initial configuration not a random partition but a computed ones, which takes into account the GO. These investigations will be the main objective of the future developments on MTGO algorithm.

Chapter 7

Stability analysis

7.1 Introduction

One of the most popular approaches for PPI Network analysis is module detection, a challenging task faced by many algorithms [17, 36]. A PPI network module (also called cluster) is represented by a group of proteins with a specific biological role, i.e. they work together to perform a specific cellular process, along with particular topological features, i.e. they share lot of connections [145]. Because of the complexity of biological systems, there are likely to be many valid clustering results, each revealing some aspect of underlying biological behavior. Therefore, clusters must be evaluated both for biological relevance and stability. Understanding and accounting for the stability of the clusters with respect to the presence of noise and uncertainty in the data is an important factor when evaluating an algorithm specific for PPI Networks [194]. In fact, the input graphs are obtained from high-throughput methods (e.g. yeast-two-hybrid, etc..) for detecting pairwise protein-protein interactions (PPIs), which are generally noisy with high false positive and false negative rates [177]. Moreover many module detection algorithms rely on a random component, thus stability of the results across different runs is considered to be an asset of the algorithm [195]. As MTGO is a non-deterministic algorithm, to evaluate its performance two different stability analyses have been executed. Firstly, one to evaluate the stability of the result over many runs starting from a same input, to consider the range of variability introduced by the random components of the algorithm; secondly, one to evaluate the robustness of the output clusters when the input is affected by noise and uncertainty. For both analyses, module biological relevance has been taken into account. The following work has been carried out in the Department of Computer Science at Brunel Univesity in London in collaboration with the Senior Lecturer Allan Tucker. Part of these results has been presented for the workshops NETTAB 2017 *Methods, tools & platforms for Personalized Medicine in the Big Data Era* and have been published in the abstract

Stability analysis of MTopGO for module identification in PPI networks by following authors Danila Vella, Allan Tucker and Riccardo Bellazzi [196].

7.2 MTGO stability analysis

For the first analysis a PPI Network is used from human proteins [177] with 502 proteins and 4316 edges. MTGO has been run 500 times over the same input network, using the same parameters (default) and the same GO list (7909 in total, including MF, CC and BP classes). The 500 clusterings obtained have been analyzed to evaluate their consistency. Each clustering result is made of a number of clusters of size equal to 77.428 ± 10.106. In Figure 7.1 the values of Q against QGO are shown for all 500 clusterings: it can be observed that the QGO values are more stable than Q values. In fact, the 94.8% of QGO values show little variability, ranging from 0.92 to 0.97. While, the 95.2% Q values show higher variability, ranging from 0.18 to 0.24 (Figure 7.2). It is interesting to note that about 95% of the total results can be identified in two main zones of high density, the first one on the left is linked to lower Q values, *Zone 1*, and the second one on the right is linked to higher Q values, *Zone 2*, suggesting that the *Zone 1* correspond to a local Q maximum, while the *Zone 2* corresponds to higher Q maximum (see Figure 7.3). As regards the remaining 5% of the results, a third zone (*Zone 3*) of lower density can be observed. It is represented by points set along a diagonal line, with $QGO < 0.92$. *Zone 3* suggests a negative correlation between Q and QGO (see Figure 7.3). In fact, the Pearson Correlation value between Q and QGO for the *Zone 3* points is -0.78. In Figure 7.4 the Q values of the final partitions are compared with the Q values of the initial random partitions, the graph shows there is a random relation, implying that the final results are little influenced by the random initialization of the algorithm. To evaluate the agreement among all 500 clusterings the Weighted Kappa (WK) metric has been used [197]. This metric allows one to compare two partitions in terms of cluster similarity, producing a score ranging from -1 (no concordance) to +1 (total concordance). The WK score has been computed among each pair of clusterings, for a total of 124750 comparisons (the values are shown in Figure 7.5). The obtained values are in the range 0.67 ± 0.12.

The next step of the analysis was the application of Robust Clustering (RC) and Consensus Clustering (CC) algorithm [198]. Robust Clustering is used to compute robust clusters, i.e. groups of proteins allocated to the same cluster in all of the repeated 500 clusterings. These can be construed as the most stable clusters with respect to the randomness within the MTGO algorithm. As a result, 28 robust clusters have been found, covering 85/502 proteins. The mean

and standard deviation of robust cluster size are 3.03 ± 2.16. The robust clustering algorithm is useful for creating clusters with high confidence, however, robust clustering can be too restrictive and discarding many proteins that do not have full agreement, this explains because robust clusters cover just 85/502 proteins. Consensus clustering overcomes this problem, requiring a minimum-agreement parameter to generate clusters based on the combined results of all clusterings. Thus, the Consensus Clustering allows one to compute a new clustering over all proteins by maximizing agreement over all the 500 repeated clusterings. CC uses a form of simulated annealing and therefore requires two key parameters. The starting temperature was set to 100 and the number of iterations to 1000000. The CC Q value falls in the middle of the *Zone 1* (see Figure 7.6). To compare the CC with all 500 clusterings the WK values have been computed between the CC and each clusterings, the obtained values are shown in Figure 7.7. It can be observed that the WK values are always positive, however there are some clusterings with low agreement with CC. To investigate the characteristics of these clusterings, those with WK value less than a selected threshold (0.6) have been highlighted in Figure 7.8. These results are concentrated in the *Zone 2* and *Zone 3*, it is interesting that the clusterings with low CC agreement are included in the lowest density area, *Zone 3*. In fact, the points in the *Zone 3* can be considered the less representative clusterings. As a result, the generation of CC can be used to filter out the less stable clusterings to ensure results that are closer to the more preferred *Zone 1* and Zone2. To improve the CC, the less stable clustering set (WK<0.6) has been removed and the remaining clusterings have been used to recompute a new CC. Peculiarly, this resulted in a decreasing of the value of Q (see Figure 7.9), this may be due to the removal of the *Zone 3* points, which have very low QGO but with the most high values of Q. In the light of these results, as MTGO can converge to a different local optimum, multiple runs followed by the CC application are suggested to the end-user to obtain superior quality predicted complexes. Obviously, this task is hindered by the necessary run time; for this reason, multiple runs are recommend just in case of small networks, such as involving about 500 nodes. Selecting as cut-off for *Zone 2*, the Q values ranging in $[0.21, 0.24]$ and the QGO values higher than 0.92, the percentage of clusterings falling down in *Zone 2* (i.e. global optimum) are 46%. Thus, for large networks, the probability to find a global optimum with a single execution can be estimated as 46%.

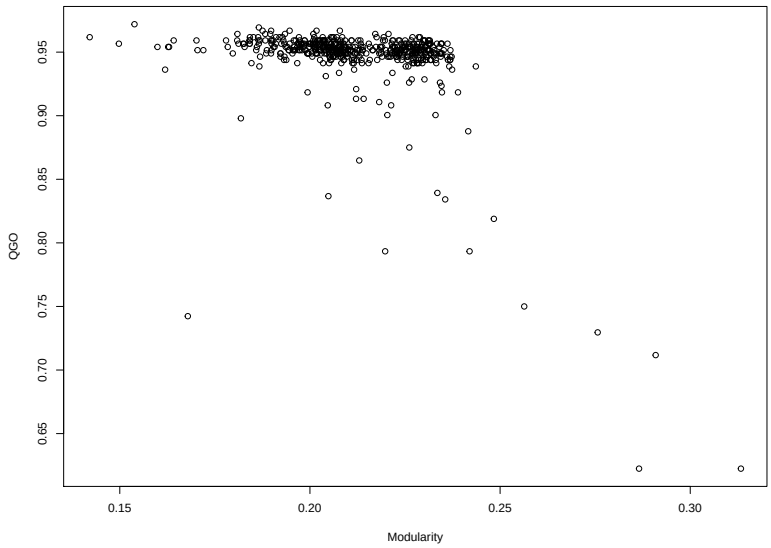


FIGURE 7.1: The values of Q and QGO for all 500 clusterings.

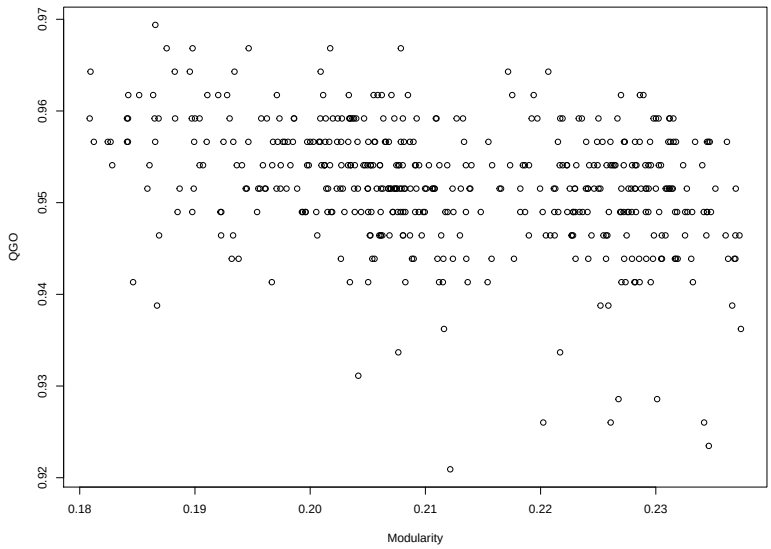


FIGURE 7.2: The graph shows the value area where about the 95% of Q and QGO values fall.

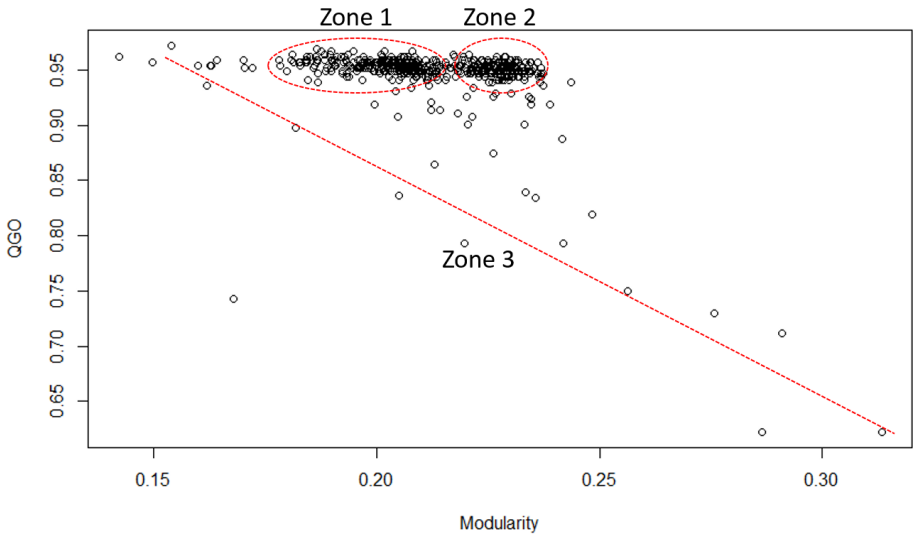


FIGURE 7.3: The graph shows the Q and QGO values for all the 500 clusterings. Three different zones of points can be individuated, underlined by red circles (*Zone 1* and *Zone 2*) and the red line (*Zone 3*).

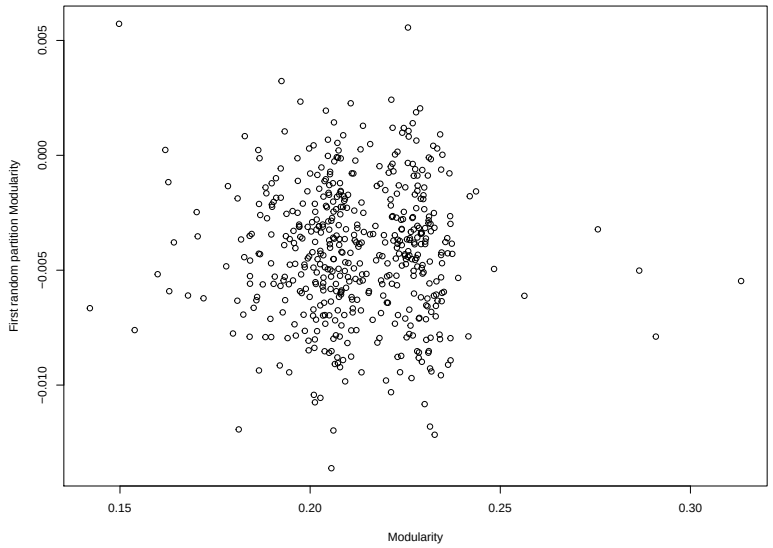


FIGURE 7.4: The Q of first random partition against Q of the final partition, for all 500 clusterings.

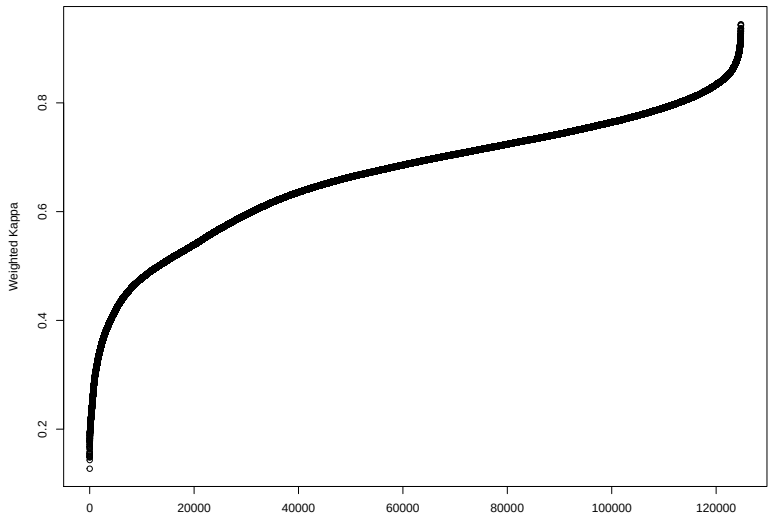


FIGURE 7.5: The graph shows the WK values obtained for each pair of clusterings. The x-axis indicates all the couple of clustering obtained,in total 125000

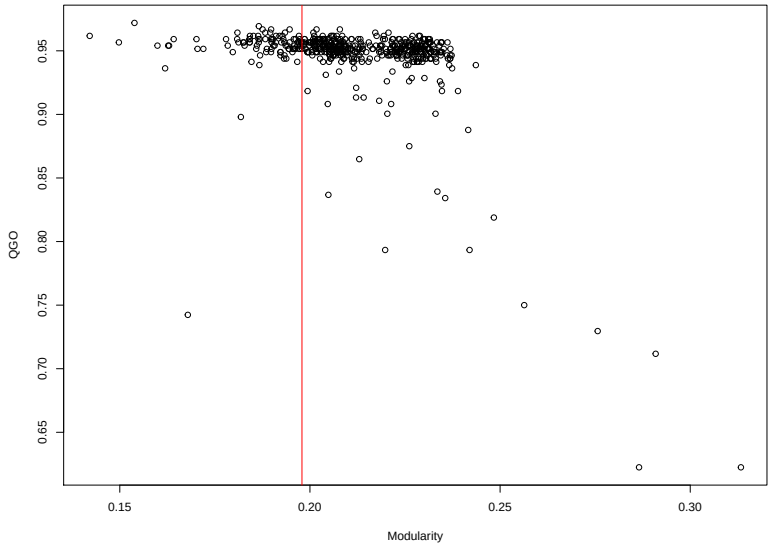


FIGURE 7.6: The graph shows the value of Q and QGO for all 500 clusterings, the red line indicates the Q value for the Consensus Clustering.

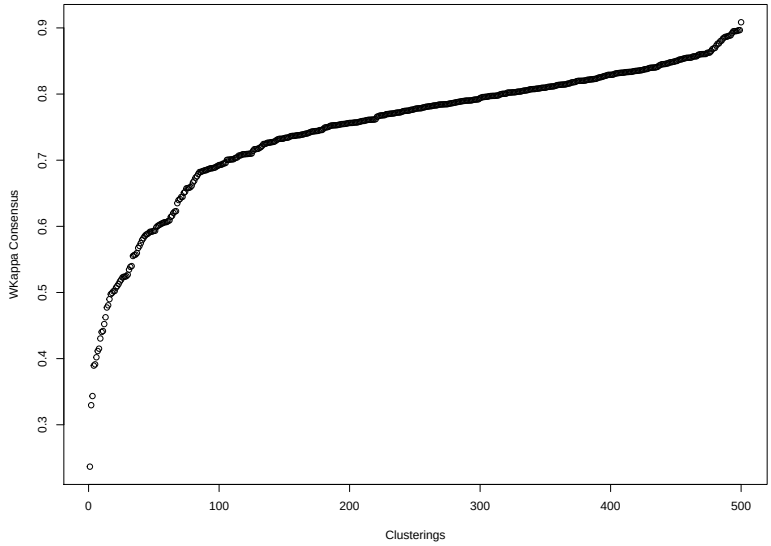


FIGURE 7.7: The graph shows the value WK obtained from the comparison between CC and all 500 clusterings.

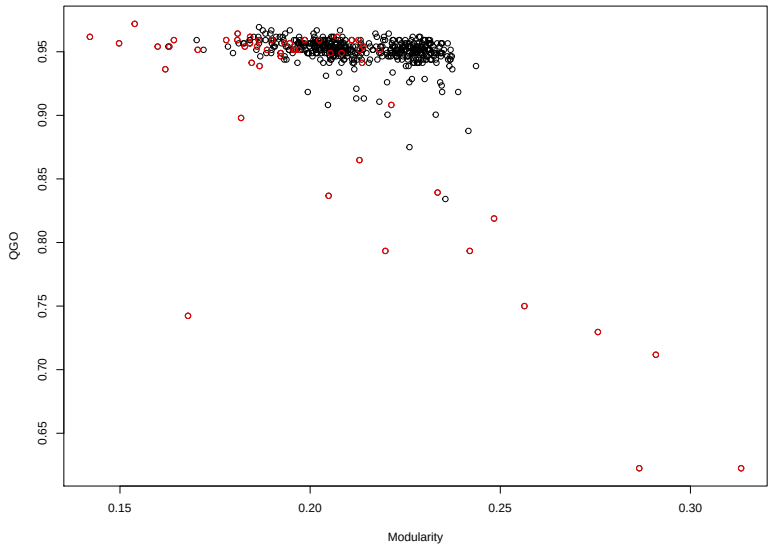


FIGURE 7.8: The red points represent the clusterings with a WK value less than 0.6 respect to the CC.

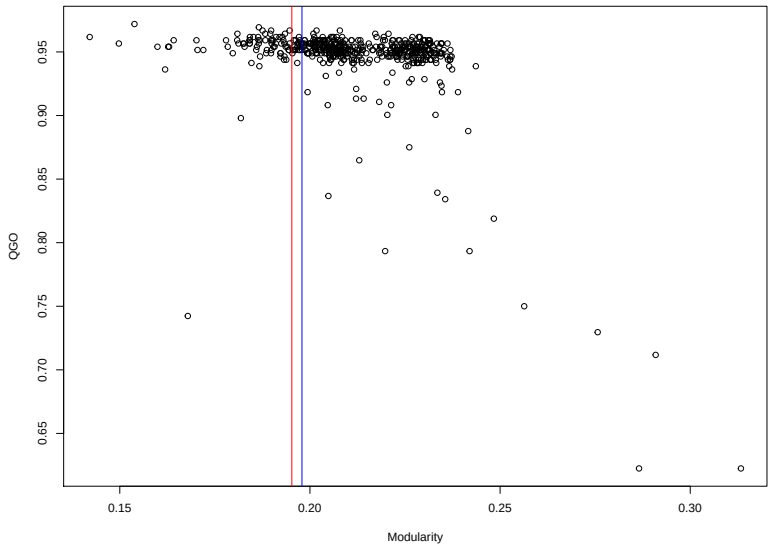


FIGURE 7.9: In the graph the blue and red lines show the Q value computed for two CCs, respectively the CC obtained considering all 500 clusterings and the CC obtained after the removal of low agreement clusterings (WK<0.6).

7.2.1 Gene Ontology stability

As regards the biological aspect, to evaluate stability, the GO terms assigned by MTGO to each cluster have been considered. To evaluate the GO terms assigned, a probability score has been used to test the significance of observing multiple proteins, belonging to a GO term assigned, in the corresponding cluster against the null hypothesis of this happening by chance [198]. The probability score was computed using the normal approximation:

$$z = \frac{x - \mu}{\sigma}; \mu = kp; \sigma = \sqrt{kpq} \tag{7.1}$$

where $p = s/n, q = 1 - p$. In this formula, s is the cluster size, n is the total number of proteins in the PPI network, k is the GO term size, lastly x is the number of proteins shared by both the GO term and the cluster. z takes high values when the probability of observing x proteins, belonging to the assigned GO term, in the cluster by chance is very small.

For all 500 clusterings and for each assigned GO the p-value was computed according Equation 1.1. A GO term is considered significant if the p- value is under the threshold 0.05. For each clustering the number of insignificant GO terms has been computed (see Figure 7.10), most of the clusterings show a percentage

of insignificant GO terms ranging from 10% to 15%. To investigate the characteristics of the clusterings with a high percentage of insignificant GO (>15%), these results have been highlighted in the main graph Q-QGO (see Figure 7.11). It is interesting to observe that these results fall mostly in *Zone 2* and *Zone 3*, the zones with the highest Q values. Moreover, for each clustering the GO p-value mean has been computed, and results are shown in Figure 7.12. Next, the results with poor mean values (mean>0.09) have been highlighted in the main graph Q-QGO (see Figure 7.13); according to previous results the corresponding points are concentrated in (*Zone 2* and Zone 3), suggesting that results with high Q values are linked to lower GO quality. It suggests that the increasing of the topological quality of a clustering (measured by Q) leads to a declining of the biological quality of the clusters. Thus, the zones with higher value of Q are linked with bad quality GO clusters (*Zone 2* and *Zone 3*), while the zones with low Q values are linked to cluster with a richer biological meaning (*Zone 1*). One of the challenge linked to clustering algorithms for PPI networks is the ability to detect small complexes [59]. This is particularly true for algorithms based on modularity optimization[64]; in fact, because of resolution limit [156] the maximum modularity partitions tend towards collecting small clusters in larger communities, leading the small module loss. MTGO faces this issue driving the cluster building process not only by topological properties but also relying on GO knowledge. To investigate this ability, for each GO term linked to a cluster the functional module size has been compared to the extraction frequency, for all 500 clusterings. In Figure 7.14 the MTGO ability to detect both big and small functional modules can be observed. In fact, the high frequency points are linked both with small and big GO sizes. Moreover, there are a lot of points concentrated in the zone with $size < 20$, confirming the ability of MTGO to detect small functional modules. It is interesting that very big GO ($size > 50$) are linked with very high frequency, this can be explained by modularity resolution limit. According to this limit a large community combining some small sub-modules maximizes the modularity much more than the same small modules separated from each other [199]. A big size GO certainly collect into itself same smaller GO groups, because of the hierarchical GO structure, this configuration has a strong biological relevance alongside a good power to increase the global network modularity, for these reason the GO with these properties are detected in almost all the 500 clustering.

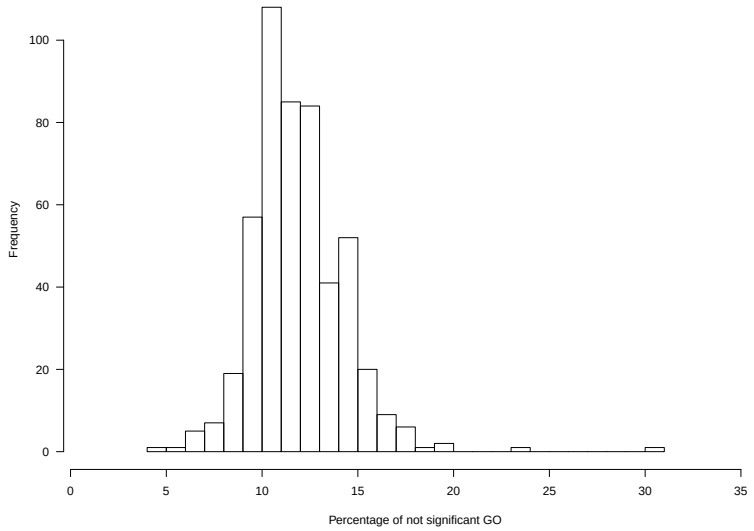


FIGURE 7.10: The histogram shows the frequency of the not significant GO percentage in a clustering

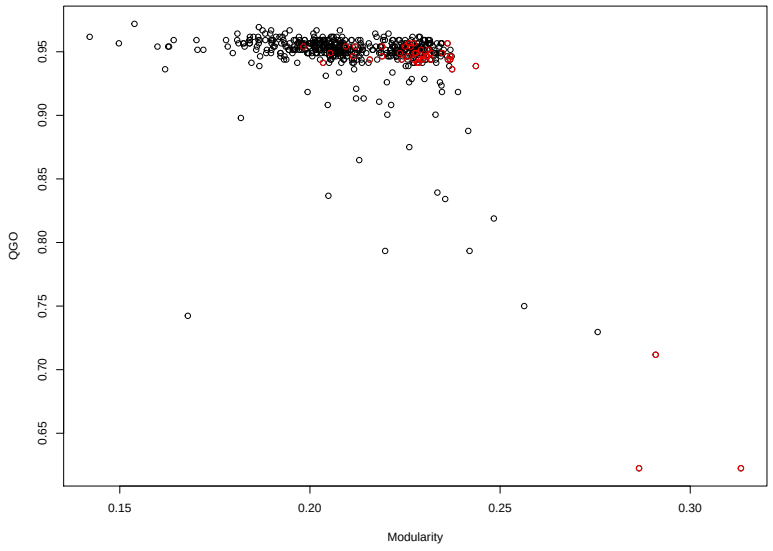


FIGURE 7.11: In this graph the red points represent the clusterings with a percentage of not significant GO $> 15\%$.

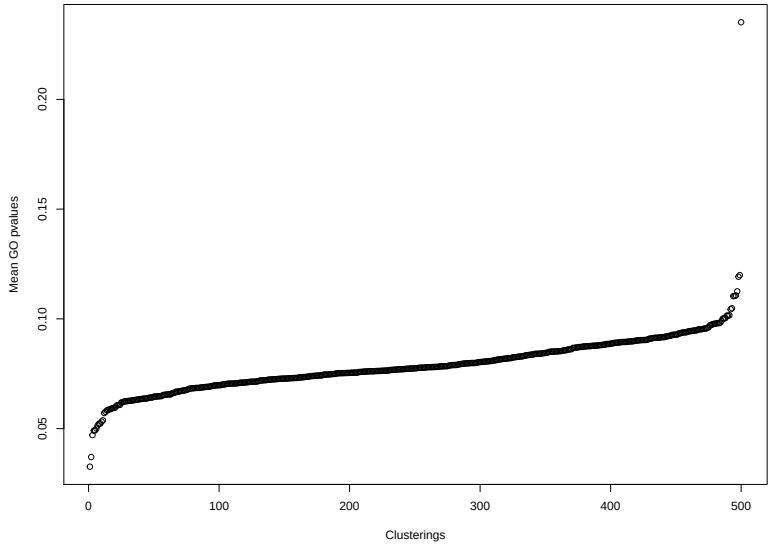


FIGURE 7.12: This graph shows the GO p-value means over the clusters for all 500 clusterings.

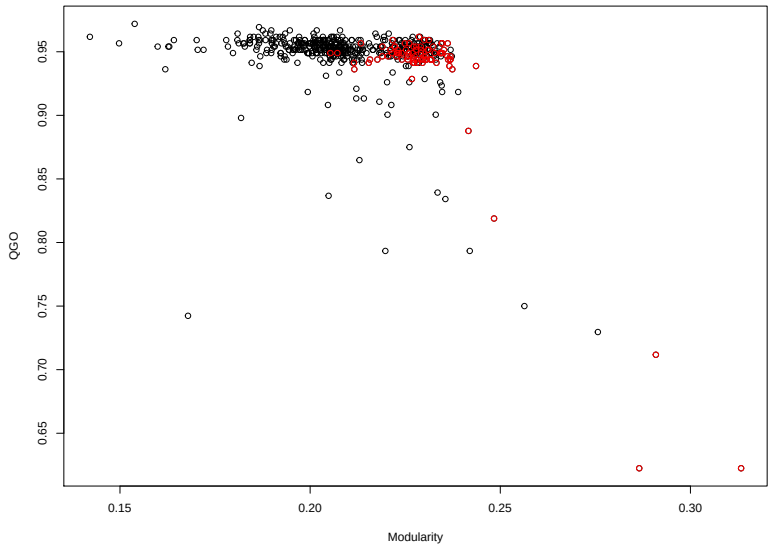


FIGURE 7.13: In this graph the red points represent the clusterings with a GO p-value mean > 0.09.

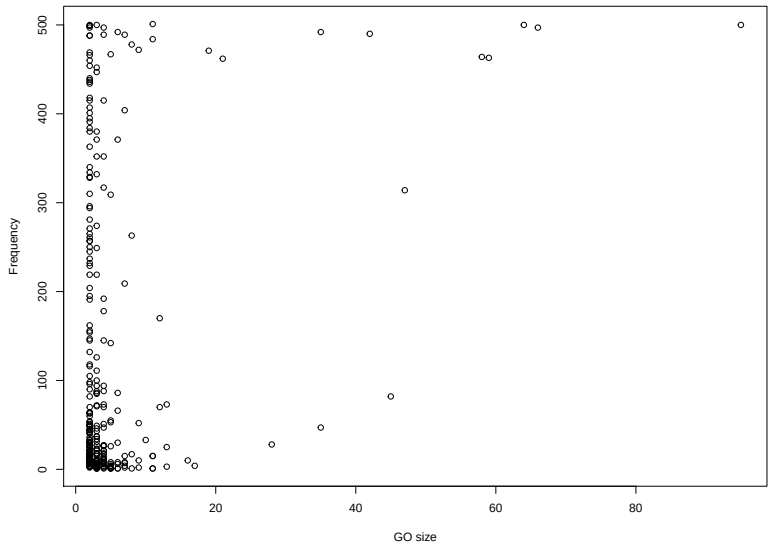


FIGURE 7.14: In this graph each point represents a GO term linked to a topological module by MTGO. The x axes indicates its size, i.e. the number of proteins/nodes linked to it. While the y axes indicates its frequency, i.e. the times a GO term appears over the 500 clusterings.

7.3 Stability analysis for perturbed networks

The PPI Networks can often be affected by false positive and false negative edges. In fact, they are often produced with the results of high-throughput experimental methods and many computational methods, able to quickly detect a large amount of novel PPIs but with relatively low quality [200]. For example, the PPIs are influenced by the experimental conditions, thus the experimental methods may detect PPIs that do not occur under physiological conditions, resulting in high false positive detection rates [201]. Moreover, the high-throughput methods may fail to detect many interactions because of their complex biological nature, for example the interactions may be transient or too weak [202], this results in false negative edges and low coverage of the interactomes. For these reasons, it is important to investi- gate the robustness of MTGO on PPI networks affected by these types of error. This is now explored by perturbing existing PPI networks and investigating the impact on MTGO. For this analysis the human network has been used to create eleven networks with different levels of alteration, keeping the same node set (502 proteins) but changing the edge set (4316 interactions). To simulate the missing and false positive edges the eleven networks have been produced randomly removing and adding edges to

different degrees. In particular, combining three different percentages (0% - 5% - 10%) of random edge removal and addition, eight different networks have been obtained.

TABLE 7.1: The table shows the edge number of each perturbed network obtained combining different percentages of random edge removal (columns) and random edge addition (rows)

% removal % addition	0	5	10
0	4316	4101	3885
5	4531	4316	4100
10	4316	4532	4316

Two additional networks have been obtained to explore more dramatic perturbations by removing and adding 50% of the edges (2159 edges and 6474 edges respectively). Finally, the last network has been obtained from the original node set adding only random edges in equal number as the original network (100% removal and 100% addition). Each network has been processed by MTGO one hundred times. Q and QGO values obtained for the 100 run for each perturbed network are shown in Figures 7.15 and 7.16. In the Figures the networks are identified through labels, nA-mR which represents the network that has been obtained by adding n% of edges and removing m% of edges. The Random label means the network is random, obtained randomly removing and adding the 100% of edges. In the Q boxplot, it can be observed in the first eight networks the modularity is quite preserved, in fact these networks are the most similar to the original. The Q distributions clearly show the MTGO algorithm is more robust against the edge removal than the edge addition, in fact a Q decreasing can be observed when the edge addiction percentage increases while the edge removal percentage is constant. Otherwise, the Q distributions for the two networks obtained through a 50% edge addition and a removal, show very similar results. It may be due to a high percentage of error introduced, that mostly destroys the modular structure of the networks so decreasing the Q values. As expected, the random network shows the lower Q values as the modular structure is completely damaged.

As regards the QGO distributions for all networks a similar trend, characterized by very high values, can be observed. To explain these results, the meaning of QGO needs to be clarified. MTGO provides as a result two paired sets, the topological modules and the functional modules. The topological modules are the clusters obtained from a network partition. While,a functional module is a group of proteins, belonging to the network, linked to a GO term. The elements

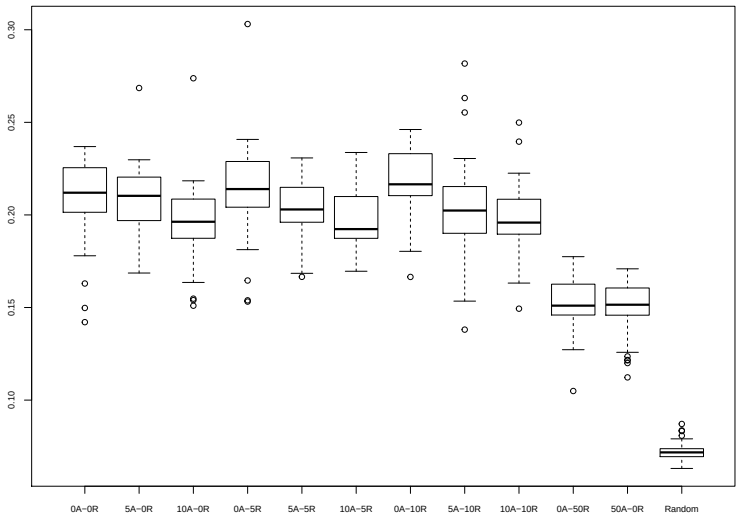


FIGURE 7.15: The boxplot shows the Q distributions computed for the 100 run for each network. The labels in x-axis identify the perturbed networks, each label *nA-mR* means the network has been obtained adding $n\%$ of edges and removing $m\%$ of edges.

of the two sets are coupled, i.e. each topological module is linked to a functional module. The QGO metric measures the agreement between each topological module, the cluster, and its corresponding functional module, i.e the GO term assigned by MTGO. While a cluster is a sub-network, with own specific topological structure, a functional module is just a node set, without any topological feature. For this reason QGO evaluates just the node/protein intersection between the clusters and the functional modules. Since the input list of GO terms is made of many redundant and overlapped GO-linked protein groups, whatever network node subset can be covered by many GO protein groups, independently by the topological features of the sub-network linked to the node set. For this reason MTGO is able to find a good set of GO terms (high QGO) for a PPI network even if this network is completely "loose" in terms of its modular structure.

To compare the clusterings obtained for each perturbed network the CC algorithm has been applied over the 100 runs to find a single clustering synthesizing the MTGO behavior. The agreement between each perturbed network CC and the original network CC has been evaluated through the WK. The results are shown in Figure 7.17. The WK values show a very similar trend to Q value distributions. In fact, they show a fast decrease respect to the increment of the edge addition percentage. While, the WK values for the two highly perturbed

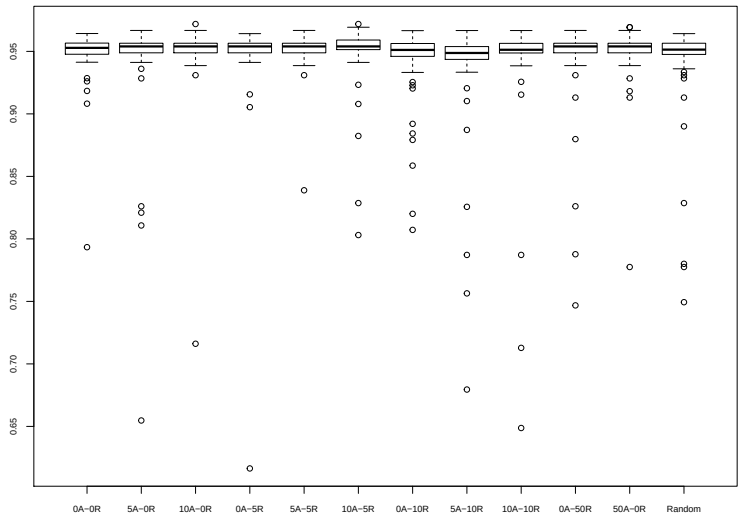


FIGURE 7.16: The boxplot shows the QGO distributions computed for the 100 run for each network. The labels in x-axis identify the perturbed networks, each label *nA-mR* means the network has been obtained adding $n\%$ of edges and removing $m\%$ of edges.

network 0A-50R and 50A-0R are almost the same. Finally, as expected, the random network shows the lowest WK values.

To evaluate the robustness of the clusters in presence of errors, the RC algorithm has been applied to find for each network a set of RCs representative of the 100 runs. The RCs derived from the perturbed networks have been compared with those from the original network using the measure Maximum Matching Ratio [153]. The result is shown in Figure 7.18, showing a similar trend as Q distribution (see Figure 1.14). In fact, the MMR decreases along with the increasing of random edge addition, confirming that MTGO is more sensitive to edge addition than edge removal. This can be due to modularity function limits; Fortunato *et al.* studying the modularity limits find that both for a random and scale-free graph the expected maximum modularity increase when the graphs gets sparser, i.e. the edge number decrease [156]. As the PPI networks often shown scale-free properties [36], this could be the reason behind the fact that the networks 0A-5R and 0A-10R , those with higher number of removed edges, show the maximum values of modularity. Moreover Fortunato *et al.* assert that modularity is extremely sensitive to even individual connections, thus the edge addition can change a lot final results [156]. In detail, modularity tends to join together two small clusters if there is even a single connection between them.

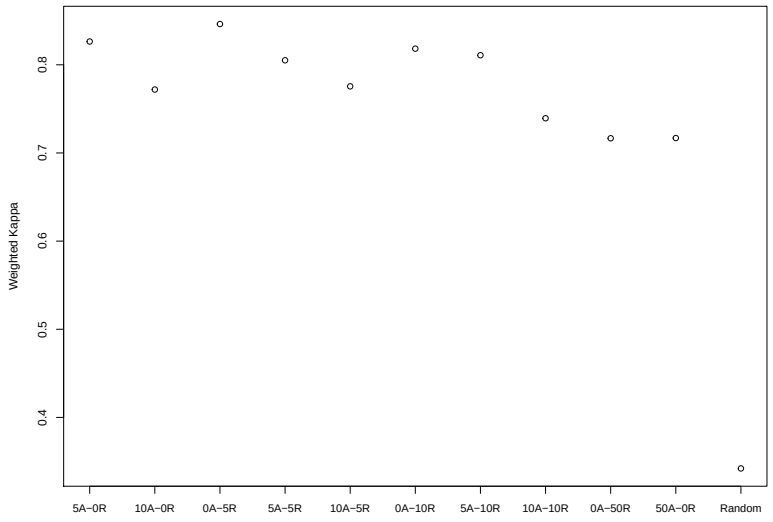


FIGURE 7.17: The figure shows the WK values computed between the original network CC and each perturbed network CC. The labels in x-axis identify the perturbed networks, each label *nA-mR* means the network has been obtained adding $n\%$ of edges and removing $m\%$ of edges.

If the connection is a false positive edge, this could bring to set nodes in wrong clusters. This could explain as MTGO is more sensitive to edge addition (i.e. false positive edges) than edge removal (i.e. false negative edges). PPI networks are affected by both false positive and false negative edges, thus to face the low robustness of MTGO against the presence of false positive edges could be useful to pre-process the network to remove them. Up to now, there are some valuable technique for edge reliability assessment, a review of the state-of-art approaches have been published by [200].

7.3.1 Gene Ontology stability

For the MTGO algorithm, an important aspect to evaluate is the ability to detect the most represented GO terms in a PPI Network, i.e. the ability to select from the whole GO list of an organism the subset of GO term linked to protein groups represented in the network as highly connected nodes. For this reason, the GO set computed for the original and perturbed networks over the 100 run have been compared. To investigate how similar are the GO set, the root mean square error (RMSE) between the GO frequency distribution has been computed for each perturbed network and the original one, the results are showed in Figure 7.19. The GO frequency distribution foe each perturbed network is computed

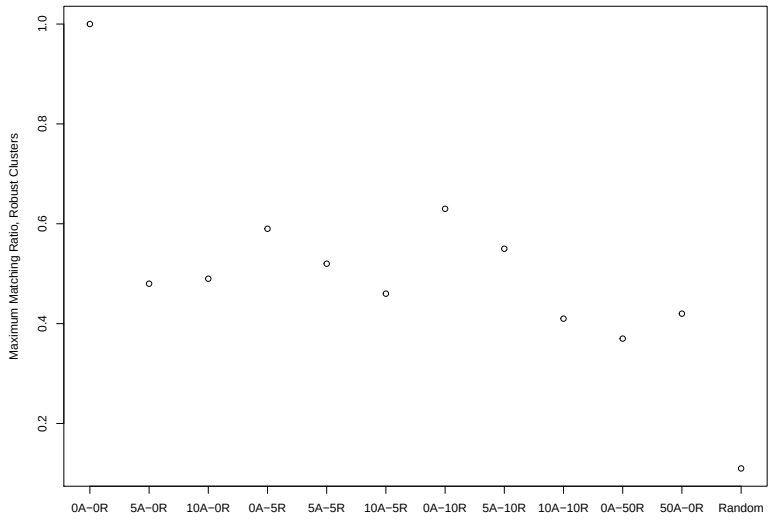


FIGURE 7.18: The figure shows the MMR values computed between the perturbed network RCs and original network RC. The labels in x-axis identify the perturbed networks, each label *nA-mR* means the network has been obtained adding $n\%$ of edges and removing $m\%$ of edges.

considering just the GO terms contained in the GO term set belonging to the original network. A slight RMSE increase can be observed as the network alteration level increase; however, it seems there isn't a different behaviour in case of edge removal and edge addition. Finally, as expected the highest error value correspond to the random network. Figure 7.20 confirms these results, it shows a good agreement between the GO frequency distributions of the lowest RMSE network (0A-5R) and the original one (green points against red points respectively); while it shows a bad agreement between the GO frequency distribution from random and original network (blue points against red points respectively).

To investigate how the GO significance is influenced by the presence of PPI Net- work errors, a probability score was computed according equation 7.1 to measure the significance of each GO term assigned by MTGO to a cluster. For each network a distribution of p-value mean values has been computed to obtain a p-value distribution representative for the network (see Figure 7.1), each value of the distribution corresponds to the mean of the p-values obtained with equation 7.1 for each pair cluster-GO in a clustering. The values follow an almost random trend, it seems there isn't a specific dependence from the edge addition and removal. However, it is interesting to view that all of the distributions are quite similar and the distribution with the lowest p-value means corresponds

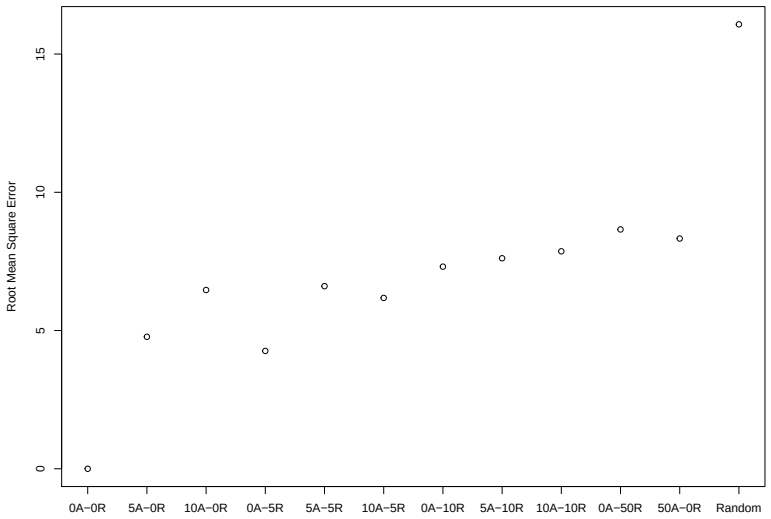


FIGURE 7.19: The figure shows the RMSE values computed between GO frequency from each perturbed network and the original one. The labels in x-axis identify the perturbed networks, each label *nA-mR* means the network has been obtained adding $n\%$ of edges and removing $m\%$ of edges.

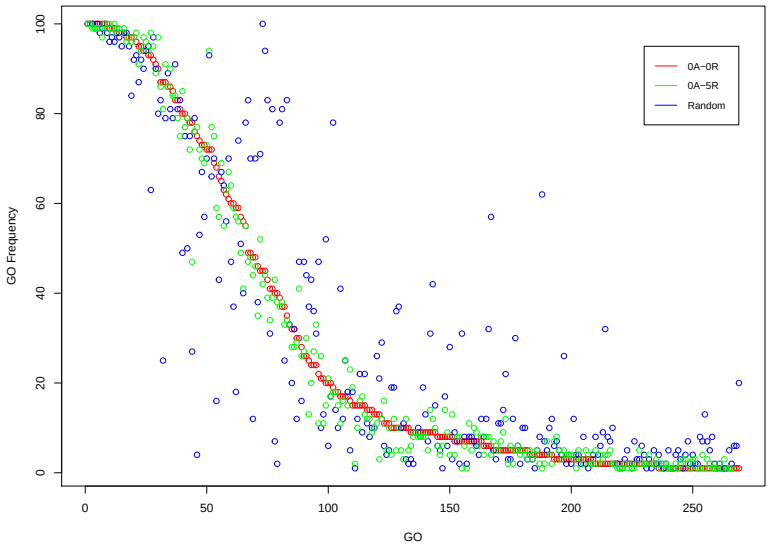


FIGURE 7.20: The figure shows the GO frequency distribution of a low error affected network (green points) and the random network (blue points) against the original one (red points)

to the random network. A similar result has been obtained computing for each network the distribution of percentage values of "GO-insignificant" clusters. A cluster is considered "GO-insignificant" if the p-value computed with equation 7.1 is bigger than threshold 0.05 (see Figure 7.22). The distribution seems to be very similar and also in this case the lowest value distribution is for the random network, confirming previous results. Moreover, this result agrees with that obtained for the QGO distributions (see Figure 7.16). These good results for the perturbed and random networks could be explained with the decreasing of network modular structure and with the GO term nature. In fact, the GO term set is redundant and the network nodes are covered by many GO terms. Because of this, independent from the network structure, it is always possible to find a subset of GO terms assuring a good overlapping between the network node groups and the GO-linked protein groups. MTGO aims to find clusters both with high topological quality and biological quality, meaning the clusters should assemble nodes that are both highly connected and highly overlapped to the proteins linked to a specific GO. To reach this scope MTGO makes a trade-off trying to build a cluster that is as overlapped as possible to the GO protein group and at the same time with its own high topological property. Thus, in a modular structure network MTGO is able to detect highly connected clusters and the search of the best GO term is limited in order to preserve the cluster topological nature, in the extreme case this could lead to GOs with low significance but at the same time cluster of high topological properties. While, in a network with scarce modular structure, the clusters lack high self-topological properties and therefore MTGO is free to search for the best representative GO for each cluster without any restrictions, this leads to cluster linked with high significant GO but low topological features.

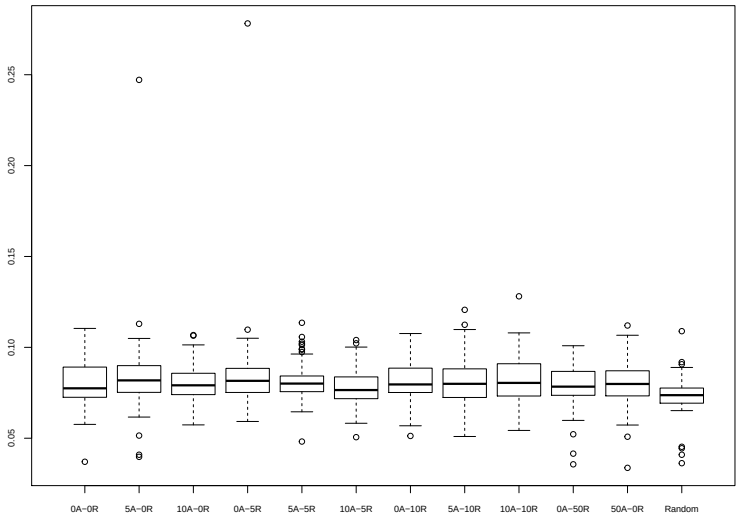


FIGURE 7.21: The figure shows for each network the distribution (obtained with the 100 MTGO run) of the mean values of the p-values obtained with equation 7.1 for each pair cluster-GO in a clustering. The labels in x-axis identify the perturbed networks, each label *nA-mR* means the network has been obtained adding $n\%$ of edges and removing $m\%$ of edges.

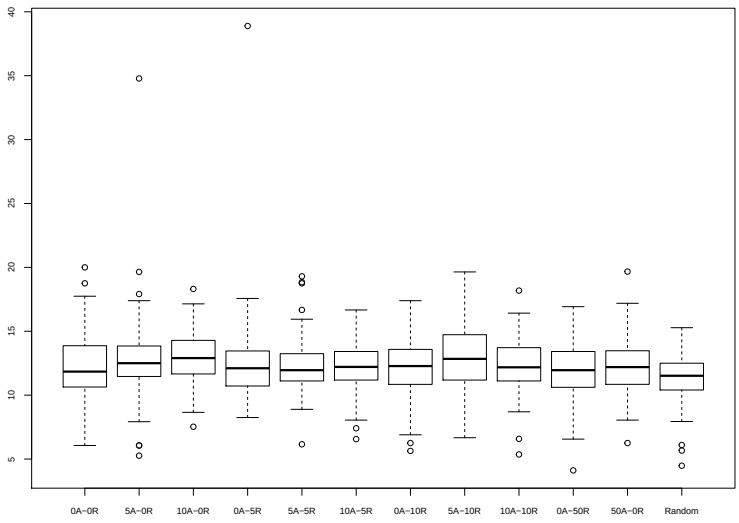


FIGURE 7.22: The figure shows for each network the distribution (obtained with the 100 MTGO run) of the percentage values of the not GO significant clusters. The labels in x-axis identify the perturbed networks, each label *nA-mR* means the network has been obtained adding $n\%$ of edges and removing $m\%$ of edges.

Chapter 8

Conclusions and future developments

8.1 Protein Co-expression Networks

Protein co-expression networks give the opportunity to represent and to evaluate biological contexts at system level, including organisms that lack information about PPIs. In fact, except for human and other few organisms, the theoretical models to describe the real-world networks are incomplete, and with a connectivity affected by false positive interactions. Although literature is yet too weak, protein co-expression networks represent a valid approach to obtain a novel overview of proteomic data and to provide new hypotheses about key molecules acting in pathological/physiological states. Of course, its real value has to be assessed by further studies, but preliminary findings make it promising.

The reviewed studies (in Paragraph 3.2) have evidenced a good relation between the topology of protein co-expression networks and the emergent phenotypes. Like PPI networks, the investigation of key proteins and modules from these graphs has proved to select the most important bio-molecules. Despite these findings, statistical methods to construct co-expression networks by processing large-scale proteomic data still need to be improved. To date, the available applications are mainly based on WGCNA framework, and studies to evaluate other approaches are needed.

The main limitation to perform the construction of protein co-expression networks may be attributed to the difficulty in measuring a proteome with enough coverage. A major consequence is the high rate of missing values that introduce loss of information and significant bias. This aspect will be surely improved by future advances of the proteomic technologies which in recent years have received a big boost from genome sequencing and from the combination of liquid

chromatography and mass spectrometry. In any case, the availability of large-scale proteomic data already offers a new range of opportunities to improve the existing network models, and in particular PPI, in understanding the mechanisms behind the emergent phenotypes.

Collection and integration of different -omics data represent essential points to perform a global evaluation of the biological systems and to improve the effectiveness of the current systems biology approaches. For these purposes, genomic and proteomic data are often used in combination with PPI networks. In this scenario, PPIs and co-expression networks provide the possibility to apply a multi-omic strategy that should improve level of significance in understanding biological mechanisms, including those related to diseases.

Computational tools are required to effectively build and integrate models to represent data from multi-omic experiments. In addition to basic research, these improvements may have important effects into clinical applications opening the way toward the use of multiple biomarkers and their relationships. These approaches represent a chance to generate new advanced diagnosis and prognosis methods which may lead toward a more preventive, predictive, and personalized medicine. These objectives are the major challenges to be addressed in the near future, and their achievement rely on the synergistic cooperation of biologists, physicists, mathematicians, and bioinformaticians.

The pipeline described in Chapter 3 aims to propose a procedure to build protein co-expression networks, facing the related critical aspects. On one hand, it shows a simple way to overcome the problem of missing data, and on the other hand, it realizes the integration of this model with the PPI data, in order to enrich the model and facilitate the understanding of the biological mechanism represented. Moreover, this pipeline presents an original method based on network models for a global investigation of amyloidosis disease, starting from proteomics data. The method includes (I) the creation of the models, (II) the topological analysis and (II) the model integration with other information, such as PPI and the edge differential analysis.

This elaboration has the final aim to provide an easy-to-understand model of the phenomenon studied, in order to facilitate the work of interpretation and understanding of the big quantity of proteomics data, in particular to support the work of biologist or clinician to extract from these rough data-sets useful information for biological/medical applications.

However, some steps of this procedure should be evaluated in more detail, for example, other approaches to face missing data or to compute network edges should be applied and compared. In addition, the hypothesis generated with the proposed models should be evaluated by targeted experiments, especially

as regards the role of the four identified proteins for the amyloidosis disease. Finally, the scripts created to perform the different steps of analysis (Python, Cytoscape, R software) should be integrated into a unique software solution to make easy-to-use the pipeline for future applications. These goals will be taken into account for future developments of this work.

8.2 MTGO: PPI network analysis via topological and functional module identification

With hundreds to thousands nodes, and even more edges, PPI networks are impossible to manually analyze in detail. For this reason, more and more algorithms have been proposed to automatically identify functional parts of these networks, i.e. groups of proteins, called modules or clusters. Traditional clustering approaches provide as result solely a set of clusters. As consequence, a second step of analysis (for example Gene Ontology (GO) enrichment analysis) is needed to investigate the biological role of the clusters.

In this dissertation a novel method to identify functional modules in PPI networks has been presented, called MTGO (Module detection via Topological information and Gene Ontology knowledge). Its theoretical architecture is based on the optimization of both GO term annotations and topology measures.

The MTGO approach differs from previous works as it integrates GO terms directly in the construction of functional modules. It provides as result both a set of clusters/communities (topological modules) and a list of functional modules, represented by GO terms, each associated to a cluster. In other words, MTGO facilitates and simplifies the PPI Network analysis coupling in a single step both the clustering analysis and the biological/GO analysis (as it provides both the clusters and their biological meaning, through the set of GO terms describing the clusters). MTGO is therefore not just a clustering algorithm but also a tool to automatically analyze the biological phenomenon described by a PPI network and guide experts' research providing clinically interpretable results.

MTGO provides both overlapping and full network coverage, two optimal features for module identification algorithms. In particular, topological modules ensures full coverage of the network, while functional modules are allowed to share nodes, *de facto* allowing overlapping. On the other hand, it must be noted that MTGO does not consider topological overlapping (i.e. the topological modules belong to a graph partition, thus they are separated by definition). MTGO heavily depends on the quality of the associated GO, therefore if this is not

well represented, or lacks information, it is biased, the results are affected negatively. To overcome this issue, MTGO gives the possibility to use the results optimized for density (see Pargraph 4.3.5). Furthermore, MTGO is tailored to search for small or sparse modules, important to shed light on the analysed network, which typically elude other approaches. Finally, MTGO has been implemented in a software version, available freely at https://gitlab.com/d1vella/MTGO.

Tested on benchmark scenarios, MTGO provides results comparable to or better than state of the art algorithms. Furthermore, by optimizing a trade-off between GO terms and topology, MTGO is extremely accurate in unveiling small and/or sparse functional modules, often missed by other algorithms. The high reliability of MTGO-retrieved modules is confirmed by GO term enriched analysis, with associated p-values comparable to or better than other GO-based state of the art algorithms. In addition, the significance of GO terms provided for describing the PPI network has been investigated using a statistical test. The percentage of topological modules associated with significant GO terms has been computed and it results greater than 80% in four different networks.

Moreover, the stability of MTGO has been explored. The results, obtained testing MTGO on perturbed network, showed that the algorithm is more stable in case of false negative edges than false positive edges (adding false edges is more damaging than removing existing links).

As future direction, the use of functional/topological module identified by MTGO should be explored in order to define the *disease modules*. This application is particularly interesting for Protein Co-expression Networks, where edges represent protein relations in the specific physiological/pathological context analyzed. MTGO has the ability to select a subset of GO terms describing a protein network, i.e. each GO term selected is biologically linked to a protein subset represented in the network in form of nodes sharing an high number of edges. For this reason, the application of MTGO on a Protein Co-expression Network allows to exploit at most its ability, because in co-expression networks the edges are directly inferred from the biological system investigated. In this way, the comparison of MTGO functional and topological sets in case (disease) vs control (healthy) networks would pinpoint the GO term difference and network rewiring characterizing the analyzed disease. In other words, explicitly addressing the disrupted/altered cellular functions. Finally, future works should be addressed to improve two aspects: the run time optimization and the building of a graphical interface. A more deep run time evaluation should be executed, in order to investigate the temporal limit of MTGO and to optimize the run time for applications on bigger networks. Furthermore, a graphical interface should

be provided to make the software easy-to-use, especially for those users, as biologists, who may not possess informatic skills.

Appendix A

A.1 Details of the Iteration phase

In this section a detailed description of the two main steps of the Iteration phase is presented.

Step 1.

Topological modules are randomly processed at each iteration k. Each topological module with $size < minSize$ (see main test Section 1.5 for a parameter description) is discarded and its nodes are added to the Temporary Node List (TNL). The rationale behind TNL is to use it as a temporary repository for discarded nodes. Discarding small topological modules is the way of MTGO to decrease H, i.e. to decrease the number of modules between two consecutive iterations. For each remaining topological module c_h^k, the following four steps are executed:

1.1 Δ_h^k is the subset of Δ containing the elements involving c_h^k nodes; Each $\delta_{p,h}^k$, $p-th$ element of Δ_h^k is evaluated by the *Selection* γ function (see function S1 in Chapter 4).

The functional module minimizing *Selection* γ (the *best* functional module $\delta_{B,h}^k$) is assigned to c_h^k.

The attribution of $\delta_{B,h}^k$ to c_h^k defines three node sets V_a, V_B and V_c, as follows:

$$V_a = \delta_{B,h}^k \cap c_h^k; V_B = c_h^k - V_a; V_c = \delta_{B,h}^k - V_a \tag{S6}$$

V_a are the nodes shared by $\delta_{B,h}^k$ and c_h^k; V_B are the nodes belonging to c_h^k but not to $\delta_{B,h}^k$; V_c are the nodes belonging to $\delta_{B,h}^k$ but not to c_h^k. Note that V_c nodes belong to other topological modules of the partition. At the end of Step 1, each c_h^k has its associated $\delta_{B,h}^k$, generating V_a, V_B and V_c.

1.2 Set V_a remains in topological module c_h^k.

1.3 Set V_B is moved to the TNL. V_B nodes are excluded from the $\delta^k_{B,h}$, and therefore they are not significantly related to the biological meaning assigned to c^k_h.

1.4 A single node $v_i \in V_c$ belongs to either another topological module of the current partition C^k, or the TNL (e.g. it has been assigned to the TNL by processing a previous c^k_m and $\delta^k_{B,m}$ pair). v_i is moved to the topological module c^k_h either if (i) it belongs to the TNL; or (ii) if $VM(c^k_h, v_i) > VM(c^k_m, v_i)$ (see function S3 in Chapter 4), where c^k_m is the other topological module of C^k containing it. In other words, v_i is added to c^k_h if the topological quality is increased by adding it. Figure 1.6 in Chapter 4 graphically describes this process.

When these four steps are executed for all the topological modules of C^k, MTGO performs another size check. Topological modules with $size < minSize$ are discarded, and their nodes added to the TNL. At this point, all nodes have been assigned either to the topological modules, or to the TNL. Finally, we need to empty the TNL.

Step 2.

In this step the nodes of the TNL are re-assigned to the topological modules:

2.1 All the TNL nodes with at least an associated δ_p, N_{GO} set, are used to create a new topological module c_{TLN} (Figure 1.7 in Chapter 4).

2.2 Each node v_i without any associated δ_p is assigned to the existing topological module maximizing its $MV(c^k_h, v_i)$ (see function S3 in Chapter 4) (Figure 1.7 in Chapter 4).

2.3 c_{TLN} is the garbage topological module made of the rejected nodes. To integrate it in the partition, *Step 1* is repeated and a second TNL is created. This procedure is the way MTGO introduces new topological modules and increases H in two consecutive iterations. Φ^{k+1} set is obtained here, by grouping all $\delta^k_{B,h}$s and their associated $l^k_{B,h}$s.

2.4 The second TNL is emptied assigning each node v_i to the topological module maximizing its $MV(c^k_h, v_i)$, regardless of having at least an associated δ_p or not.

Now all the nodes of TNL have been assigned to topological modules (full coverage). The obtained topological modules are grouped in the new partition

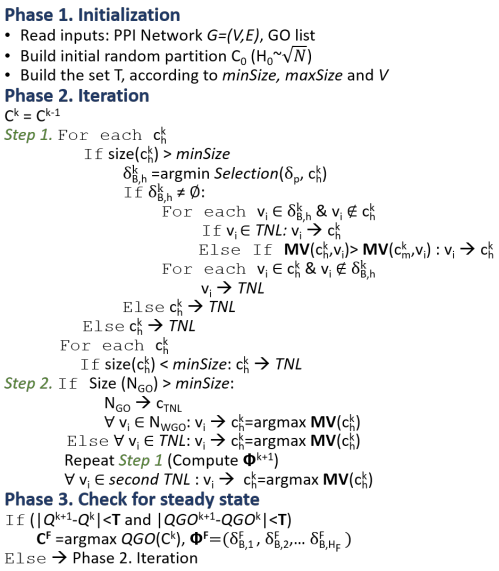


FIGURE A.1: MTGO pseudocode

C^{k+1}. While C^{k+1} is computed as described above, Φ^{k+1} is simply the set of all $\delta^k_{B,h}$s, along with their associated $l^k_{B,h}$s.

Detailed pseudocode of the Iteration phase, along with the whole MTGO algorithm are provided in the Figure A.1.

A.2 Table A

CYC2008

Krogan

	Precision	Recall	F-measure	Coverage	Positive Predictive Value
MTGO	0,32	0,36	0,34	0,63	0,46
ClusterOne	0,29	0,28	0,29	0,43	0,70
GMFTP	0,33	0,27	0,30	0,50	0,73
MCL	0,15	0,24	0,18	0,55	0,68
MCODE	0,36	0,07	0,12	0,25	0,58
DCAFP	0,37	0,21	0,27	0,35	0,65
COACH	0,11	0,09	0,10	0,58	0,25
Cfinder	0,51	0,09	0,15	0,35	0,53

Collins

	Precision	Recall	F-measure	Coverage	Positive Predictive Value
MTGO	0,41	0,38	0,39	0,67	0,43
ClusterOne	0,45	0,36	0,40	0,67	0,64
GMFTP	0,52	0,28	0,37	0,54	0,67
MCL	0,45	0,36	0,40	0,67	0,64
MCODE	0,57	0,15	0,24	0,43	0,65
DCAFP	0,57	0,15	0,24	0,43	0,65
COACH	0,38	0,17	0,24	0,60	0,43
Cfinder	0,39	0,14	0,21	0,51	0,53

Gavin

	Precision	Recall	F-measure	Coverage	Positive Predictive Value
MTGO	0,35	0,33	0,34	0,61	0,48
ClusterOne	0,24	0,21	0,22	0,54	0,62
GMFTP	0,41	0,27	0,33	0,51	0,68
MCL	0,19	0,17	0,18	0,57	0,59
MCODE	0,43	0,13	0,20	0,41	0,57
DCAFP	0,38	0,19	0,26	0,42	0,59
COACH	0,19	0,11	0,14	0,60	0,40
Cfinder	0,39	0,14	0,21	0,51	0,53

MIPS + SGD

Krogan

	Precision	Recall	F-measure	Coverage	Positive Predictive Value
MTGO	0,31	0,39	0,34	0,52	0,23
ClusterOne	0,24	0,25	0,24	0,33	0,32
GMFTP	0,33	0,33	0,33	0,41	0,32
MCL	0,12	0,21	0,15	0,42	0,31
MCODE	0,39	0,09	0,15	0,24	0,24
DCAFP	0,33	0,24	0,28	0,31	0,28
COACH	0,12	0,11	0,11	0,47	0,12
Cfinder	0,51	0,11	0,18	0,31	0,24

Collins

	Precision	Recall	F-measure	Coverage	Positive Predictive Value
MTGO	0,39	0,39	0,39	0,54	0,24
ClusterOne	0,36	0,37	0,36	0,49	0,32
GMFTP	0,51	0,35	0,41	0,45	0,32
MCL	0,36	0,34	0,35	0,50	0,31
MCODE	0,57	0,20	0,30	0,37	0,28
DCAFP	0,37	0,26	0,30	0,41	0,37
COACH	0,42	0,24	0,31	0,50	0,27
Cfinder	0,60	0,16	0,25	0,43	0,26

avin

	Precision	Recall	F-measure	Coverage	Positive Predictive Value
MTGO	0,35	0,37	0,36	0,50	0,24
ClusterOne	0,22	0,23	0,22	0,43	0,29
GMFTP	0,36	0,30	0,33	0,41	0,30
MCL	0,18	0,20	0,19	0,46	0,29

Ga		**MCODE**	0,41	0,17	0,24	0,36	0,25
		DCAFP	0,34	0,24	0,28	0,36	0,27
		COACH	0,18	0,13	0,15	0,49	0,22
		Cfinder	0,42	0,19	0,27	0,43	0,25
Human	**HSAPI**		**Precision**	**Recall**	**F-measure**	**Coverage**	**Positive Predictive Value**
		MTGO	0,10	0,12	0,11	0,48	0,05
		ClusterOne	0,10	0,05	0,07	0,24	0,15
		GMFTP	NA	NA	NA	NA	NA
		MCL	0,07	0,06	0,07	0,36	0,15
		MCODE	0,21	0,02	0,04	0,14	0,10
		DCAFP	0,17	0,03	0,05	0,14	0,14
		COACH	0,08	0,01	0,02	0,42	0,05
		Cfinder	0,40	0,02	0,04	0,13	0,10

		Accuracy	MMR	Composite Score	PC	TC	OTC	OPC
Krogan	CYC2008	0,54	0,40	1,30	453	408	146	143
		0,55	0,29	1,12	391	408	113	115
		0,60	0,28	1,15	319	408	111	105
		0,61	0,32	1,17	645	408	98	94
		0,38	0,10	0,55	77	408	28	28
		0,48	0,21	0,90	618	408	86	231
		0,38	0,14	0,61	963	408	37	109
		0,43	0,10	0,61	67	408	35	34
Collins		Accuracy	MMR	Composite Score	PC	TC	OTC	OPC
		0,54	0,40	1,31	353	408	153	145
		0,66	0,41	1,42	330	408	163	150
		0,61	0,29	1,18	199	408	116	104
		0,66	0,39	1,40	297	408	146	134
		0,53	0,17	0,84	95	408	61	54
		0,53	0,22	0,89	649	408	92	256
		0,51	0,21	0,89	1098	408	70	422
		0,52	0,15	0,81	75	408	51	45
Gavin		Accuracy	MMR	Composite Score	PC	TC	OTC	OPC
		0,54	0,35	1,22	362	408	133	127
		0,58	0,26	1,05	310	408	85	75
		0,59	0,27	1,13	246	408	111	101
		0,58	0,24	0,99	324	408	62	70
		0,48	0,15	0,76	109	408	52	47
		0,50	0,21	0,90	702	408	79	267
		0,49	0,17	0,77	1301	408	45	242
		0,52	0,17	0,84	137	408	59	54
Krogan	MIPS + SGD	Accuracy	MMR	Composite Score	PC	TC	OTC	OPC
		0,34	0,43	1,16	453	509	198	140
		0,33	0,28	0,85	391	509	125	92
		0,36	0,33	1,01	319	509	166	104
		0,36	0,30	0,87	645	509	106	76
		0,24	0,14	0,47	77	509	46	30
		0,30	0,26	0,80	618	509	124	206
		0,24	0,17	0,52	963	509	55	117
		0,27	0,14	0,52	67	509	56	34
Collins		Accuracy	MMR	Composite Score	PC	TC	OTC	OPC
		0,36	0,43	1,18	353	509	199	136
		0,40	0,40	1,16	330	509	188	118
		0,38	0,36	1,08	199	509	178	101
		0,39	0,38	1,11	297	509	171	107
		0,32	0,24	0,76	95	509	102	54
		0,39	0,28	0,92	649	509	132	239
		0,37	0,28	0,89	1098	509	122	465
		0,33	0,19	0,68	75	509	81	45
avin		Accuracy	MMR	Composite Score	PC	TC	OTC	OPC
		0,34	0,39	1,10	362	509	187	125
		0,35	0,28	0,86	310	509	117	68
		0,35	0,31	0,96	246	509	153	89
		0,37	0,26	0,82	324	509	100	58

Ga		0,30	0,20		0,66	109	509	85	45
		0,31	0,26		0,81	702	509	121	239
		0,32	0,21		0,66	1301	509	67	240
		0,33	0,22		0,74	137	509	99	58
Human	HSAPI	**Accuracy**	**MMR**	**Composite Score**	**PC**	**TC**	**OTC**	**OPC**	
		0,15	0,22		0,49	691	1765	203	70
		0,19	0,13		0,38	441	1765	93	46
		NA	NA	NA	NA		1765	NA	NA
		0,23	0,19		0,48	698	1765	111	51
		0,12	0,05		0,19	57	1765	40	12
		0,14	0,08		0,25	227	1765	47	39
		0,14	0,09		0,24	441	1765	23	44
		0,11	0,05		0,19	40	1765	41	16

A.3 Table B

			Percentage		Number of found GO	
P-value threshold			1,00E-03	1,00E-10	1,00E-03	1,00E-10
	P	DCAFP	72,49191	25,56634	9886	2106
	P	GMFTP	60,81505	15,98746	3336	590
	P	MTGO	92,27373	36,20309	9882	2511
	C	DCAFP	77,6699	39,4822	5053	1648
Krogan	C	GMFTP	53,9185	25,07837	1505	440
	C	MTGO	69,31567	31,56733	3351	1048
	F	DCAFP	57,60518	18,60841	2142	486
	F	GMFTP	44,20063	9,717868	719	137
	F	MTGO	74,39294	18,7638	2045	382
	P	DCAFP	88,90601	53,77504	15968	5359
	P	GMFTP	88,94472	36,1809	4069	797
	P	MTGO	91,21813	40,7932	7787	2103
	C	DCAFP	91,52542	64,71495	9151	4316
Collins	C	GMFTP	85,42714	48,24121	1841	593
	C	MTGO	82,71955	42,49292	3382	1205
	F	DCAFP	76,73344	37,44222	3108	807
	F	GMFTP	75,37688	20,1005	812	169
	F	MTGO	76,77054	22,94618	1622	325
	P	DCAFP	83,04843	43,01994	13865	3281
	P	GMFTP	71,95122	26,01626	3587	644
	P	MTGO	95,02762	40,88398	7900	1955
	C	DCAFP	85,18519	55,55556	7486	2778
Gavin	C	GMFTP	67,47967	35,77236	1635	517
	C	MTGO	79,83425	37,01657	3074	1092
	F	DCAFP	68,37607	26,21083	2778	642
	F	GMFTP	58,53659	16,66667	774	156
	F	MTGO	79,28177	24,86188	1858	414
	P	DCAFP	62,55507	9,251101	1526	130
	P	GMFTP	/	/	/	/
	P	MTGO	91,0275	42,54703	32512	7865
	C	DCAFP	57,70925	15,4185	718	84
Human	C	GMFTP	/	/	/	/
	C	MTGO	65,41245	27,6411	5131	1336
	F	DCAFP	43,17181	8,810573	507	47
	F	GMFTP	/	/	/	/
	F	MTGO	81,76556	28,65412	7113	1197

A.4 Table C

Nodes	(*minSize* 5-*maxSize*	Description (*minSize* 5-*maxSize* 30)
ARF4	6414	translational elongation
PHB2	6414	translational elongation
DDOST	6414	translational elongation
CACNA1C	6414	translational elongation
CAND1	6414	translational elongation
CISD1	6414	translational elongation
DDX55	6414	translational elongation
EEF1A1	6414	translational elongation
EEF1A2	6414	translational elongation
EEF1G	6414	translational elongation
EIF4A2	6414	translational elongation
EIF5	6414	translational elongation
EIF5A2	6414	translational elongation
EIF6	6414	translational elongation
GLUD2	6414	translational elongation
PABPC4	6414	translational elongation
RPL10A	6414	translational elongation
RPL14	6414	translational elongation
RPL17	6414	translational elongation
RPL19	6414	translational elongation
RPL21	6414	translational elongation
RPL23	6414	translational elongation
RPL27	6414	translational elongation
RPL3	6414	translational elongation
RPL38	6414	translational elongation
RPL5	6414	translational elongation
RPL6	6414	translational elongation
RPL9	6414	translational elongation
RPLP2	6414	translational elongation
RPS10	6414	translational elongation
RPS13	6414	translational elongation
RPS15	6414	translational elongation
RPS15A	6414	translational elongation
RPS2	6414	translational elongation
RPS20	6414	translational elongation
RPS26	6414	translational elongation
RPS3A	6414	translational elongation
RPS4X	6414	translational elongation
RPS6	6414	translational elongation
SND1	6414	translational elongation
SOD2	6414	translational elongation
TUFM	6414	translational elongation
VARS	6414	translational elongation
GSN	10038	response to metal ion
ALDH3A2	10038	response to metal ion
SLC25A12	10038	response to metal ion
SLC25A13	10038	response to metal ion
FGA	10038	response to metal ion
FGB	10038	response to metal ion

FGG	10038	response to metal ion
BSG	10038	response to metal ion
COL1A1	30155	regulation of cell adhesion
NACA	30155	regulation of cell adhesion
NID2	30155	regulation of cell adhesion
PPP2CA	30155	regulation of cell adhesion
PPP2R1A	30155	regulation of cell adhesion
NIT2	30155	regulation of cell adhesion
NPTN	30155	regulation of cell adhesion
BGN	30155	regulation of cell adhesion
LTBP1	30155	regulation of cell adhesion
TGFBI	30155	regulation of cell adhesion
THBS1	30155	regulation of cell adhesion
VTN	30155	regulation of cell adhesion
RAB7A	30155	regulation of cell adhesion
ARL8B	15992	proton transport
ASAH1	15992	proton transport
ATP5A1	15992	proton transport
ATP5B	15992	proton transport
ATP5C1	15992	proton transport
ATP5F1	15992	proton transport
ATP5J2	15992	proton transport
ATP5O	15992	proton transport
ATP6V1A	15992	proton transport
COX5A	15992	proton transport
COX5B	15992	proton transport
COX6B1	15992	proton transport
COX6C	15992	proton transport
GBP1	15992	proton transport
ITIH2	15992	proton transport
SLC25A4	15992	proton transport
TIPRL	15992	proton transport
USMG5	15992	proton transport
NNT	15992	proton transport
ATP6V1E1	15992	proton transport
AKAP12	6605	protein targeting
CALR	6605	protein targeting
CAPN2	6605	protein targeting
GMFB	6605	protein targeting
MAPK1	6605	protein targeting
NDUFA13	6605	protein targeting
PEA15	6605	protein targeting
CDC42	6605	protein targeting
ATP2A2	6605	protein targeting
CAPNS1	6605	protein targeting
ESYT1	6605	protein targeting
SUB1	6605	protein targeting
ABLIM1	6605	protein targeting
ACTR1A	6605	protein targeting
CLIC4	6605	protein targeting

CPNE1	6605	protein targeting
CTNNA1	6605	protein targeting
DCTN1	6605	protein targeting
DYNC1H1	6605	protein targeting
IPO5	6605	protein targeting
KPNB1	6605	protein targeting
LMNA	6605	protein targeting
LMNB1	6605	protein targeting
LUC7L2	6605	protein targeting
MYOF	6605	protein targeting
NUTF2	6605	protein targeting
PDXK	6605	protein targeting
PEX19	6605	protein targeting
PKP2	6605	protein targeting
PRKAR2B	6605	protein targeting
SPTBN1	6605	protein targeting
TUBA3C	6605	protein targeting
YWHAG	6605	protein targeting
YWHAZ	6605	protein targeting
CANX	6457	protein folding
PDIA4	6457	protein folding
HSPA1A	6457	protein folding
NQO1	6457	protein folding
PPIF	6457	protein folding
ST13	6457	protein folding
CCT5	6457	protein folding
CCT6A	6457	protein folding
CCT7	6457	protein folding
CCT8	6457	protein folding
PPP2R2A	6457	protein folding
PPP2R2B	6457	protein folding
HSPA9	6457	protein folding
PPIB	6457	protein folding
ERO1L	6457	protein folding
S100A10	48585	negative regulation of response to stimulus
A2M	48585	negative regulation of response to stimulus
AMBP	48585	negative regulation of response to stimulus
CTSB	48585	negative regulation of response to stimulus
INPP5D	48585	negative regulation of response to stimulus
RTN4	48585	negative regulation of response to stimulus
COL1A2	48585	negative regulation of response to stimulus
COL3A1	48585	negative regulation of response to stimulus
COL6A1	48585	negative regulation of response to stimulus
COL6A3	48585	negative regulation of response to stimulus
SDC2	48585	negative regulation of response to stimulus
SERPINC1	48585	negative regulation of response to stimulus
SPARC	48585	negative regulation of response to stimulus
TNC	48585	negative regulation of response to stimulus
APOE	48585	negative regulation of response to stimulus
ACAA2	48585	negative regulation of response to stimulus

H2AFZ	48585	negative regulation of response to stimulus
CAMK2B	3012	muscle system process
CAMK2G	3012	muscle system process
MYOM1	3012	muscle system process
ACTA2	3012	muscle system process
ACTG2	3012	muscle system process
ACTN2	3012	muscle system process
CACNA1S	3012	muscle system process
CAPN1	3012	muscle system process
CKM	3012	muscle system process
CRYAB	3012	muscle system process
DES	3012	muscle system process
DMD	3012	muscle system process
LDB3	3012	muscle system process
MYBPC3	3012	muscle system process
MYH10	3012	muscle system process
MYH11	3012	muscle system process
MYH2	3012	muscle system process
MYH3	3012	muscle system process
MYH4	3012	muscle system process
MYH7	3012	muscle system process
MYL1	3012	muscle system process
MYL12A	3012	muscle system process
MYL12B	3012	muscle system process
MYL6	3012	muscle system process
MYL6B	3012	muscle system process
MYLK	3012	muscle system process
MYOM2	3012	muscle system process
MYOZ2	3012	muscle system process
PGM5	3012	muscle system process
SLMAP	3012	muscle system process
TMOD1	3012	muscle system process
TTN	3012	muscle system process
UTRN	3012	muscle system process
VASP	3012	muscle system process
WAS	3012	muscle system process
CAMK2D	3012	muscle system process
MYL4	3012	muscle system process
MYL9	3012	muscle system process
SNCG	3012	muscle system process
HSPB2	3012	muscle system process
HSPB3	3012	muscle system process
CHCHD3	16071	mRNA metabolic process
DCPS	16071	mRNA metabolic process
DDX17	16071	mRNA metabolic process
DHX9	16071	mRNA metabolic process
FLNC	16071	mRNA metabolic process
HDGF	16071	mRNA metabolic process
HNRNPA1	16071	mRNA metabolic process
HNRNPA2B1	16071	mRNA metabolic process

HNRNPA3	16071	mRNA metabolic process
HNRNPAB	16071	mRNA metabolic process
HNRNPD	16071	mRNA metabolic process
HNRNPF	16071	mRNA metabolic process
HNRNPK	16071	mRNA metabolic process
HNRNPU	16071	mRNA metabolic process
IMMT	16071	mRNA metabolic process
MOV10	16071	mRNA metabolic process
PABPC1	16071	mRNA metabolic process
RBM39	16071	mRNA metabolic process
RBM8A	16071	mRNA metabolic process
SF3B1	16071	mRNA metabolic process
SFPQ	16071	mRNA metabolic process
SNRNP200	16071	mRNA metabolic process
SNRPE	16071	mRNA metabolic process
SNRPF	16071	mRNA metabolic process
SSB	16071	mRNA metabolic process
VDAC2	16071	mRNA metabolic process
YBX1	16071	mRNA metabolic process
PHB	16071	mRNA metabolic process
HDLBP	43632	modification-dependent macromolecule catabolic process
CAST	43632	modification-dependent macromolecule catabolic process
CAT	43632	modification-dependent macromolecule catabolic process
PSMA1	43632	modification-dependent macromolecule catabolic process
PSMA2	43632	modification-dependent macromolecule catabolic process
PSMA3	43632	modification-dependent macromolecule catabolic process
PSMA6	43632	modification-dependent macromolecule catabolic process
PSMB1	43632	modification-dependent macromolecule catabolic process
PSMB10	43632	modification-dependent macromolecule catabolic process
PSMB3	43632	modification-dependent macromolecule catabolic process
PSMB7	43632	modification-dependent macromolecule catabolic process
PSMB9	43632	modification-dependent macromolecule catabolic process
PSMC1	43632	modification-dependent macromolecule catabolic process
PSMD12	43632	modification-dependent macromolecule catabolic process
PSMD2	43632	modification-dependent macromolecule catabolic process
PSMD3	43632	modification-dependent macromolecule catabolic process
PSMD6	43632	modification-dependent macromolecule catabolic process
PSME2	43632	modification-dependent macromolecule catabolic process
RPN1	43632	modification-dependent macromolecule catabolic process
RPN2	43632	modification-dependent macromolecule catabolic process
STT3B	43632	modification-dependent macromolecule catabolic process
SUCLA2	43632	modification-dependent macromolecule catabolic process
ACADL	34440	lipid oxidation
ACADS	34440	lipid oxidation
ACADVL	34440	lipid oxidation
ETFA	34440	lipid oxidation
ETFB	34440	lipid oxidation
ETFDH	34440	lipid oxidation
GBAS	34440	lipid oxidation
HADH	34440	lipid oxidation

HADHA	34440	lipid oxidation
HADHB	34440	lipid oxidation
AP2B1	48193	Golgi vesicle transport
CKMT2	48193	Golgi vesicle transport
COPG2	48193	Golgi vesicle transport
HLA-A	48193	Golgi vesicle transport
NAPA	48193	Golgi vesicle transport
NAPB	48193	Golgi vesicle transport
SAR1A	48193	Golgi vesicle transport
SEC22B	48193	Golgi vesicle transport
TMED2	48193	Golgi vesicle transport
TMED9	48193	Golgi vesicle transport
USO1	48193	Golgi vesicle transport
VPS29	48193	Golgi vesicle transport
HN1	48193	Golgi vesicle transport
SEPT2	48193	Golgi vesicle transport
SEPT6	48193	Golgi vesicle transport
SEPT9	48193	Golgi vesicle transport
AIFM1	48193	Golgi vesicle transport
HSPA2	48193	Golgi vesicle transport
TXNDC5	48193	Golgi vesicle transport
ZYX	48193	Golgi vesicle transport
COPA	48193	Golgi vesicle transport
COPB1	48193	Golgi vesicle transport
COPB2	48193	Golgi vesicle transport
COPG	48193	Golgi vesicle transport
COPZ1	48193	Golgi vesicle transport
CLTA	48193	Golgi vesicle transport
CLTC	48193	Golgi vesicle transport
DNM2	48193	Golgi vesicle transport
TWF2	6006	glucose metabolic process
NCL	6006	glucose metabolic process
TROVE2	6006	glucose metabolic process
VDAC3	6006	glucose metabolic process
NEBL	6006	glucose metabolic process
ACLY	6006	glucose metabolic process
ALDH2	6006	glucose metabolic process
ACTN1	6006	glucose metabolic process
GNAI2	6006	glucose metabolic process
GNAI3	6006	glucose metabolic process
MDH2	6006	glucose metabolic process
PRDX5	6006	glucose metabolic process
SLC25A11	6006	glucose metabolic process
GNAQ	6006	glucose metabolic process
AGL	6006	glucose metabolic process
AHCY	6006	glucose metabolic process
ALDOA	6006	glucose metabolic process
ALDOC	6006	glucose metabolic process
CFL2	6006	glucose metabolic process
CKB	6006	glucose metabolic process

DDX3X	6006	glucose metabolic process
DLAT	6006	glucose metabolic process
ENO1	6006	glucose metabolic process
ENO3	6006	glucose metabolic process
FBP1	6006	glucose metabolic process
GANAB	6006	glucose metabolic process
GAPDH	6006	glucose metabolic process
GLO1	6006	glucose metabolic process
GPI	6006	glucose metabolic process
HEXA	6006	glucose metabolic process
HIBADH	6006	glucose metabolic process
HK1	6006	glucose metabolic process
IQGAP1	6006	glucose metabolic process
LDHA	6006	glucose metabolic process
LDHAL6B	6006	glucose metabolic process
LDHB	6006	glucose metabolic process
MAP4	6006	glucose metabolic process
MDH1	6006	glucose metabolic process
OGDH	6006	glucose metabolic process
PDHB	6006	glucose metabolic process
PDHX	6006	glucose metabolic process
PFKM	6006	glucose metabolic process
PGAM2	6006	glucose metabolic process
PGD	6006	glucose metabolic process
PHGDH	6006	glucose metabolic process
PKM2	6006	glucose metabolic process
PLS3	6006	glucose metabolic process
PPA1	6006	glucose metabolic process
PYGB	6006	glucose metabolic process
PYGM	6006	glucose metabolic process
SORBS2	6006	glucose metabolic process
STIP1	6006	glucose metabolic process
TKT	6006	glucose metabolic process
TPI1	6006	glucose metabolic process
UGP2	6006	glucose metabolic process
RAB14	16197	endosome transport
RHOB	16197	endosome transport
GC	16197	endosome transport
DIABLO	16197	endosome transport
LGMN	16197	endosome transport
ECH1	16197	endosome transport
SERPINB9	16197	endosome transport
DPY30	16197	endosome transport
DBNL	16197	endosome transport
EHD1	16197	endosome transport
HLA-DMA	16197	endosome transport
HLA-DQB1	16197	endosome transport
HLA-DRB1	16197	endosome transport
PICALM	16197	endosome transport
TOM1	16197	endosome transport

ACP6	16197	endosome transport
RSU1	16197	endosome transport
GFPT1	6952	defense response
HLA-B	6952	defense response
KCTD12	6952	defense response
VDAC1	6952	defense response
ALB	6952	defense response
CLU	6952	defense response
MIF	6952	defense response
PON1	6952	defense response
PRDX6	6952	defense response
TUBB1	6952	defense response
APOA2	6952	defense response
APOA4	6952	defense response
FH	6952	defense response
CLIC1	6952	defense response
FBN1	6952	defense response
HSPG2	6952	defense response
RAB5C	6952	defense response
SERPIND1	6952	defense response
SERPING1	6952	defense response
AHSG	6952	defense response
C1QA	6952	defense response
C4B	6952	defense response
CST3	6952	defense response
GCN1L1	6952	defense response
IGHV@	6952	defense response
KRT1	6952	defense response
KRT10	6952	defense response
KRT9	6952	defense response
SAMHD1	6952	defense response
PLEC	6952	defense response
CFH	6952	defense response
C6	19835	cytolysis
C5	19835	cytolysis
C7	19835	cytolysis
C8A	19835	cytolysis
C8B	19835	cytolysis
C8G	19835	cytolysis
C9	19835	cytolysis
SERPINH1	6333	chromatin assembly or disassembly
NME1	6333	chromatin assembly or disassembly
BDH1	6333	chromatin assembly or disassembly
OXCT1	6333	chromatin assembly or disassembly
CBX3	6333	chromatin assembly or disassembly
H1F0	6333	chromatin assembly or disassembly
H2AFY	6333	chromatin assembly or disassembly
HIST1H1A	6333	chromatin assembly or disassembly
HIST1H1B	6333	chromatin assembly or disassembly
HIST1H2BN	6333	chromatin assembly or disassembly

NAP1L1	6333	chromatin assembly or disassembly
SET	6333	chromatin assembly or disassembly
CMPK1	6333	chromatin assembly or disassembly
MYH9	6928	cellular component movement
PTPN6	6928	cellular component movement
RBM3	6928	cellular component movement
CALD1	6928	cellular component movement
COL12A1	6928	cellular component movement
FLNB	6928	cellular component movement
TLN1	6928	cellular component movement
ACTR2	6928	cellular component movement
ANP32B	6928	cellular component movement
ARPC1B	6928	cellular component movement
ARPC2	6928	cellular component movement
ARPC3	6928	cellular component movement
ARPC5	6928	cellular component movement
CALU	6928	cellular component movement
ARHGAP1	6928	cellular component movement
ARHGDIA	6928	cellular component movement
HBB	6928	cellular component movement
HBD	6928	cellular component movement
HBE1	6928	cellular component movement
HMGB1	6928	cellular component movement
HSPB1	6928	cellular component movement
RAC1	6928	cellular component movement
CDH2	6928	cellular component movement
CNN1	6928	cellular component movement
TPM3	6928	cellular component movement
COL5A1	6928	cellular component movement
COMP	6928	cellular component movement
DECR1	6928	cellular component movement
FBLN1	6928	cellular component movement
FMOD	6928	cellular component movement
FN1	6928	cellular component movement
ACAN	6928	cellular component movement
APOA1	6928	cellular component movement
VCAN	6928	cellular component movement
KRT2	6928	cellular component movement
HMGB2	6928	cellular component movement
MRPS18B	6928	cellular component movement
CRK	6928	cellular component movement
KRT77	6928	cellular component movement
KRT79	6928	cellular component movement
CORO1C	6928	cellular component movement
MFGE8	6928	cellular component movement
CYCS	42773	ATP synthesis coupled electron transport
MT-ND1	42773	ATP synthesis coupled electron transport
MT-ND4	42773	ATP synthesis coupled electron transport
NDUFA9	42773	ATP synthesis coupled electron transport
NDUFB6	42773	ATP synthesis coupled electron transport

NDUFB8	42773	ATP synthesis coupled electron transport
NDUFS1	42773	ATP synthesis coupled electron transport
NDUFS2	42773	ATP synthesis coupled electron transport
NDUFS3	42773	ATP synthesis coupled electron transport
NDUFS4	42773	ATP synthesis coupled electron transport
NDUFS7	42773	ATP synthesis coupled electron transport
NDUFV1	42773	ATP synthesis coupled electron transport
NDUFV2	42773	ATP synthesis coupled electron transport
SDHA	42773	ATP synthesis coupled electron transport
SDHB	42773	ATP synthesis coupled electron transport
SDHD	42773	ATP synthesis coupled electron transport
SUCLG2	42773	ATP synthesis coupled electron transport
UQCR10	42773	ATP synthesis coupled electron transport
UQCR11	42773	ATP synthesis coupled electron transport
UQCRB	42773	ATP synthesis coupled electron transport
UQCRC1	42773	ATP synthesis coupled electron transport
ALDH6A1	9310	amine catabolic process
COMT	9310	amine catabolic process
GOT2	9310	amine catabolic process
HIBCH	9310	amine catabolic process
IVD	9310	amine catabolic process
MTHFD1	9310	amine catabolic process
PACSIN3	9310	amine catabolic process
PCCB	9310	amine catabolic process
PDCD6IP	7015	actin filament organization
PTPN23	7015	actin filament organization
ACTC1	7015	actin filament organization
AHNAK	7015	actin filament organization
ANK1	7015	actin filament organization
ARPC4	7015	actin filament organization
EZR	7015	actin filament organization
PRKACB	7015	actin filament organization
PRKAR2A	7015	actin filament organization
SPTA1	7015	actin filament organization
TPT1	7015	actin filament organization
DBN1	7015	actin filament organization
CNN2	7015	actin filament organization
FSCN1	7015	actin filament organization
LCP1	7015	actin filament organization
FLNA	7015	actin filament organization
AK1	6103	2-oxoglutarate metabolic process
RCN1	6103	2-oxoglutarate metabolic process
ACO2	6103	2-oxoglutarate metabolic process
AK2	6103	2-oxoglutarate metabolic process
DLST	6103	2-oxoglutarate metabolic process
EFHD2	6103	2-oxoglutarate metabolic process
GOT1	6103	2-oxoglutarate metabolic process
IDH1	6103	2-oxoglutarate metabolic process
IDH3A	6103	2-oxoglutarate metabolic process
IDH3B	6103	2-oxoglutarate metabolic process

IDH3G	6103	2-oxoglutarate metabolic process
MTAP	6103	2-oxoglutarate metabolic process
RRBP1	6103	2-oxoglutarate metabolic process

Bibliography

[1] Laura Beretta. "Proteomics from the clinical perspective: many hopes and much debate: mass spectrometry has been rapidly maturing as the core technology at the heart of proteomics. The application of these powerful methods to the study of human diseases and their translation to the clinic, however, has been beset with unique challenges". In: *Nature Methods* 4.10 (2007), pp. 785–787.

[2] Lei Xie et al. "Towards structural systems pharmacology to study complex diseases and personalized medicine". In: *PLoS computational biology* 10.5 (2014), e1003554.

[3] Nigel Williams. "Biologists cut reductionist approach down to size". In: *Science* 277.5325 (1997), pp. 476–477.

[4] Marc HV Van Regenmortel. "Reductionism and complexity in molecular biology". In: *EMBO reports* 5.11 (2004), pp. 1016–1020.

[5] Leland H Hartwell et al. "From molecular to modular cell biology". In: *Nature* 402.6761 (1999), p. C47.

[6] Albert-Laszlo Barabasi and Zoltan N Oltvai. "Network biology: understanding the cell's functional organization". In: *Nature reviews genetics* 5.2 (2004), pp. 101–113.

[7] Ariel Bensimon, Albert JR Heck, and Ruedi Aebersold. "Mass spectrometry–based proteomics and network biology". In: *Annual review of biochemistry* 81 (2012), pp. 379–405.

[8] Fred D Mast, Alexander V Ratushny, and John D Aitchison. "Systems cell biology". In: *J Cell Biol* 206.6 (2014), pp. 695–706.

[9] Leonhard Euler. "Solutio problematis ad geometriam situs pertinentis". In: *Commentarii academiae scientiarum Petropolitanae* 8 (1741), pp. 128–140.

[10] Luonan Chen, Rui-Sheng Wang, and Xiang-Sun Zhang. *Biomolecular networks: methods and applications in systems biology*. Vol. 10. John Wiley & Sons, 2009.

[11] Jukka Westermarck, Johanna Ivaska, and Garry L Corthals. "Identification of protein interactions involved in cellular signaling". In: *Molecular & Cellular Proteomics* 12.7 (2013), pp. 1752–1763.

[12] Roded Sharan, Igor Ulitsky, and Ron Shamir. "Network-based prediction of protein function". In: *Molecular systems biology* 3.1 (2007), p. 88.

[13] Claudio Procaccini et al. "The proteomic landscape of human ex vivo regulatory and conventional T cells reveals specific metabolic requirements". In: *Immunity* 44.2 (2016), pp. 406–421.

[14] Mika Gustafsson et al. "Modules, networks and systems medicine for understanding disease and aiding diagnosis". In: *Genome medicine* 6.10 (2014), p. 82.

[15] Albert-László Barabási. "Scale-free networks: a decade and beyond". In: *science* 325.5939 (2009), pp. 412–413.

[16] Kai Sun et al. "Predicting disease associations via biological network analysis". In: *BMC bioinformatics* 15.1 (2014), p. 304.

[17] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease". In: *Nature reviews. Genetics* 12.1 (2011), p. 56.

[18] Marc Vidal, Michael E Cusick, and Albert-László Barabási. "Interactome networks and human disease". In: *Cell* 144.6 (2011), pp. 986–998.

[19] Laura I Furlong. "Human diseases through the lens of network biology". In: *Trends in Genetics* 29.3 (2013), pp. 150–159.

[20] Maxime Rotival and Enrico Petretto. "Leveraging gene co-expression networks to pinpoint the regulation of complex traits and disease, with a focus on cardiovascular traits". In: *Briefings in functional genomics* 13.1 (2013), pp. 66–78.

[21] Virja Mehta and Laura Trinkle-Mulcahy. "Recent advances in large-scale protein interactome mapping". In: *F1000Research* 5 (2016).

[22] Harry Jubb et al. "Structural biology and drug discovery for protein–protein interactions". In: *Trends in pharmacological sciences* 33.5 (2012), pp. 241–248.

[23] Hui Ge. "UPA, a universal protein array system for quantitative detection of protein–protein, protein–DNA, protein–RNA and protein–ligand interactions". In: *Nucleic Acids Research* 28.2 (2000), e3–e3.

[24] Archana Pan et al. "Computational analysis of protein interaction networks for infectious diseases". In: *Briefings in bioinformatics* 17.3 (2015), pp. 517–526.

[25] Roberto Mosca et al. "Towards a detailed atlas of protein–protein interactions". In: *Current opinion in structural biology* 23.6 (2013), pp. 929–940.

[26] Tong Hao et al. "Reconstruction and application of protein–protein interaction network". In: *International journal of molecular sciences* 17.6 (2016), p. 907.

[27] Rod K Nibbe, Mehmet Koyutürk, and Mark R Chance. "An integrative-omics approach to identify functional sub-networks in human colorectal cancer". In: *PLoS computational biology* 6.1 (2010), e1000639.

[28] Sangchul Rho et al. "From proteomics toward systems biology: integration of different types of proteomics data into network models". In: *BMB Rep* 41.3 (2008), pp. 184–193.

[29] Matthias Gstaiger and Ruedi Aebersold. "Applying mass spectrometry-based proteomics to genetics, genomics and network biology". In: *Nature Reviews Genetics* 10.9 (2009), pp. 617–627.

[30] Serene WH Wong, Nick Cercone, and Igor Jurisica. "Comparative network analysis via differential graphlet communities". In: *Proteomics* 15.2-3 (2015), pp. 608–617.

[31] Michelle Girvan and Mark EJ Newman. "Community structure in social and biological networks". In: *Proc. Natl. Acad. Sci. USA* 99.cond-mat/0112110 (2001), pp. 8271–8276.

[32] Avi Ma'ayan. "Network integration and graph analysis in mammalian molecular systems biology". In: *IET systems biology* 2.5 (2008), pp. 206–221.

[33] Peter Grindrod and Milla Kibble. "Review of uses of network and graph theory concepts within proteomics". In: *Expert review of proteomics* 1.2 (2004), pp. 229–238.

[34] Paul Erdos and Alfréd Rényi. "On the evolution of random graphs". In: *Publ. Math. Inst. Hung. Acad. Sci* 5.1 (1960), pp. 17–60.

[35] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *science* 286.5439 (1999), pp. 509–512.

[36] Danila Vella et al. "From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data". In: *EURASIP Journal on Bioinformatics and Systems Biology* 2017.1 (2017), p. 6.

[37] Attila Gursoy, Ozlem Keskin, and Ruth Nussinov. *Topological properties of protein interaction networks from a structural perspective*. 2008.

[38] Hunter B Fraser. "Modularity and evolutionary constraint on proteins". In: *Nature genetics* 37.4 (2005), p. 351.

[39] Victor Spirin and Leonid A. Mirny. "Protein complexes and functional modules in molecular networks." eng. In: *Proc Natl Acad Sci U S A* 100.21 (2003), pp. 12123–12128.

[40] David L Gibbs et al. "Multi-omic network signatures of disease". In: *Frontiers in genetics* 4 (2013).

[41] David L Gibbs et al. "Protein co-expression network analysis (ProCoNA)". In: *Journal of clinical bioinformatics* 3.1 (2013), p. 11.

[42] Chang Guo et al. "Characterization of protein species and weighted protein co-expression network regulation of escherichia coli in response to serum killing using a 2-de based proteomics approach". In: *Molecular BioSystems* 10.3 (2014), pp. 475–484.

[43] Duojiao Wu et al. "Network analysis reveals roles of inflammatory factors in different phenotypes of kidney transplant patients". In: *Journal of theoretical biology* 362 (2014), pp. 62–68.

[44] Matthew L MacDonald et al. "Altered glutamate protein co-expression network topology linked to spine loss in the auditory cortex of schizophrenia". In: *Biological psychiatry* 77.11 (2015), pp. 959–968.

[45] Evangelos I Kanonidis et al. "Protein co-expression analysis as a strategy to complement a standard quantitative proteomics approach: Case of a glioblastoma multiforme study". In: *PloS one* 11.8 (2016), e0161828.

[46] Xuexin Yu et al. "Quantitative proteomics reveals the novel co-expression signatures in early brain development for prognosis of glioblastoma multiforme". In: *Oncotarget* 7.12 (2016), p. 14161.

[47] Francesca Brambilla et al. "Reliable typing of systemic amyloidoses through proteomic analysis of subcutaneous adipose tissue". In: *Blood* (2011), blood–2011.

[48] I. Zoppis et al. "Analysis of correlation structures in renal cell carcinoma patient data". In: *BIOINFORMATICS 2012 - Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (2012), pp. 251–256.

[49] C Cava et al. "Combination of gene expression and genome copy number alteration has a prognostic value for breast cancer". In: *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*. IEEE. 2013, pp. 608–611.

[50] Jennyfer Bultinck, Sam Lievens, and Jan Tavernier. "Protein-protein interactions: network analysis and applications in drug discovery". In: *Current pharmaceutical design* 18.30 (2012), pp. 4619–4629.

[51] Aravind Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

[52] Francesca Brambilla et al. "Shotgun protein profile of human adipose tissue and its changes in relation to systemic amyloidoses". In: *Journal of proteome research* 12.12 (2013), pp. 5642–5655.

[53] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". In: *Nucleic acids research* 37.1 (2008), pp. 1–13.

[54] Michael Ashburner et al. "Gene Ontology: tool for the unification of biology". In: *Nature genetics* 25.1 (2000), p. 25.

[55] Chao Zhang et al. "NOA: a cytoscape plugin for network ontology analysis." eng. In: *Bioinformatics* 29.16 (2013), pp. 2066–2067.

[56] Andrey Alexeyenko et al. "Network enrichment analysis: extension of gene-set enrichment analysis to gene networks". In: *BMC bioinformatics* 13.1 (2012), p. 226.

[57] Pietro Di Lena et al. "NET-GE: a novel NETwork-based Gene Enrichment for detecting biological processes associated to Mendelian diseases". In: *BMC genomics* 16.8 (2015), S6.

[58] Allen L Hu and Keith CC Chan. "Utilizing both topological and attribute information for protein complex identification in ppi networks". In: *IEEE/ACM transactions on computational biology and bioinformatics* 10.3 (2013), pp. 780–792.

[59] Sriganesh Srihari and Hon Wai Leong. "A survey of computational methods for protein complex prediction from protein interaction networks". In: *Journal of bioinformatics and computational biology* 11.02 (2013), p. 1230002.

[60] Xiao-Fei Zhang et al. "Detecting overlapping protein complexes based on a generative model with functional and topological properties." eng. In: *BMC Bioinformatics* 15 (2014), p. 186.

[61] Lun Hu and Keith CC Chan. "A density-based clustering approach for identifying overlapping protein complexes with functional preferences". In: *BMC bioinformatics* 16.1 (2015), p. 174.

[62] Jian Wang et al. "Filtering Gene Ontology semantic similarity for identifying protein complexes in large protein interaction networks". In: *Proteome science* 10.1 (2012), S18.

[63] Morteza Kouhsar, Fatemeh Zare-Mirakabad, and Yousef Jamali. "WCOACH: Protein complex prediction in weighted PPI networks". In: *Genes & genetic systems* 90.5 (2015), pp. 317–324.

[64] Mark EJ Newman and Michelle Girvan. "Finding and evaluating community structure in networks". In: *Physical review E* 69.2 (2004), p. 026113.

[65] Elizabeth I Boyle et al. "GO:: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes". In: *Bioinformatics* 20.18 (2004), pp. 3710–3715.

[66] Harvey Lodish. *Molecular cell biology*. Macmillan, 2008.

[67] David L Nelson, Albert L Lehninger, and Michael M Cox. *Lehninger principles of biochemistry*. Macmillan, 2008.

[68] Michael PH Stumpf et al. "Estimating the size of the human interactome". In: *Proceedings of the National Academy of Sciences* 105.19 (2008), pp. 6959–6964.

[69] James D Watson and James D Watson. *Molecular biology of the gene*. Sirsi) i9780805395921. 2008.

[70] George W Beadle and Edward L Tatum. "Genetic control of biochemical reactions in Neurospora". In: *proceedings of the National Academy of Sciences* 27.11 (1941), pp. 499–506.

[71] Mark Schena et al. "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes". In: *Proceedings of the National Academy of Sciences* 93.20 (1996), pp. 10614–10619.

[72] Mike Tyers and Matthias Mann. "From genomics to proteomics". In: *Nature* 422.6928 (2003), pp. 193–197.

[73] Gene Ontology Consortium et al. "The gene ontology project in 2008". In: *Nucleic acids research* 36.suppl 1 (2008), pp. D440–D444.

[74] Réka Albert. "Network inference, analysis, and modeling in systems biology". In: *The Plant Cell Online* 19.11 (2007), pp. 3327–3338.

[75] Oliver Mason and Mark Verwoerd. "Graph theory and networks in biology". In: *IET systems biology* 1.2 (2007), pp. 89–119.

[76] Donald Petrey and Barry Honig. "Structural bioinformatics of the interactome". In: *Annual review of biophysics* 43 (2014), pp. 193–210.

[77] Havva Kohestani and Alessandro Giuliani. "Organization principles of biological networks: an explorative study". In: *Biosystems* 141 (2016), pp. 31–39.

[78] Zaynab Mousavian, José Díaz, and Ali Masoudi-Nejad. "Information theory in systems biology. Part II: Protein–protein interaction and signaling networks". In: *Seminars in cell & developmental biology*. Vol. 51. Elsevier. 2016, pp. 14–23.

[79] Christopher E Mason, Sandra G Porter, and Todd M Smith. "Characterizing multi-omic data in systems biology". In: *Systems Analysis of Human Multigene Disorders*. Springer, 2014, pp. 15–38.

[80] BFMSD Di Silvestre and P Mauri. "Evaluation of Proteomic Data: From Profiling to Network Analysis by Way of Biomarker Discovery". In: *Biomarker Validation, Technological, Clinical and Commercial Aspects*. Wiley-VCH Verlag GmbH & Co. KGaA, 2015.

[81] Francisco Azuaje, Yvan Devaux, and Daniel R Wagner. "Coordinated modular functionality and prognostic potential of a heart failure biomarker-driven interaction network". In: *BMC systems biology* 4.1 (2010), p. 60.

[82] Jiny Nair et al. "Network analysis of inflammatory genes and their transcriptional regulators in coronary artery disease". In: *PloS one* 9.4 (2014), e94328.

[83] Javier De Las Rivas and Celia Fontanillo. "Protein–protein interactions essentials: key concepts to building and analyzing interactome networks". In: *PLoS computational biology* 6.6 (2010), e1000807.

[84] Julian Mintseris and Zhiping Weng. "Structure, function, and evolution of transient and obligate protein–protein interactions". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.31 (2005), pp. 10930–10935.

[85] Emmanuel D Levy, Christian R Landry, and Stephen W Michnick. "How perfect can protein interactomes be". In: *Sci Signal* 2.60 (2009), e11.

[86] Ngounou Wetie et al. “Investigation of stable and transient protein–protein interactions: past, present, and future”. In: *Proteomics* 13.3-4 (2013), pp. 538–557.

[87] David La et al. “Predicting permanent and transient protein–protein interfaces”. In: *Proteins: Structure, Function, and Bioinformatics* 81.5 (2013), pp. 805–818.

[88] Arunachalam Vinayagam et al. “Integrating protein-protein interaction networks with phenotypes reveals signs of interactions”. In: *Nature methods* 11.1 (2014), pp. 94–99.

[89] Benjamin A Shoemaker and Anna R Panchenko. “Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners”. In: *PLoS computational biology* 3.4 (2007), e43.

[90] Arnaud Ceol et al. “MINT, the molecular interaction database: 2009 update”. In: *Nucleic acids research* 38.suppl_1 (2009), pp. D532–D539.

[91] Samuel Kerrien et al. “IntAct—open source resource for molecular interaction data”. In: *Nucleic acids research* 35.suppl_1 (2006), pp. D561–D565.

[92] Andrea Franceschini et al. “STRING v9. 1: protein-protein interaction networks, with increased coverage and integration”. In: *Nucleic acids research* 41.D1 (2012), pp. D808–D815.

[93] TS Keshava Prasad et al. “Human protein reference database—2009 update”. In: *Nucleic acids research* 37.suppl_1 (2008), pp. D767–D772.

[94] Gary D Bader, Michael P Cary, and Chris Sander. “Pathguide: a pathway resource list”. In: *Nucleic acids research* 34.suppl_1 (2006), pp. D504–D506.

[95] Rintaro Saito et al. “A travel guide to Cytoscape plugins”. In: *Nature methods* 9.11 (2012), pp. 1069–1076.

[96] Zhenjun Hu et al. “VisANT: an online visualization and analysis tool for biological interaction data”. In: *BMC bioinformatics* 5.1 (2004), p. 17.

[97] Yijun Ding et al. “atBioNet–an integrated network analysis tool for genomics and biomarker discovery”. In: *BMC genomics* 13.1 (2012), p. 325.

[98] Jianmin Wu et al. “Integrated network analysis platform for protein-protein interactions.” eng. In: *Nat Methods* 6.1 (2009), pp. 75–77.

[99] *QIAGEN’s Ingenuity Pathway Analysis*. URL: `https://www.ingenuity.com/`.

[100] Guanming Wu et al. “ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis”. In: *F1000Research* 3 (2014).

[101] Dong-Yeon Cho, Yoo-Ah Kim, and Teresa M Przytycka. "Network biology approach to complex diseases". In: *PLoS computational biology* 8.12 (2012), e1002820.

[102] Lin Song, Peter Langfelder, and Steve Horvath. "Comparison of co-expression measures: mutual information, correlation, and model based indices". In: *BMC bioinformatics* 13.1 (2012), p. 328.

[103] Cecily J Wolfe, Isaac S Kohane, and Atul J Butte. "Systematic survey reveals general applicability of" guilt-by-association" within gene coexpression networks". In: *BMC bioinformatics* 6.1 (2005), p. 227.

[104] Zhi-Ping Liu. "Reverse engineering of genome-wide gene regulatory networks from gene expression data". In: *Current genomics* 16.1 (2015), pp. 3–22.

[105] Tobias Maier, Marc Güell, and Luis Serrano. "Correlation of mRNA and protein in complex biological samples." eng. In: *FEBS Lett* 583.24 (2009), pp. 3966–3973.

[106] Bjoern Usadel et al. "Co-expression tools for plant biology: opportunities for hypothesis generation and caveats". In: *Plant, cell & environment* 32.12 (2009), pp. 1633–1651.

[107] Feng Luo et al. "Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory". In: *BMC bioinformatics* 8.1 (2007), p. 299.

[108] Laura L Elo et al. "Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process". In: *Bioinformatics* 23.16 (2007), pp. 2096–2103.

[109] Andrea Gobbi and Giuseppe Jurman. "A null model for Pearson coexpression networks". In: *PloS one* 10.6 (2015), e0128115.

[110] *ExpressionCorrelation*. URL: `http://www.baderlab.org/Software/ExpressionCorrelation`.

[111] Peter Langfelder and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis". In: *BMC bioinformatics* 9.1 (2008), p. 559.

[112] Yinghua Guo and Yonghua Xing. "Weighted gene co-expression network analysis of pneumocytes under exposure to a carcinogenic dose of chloroprene". In: *Life sciences* 151 (2016), pp. 339–347.

[113] Peter Langfelder and Steve Horvath. "Fast R functions for robust correlations and hierarchical clustering". In: *Journal of statistical software* 46.11 (2012).

[114] John D Storey and Robert Tibshirani. "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445.

[115] Bin Zhang and Steve Horvath. "A general framework for weighted gene co-expression network analysis". In: *Statistical applications in genetics and molecular biology* 4.1 (2005).

[116] Peter Langfelder and Steve Horvath. "Eigengene networks for studying the relationships between co-expression modules." eng. In: *BMC Syst Biol* 1 (2007), p. 54.

[117] Lei Nie et al. "Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications". In: *Critical reviews in biotechnology* 27.2 (2007), pp. 63–75.

[118] Lan Zhang et al. "Network-based proteomic analysis for postmenopausal osteoporosis in Caucasian females". In: *Proteomics* 16.1 (2016), pp. 12–28.

[119] Cosmin Lazar et al. "Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies". In: *Journal of proteome research* 15.4 (2016), pp. 1116–1125.

[120] Paulo C Carvalho et al. "Identifying differences in protein expression levels by spectral counting and feature selection". In: *Genetics and molecular research: GMR* 7.2 (2008), p. 342.

[121] Giovanni Scardoni, Michele Petterlini, and Carlo Laudanna. "Analyzing biological network parameters with CentiScaPe." eng. In: *Bioinformatics* 25.21 (2009), pp. 2857–2859.

[122] Huijuan Wang, Javier Martin Hernandez, and Piet Van Mieghem. "Betweenness centrality in a weighted network". In: *Physical Review E* 77.4 (2008), p. 046105.

[123] Xionglei He and Jianzhi Zhang. "Why do hubs tend to be essential in protein networks?" In: *PLoS genetics* 2.6 (2006), e88.

[124] M E J. Newman. "Modularity and community structure in networks." eng. In: *Proc Natl Acad Sci U S A* 103.23 (2006), pp. 8577–8582.

[125] D. J. Watts and S. H. Strogatz. "Collective dynamics of 'small-world' networks." eng. In: *Nature* 393.6684 (1998), pp. 440–442.

[126] R. Milo et al. "Network motifs: simple building blocks of complex networks." eng. In: *Science* 298.5594 (2002), pp. 824–827.

[127] Jianxin Wang et al. "Identification of essential proteins based on edge clustering coefficient". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9.4 (2012), pp. 1070–1080.

[128] Chia-Hao Chin et al. "cytoHubba: identifying hub objects and sub-networks from complex interactome". In: *BMC systems biology* 8.4 (2014), S11.

[129] Nadezhda T Doncheva et al. "Topological analysis and interactive visualization of biological networks and protein structures". In: *Nature protocols* 7.4 (2012), p. 670.

[130] Yu Tang et al. "CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks". In: *Biosystems* 127 (2015), pp. 67–72.

[131] *ELIXIR A distributed infrastructure for life-science information*. URL: `http://160.80.34.9/elixir2015/`.

[132] Josep Díaz et al. "Convergence theorems for some layout measures on random lattice and random geometric graphs". In: *Combinatorics, Probability and Computing* 9.6 (2000), pp. 489–511.

[133] Réka Albert, Hawoong Jeong, and Albert-László Barabási. "Error and attack tolerance of complex networks". In: *arXiv preprint cond-mat/0008064* (2000).

[134] Hawoong Jeong et al. "Lethality and centrality in protein networks". In: *arXiv preprint cond-mat/0105306* (2001).

[135] Jing-Dong J Han et al. "Effect of sampling on topology predictions of protein-protein interaction networks". In: *Nature biotechnology* 23.7 (2005), p. 839.

[136] N. Przulj, D. G. Corneil, and I. Jurisica. "Modeling interactome: scale-free or geometric?" eng. In: *Bioinformatics* 20.18 (2004), pp. 3508–3515.

[137] Natasa Przulj. "Biological network comparison using graphlet degree distribution." eng. In: *Bioinformatics* 23.2 (2007), e177–e183.

[138] Vuk Janjić and Nataša Pržulj. "The topology of the growing human interactome data". In: *Journal of Integrative Bioinformatics (JIB)* 11.2 (2014), pp. 27–42.

[139] Bader Al-Anzi et al. "Experimental and computational analysis of a large protein network that controls fat storage reveals the design principles of a signaling network". In: *PLoS computational biology* 11.5 (2015), e1004264.

[140] Jing-Dong J Han et al. "Evidence for dynamically organized modularity in the yeast protein–protein interaction network". In: *Nature* 430.6995 (2004), pp. 88–93.

[141] Günter P Wagner, Mihaela Pavlicev, and James M Cheverud. "The road to modularity". In: *Nature Reviews Genetics* 8.12 (2007), pp. 921–931.

[142] Zhi Wang and Jianzhi Zhang. "In search of the biological significance of modular structures in protein networks". In: *PLoS computational biology* 3.6 (2007), e107.

[143] Andre XCN Valente and Michael E Cusick. "Yeast Protein Interactome topology provides framework for coordinated-functionality". In: *Nucleic acids research* 34.9 (2006), pp. 2812–2819.

[144] Jingchun Chen and Bo Yuan. "Detecting functional modules in the yeast protein–protein interaction network". In: *Bioinformatics* 22.18 (2006), pp. 2283–2290.

[145] Sourav S Bhowmick and Boon Siew Seah. "Clustering and summarizing protein-protein interaction networks: a survey". In: *IEEE Transactions on Knowledge and Data Engineering* 28.3 (2016), pp. 638–658.

[146] M E J. Newman. "Fast algorithm for detecting community structure in networks." eng. In: *Phys Rev E Stat Nonlin Soft Matter Phys* 69.6 Pt 2 (2004), p. 066133.

[147] Luca Donetti and Miguel A Munoz. "Detecting network communities: a new systematic and efficient algorithm". In: *Journal of Statistical Mechanics: Theory and Experiment* 2004.10 (2004), P10012.

[148] Min Wu et al. "A core-attachment based method to detect protein complexes in PPI networks." eng. In: *BMC Bioinformatics* 10 (2009), p. 169.

[149] Balázs Adamcsek et al. "CFinder: locating cliques and overlapping modules in biological networks". In: *Bioinformatics* 22.8 (2006), pp. 1021–1023.

[150] Gary D Bader and Christopher WV Hogue. "An automated method for finding molecular complexes in large protein interaction networks". In: *BMC bioinformatics* 4.1 (2003), p. 2.

[151] Sourav S Bhowmick and Boon Siew Seah. "Clustering and summarizing protein-protein interaction networks: a survey". In: *IEEE Transactions on Knowledge and Data Engineering* 28.3 (2016), pp. 638–658.

[152] Jianxin Wang et al. "Recent advances in clustering methods for protein interaction networks". In: *BMC genomics* 11.Suppl 3 (2010), S10.

[153] Tamás Nepusz, Haiyuan Yu, and Alberto Paccanaro. "Detecting overlapping protein complexes in protein-protein interaction networks". In: *Nature methods* 9.5 (2012), pp. 471–472.

[154] S VAN Dongen. "Graph clustering by flow simulation". PhD thesis. Standardization and Knowledge Transfer, 2000.

[155] Shailesh Tripathi et al. "Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules". In: *BMC bioinformatics* 17.1 (2016), p. 129.

[156] Santo Fortunato. "Community detection in graphs". In: *Physics reports* 486.3 (2010), pp. 75–174.

[157] Junzhong Ji et al. "Survey: Functional module detection from protein-protein interaction networks". In: *IEEE Transactions on Knowledge and Data Engineering* 26.2 (2014), pp. 261–277.

[158] Zelmina Lubovac, Jonas Gamalielsson, and Björn Olsson. "Combining functional and topological properties to identify core modules in protein interaction networks". In: *Proteins: Structure, Function, and Bioinformatics* 64.4 (2006), pp. 948–959.

[159] Ioannis A Maraziotis, Konstantina Dimitrakopoulou, and Anastasios Bezerianos. "Growing functional modules from a seed protein via integration of protein interaction and gene expression data". In: *Bmc Bioinformatics* 8.1 (2007), p. 408.

[160] Sriganesh Srihari et al. "Methods for protein complex prediction and their contributions towards understanding the organisation, function and dynamics of complexes". In: *FEBS letters* 589.19 (2015), pp. 2590–2602.

[161] Jian Wang et al. "Filtering Gene Ontology semantic similarity for identifying protein complexes in large protein interaction networks". In: *Proteome Science*. Vol. 10. S1. BioMed Central. 2012, S18.

[162] Yijia Zhang et al. "Construction of ontology augmented networks for protein complex prediction". In: *PloS one* 8.5 (2013), e62077.

[163] Joseph R Nevins and Anil Potti. "Mining gene expression profiles: expression signatures as cancer phenotypes". In: *Nature Reviews Genetics* 8.8 (2007), pp. 601–609.

[164] Pierluigi Mauri et al. "Proteomics of bronchial biopsies: galectin-3 as a predictive biomarker of airway remodelling modulation in omalizumab-treated severe asthma patients". In: *Immunology letters* 162.1 (2014), pp. 2–10.

[165] Giampaolo Merlini and Vittorio Bellotti. "Molecular mechanisms of amyloidosis". In: *New England Journal of Medicine* 349.6 (2003), pp. 583–596.

[166] Giampaolo Merlini, David C Seldin, and Morie A Gertz. "Amyloidosis: pathogenesis and new therapeutic options". In: *Journal of Clinical Oncology* 29.14 (2011), pp. 1924–1933.

[167] Francesca Brambilla et al. "Reliable typing of systemic amyloidoses through proteomic analysis of subcutaneous adipose tissue". In: *Blood* (2011), blood–2011.

[168] Dario Di Silvestre, Francesca Brambilla, and Pier Luigi Mauri. "Multidimensional protein identification technology for direct-tissue proteomics of heart". In: *Heart Proteomics: Methods and Protocols* (2013), pp. 25–38.

[169] Noemi del Toro et al. "A new reference implementation of the PSICQUIC web service". In: *Nucleic acids research* 41.W1 (2013), W601–W606.

[170] PSICQUIC. URL: `http://psicquic.googlecode.com/`.

[171] David Croft et al. "The Reactome pathway knowledgebase". In: *Nucleic acids research* 42.D1 (2013), pp. D472–D477.

[172] Danila Vella et al. *MTopGO: a tool for module identification in PPI Networks*. Tech. rep. PeerJ Preprints, 2017.

[173] Wendy Weijia Soon, Manoj Hariharan, and Michael P Snyder. "High-throughput sequencing for biology and medicine". In: *Molecular systems biology* 9.1 (2013), p. 640.

[174] Tommy Nilsson et al. "Mass spectrometry in high-throughput proteomics: ready for the big time". In: *Nature methods* 7.9 (2010), pp. 681–685.

[175] Paul Shannon et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". In: *Genome research* 13.11 (2003), pp. 2498–2504.

[176] Steven Maere, Karel Heymans, and Martin Kuiper. "BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks". In: *Bioinformatics* 21.16 (2005), pp. 3448–3449.

[177] Dario Di Silvestre et al. "Proteomics-based network analysis characterizes biological processes and pathways activated by preconditioned mesenchymal stem cells in cardiac repair mechanisms". In: *Biochimica et Biophysica Acta (BBA)-General Subjects* 1861.5 (2017), pp. 1190–1199.

[178] Elli Papaemmanuil et al. "Genomic classification and prognosis in acute myeloid leukemia". In: *New England Journal of Medicine* 374.23 (2016), pp. 2209–2221.

[179] Gene Ontology Annotations. URL: `http://geneontology.org/page/download-annotations`.

[180] Clara Pizzuti and Simona E Rombo. "Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods". In: *Bioinformatics* 30.10 (2014), pp. 1343–1352.

[181] Xiaoli Li et al. "Computational approaches for detecting protein complexes from protein interaction networks: a survey". In: *BMC genomics* 11.1 (2010), S3.

[182] Nevan J Krogan et al. "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae". In: *Nature* 440.7084 (2006), p. 637.

[183] Anne-Claude Gavin et al. "Proteome survey reveals modularity of the yeast cell machinery". In: *Nature* 440.7084 (2006), p. 631.

[184] Sean R Collins et al. "Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae". In: *Molecular & Cellular Proteomics* 6.3 (2007), pp. 439–450.

[185] Ioannis Xenarios et al. "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions". In: *Nucleic acids research* 30.1 (2002), pp. 303–305.

[186] Anne-Claude Gingras, Ruedi Aebersold, and Brian Raught. "Advances in protein complex analysis using mass spectrometry". In: *The Journal of physiology* 563.1 (2005), pp. 11–21.

[187] Shuye Pu et al. "Up-to-date catalogues of yeast protein complexes". In: *Nucleic acids research* 37.3 (2008), pp. 825–831.

[188] Hans-Werner Mewes et al. "MIPS: analysis and annotation of proteins from whole genomes". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D41–D44.

[189] Eurie L Hong et al. "Gene Ontology annotations at SGD: new data sources and annotation methods". In: *Nucleic acids research* 36.suppl_1 (2007), pp. D577–D581.

[190] Andreas Ruepp et al. "CORUM: the comprehensive resource of mammalian protein complexes—2009". In: *Nucleic acids research* 38.suppl_1 (2009), pp. D497–D501.

[191] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. "Finding community structure in very large networks". In: *Physical review E* 70.6 (2004), p. 066111.

[192] Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. "An efficient algorithm for large-scale detection of protein families". In: *Nucleic acids research* 30.7 (2002), pp. 1575–1584.

[193] Quanzhong Liu, Jiangning Song, and Jinyan Li. "Using contrast patterns between true complexes and random subgraphs in PPI networks to predict unknown protein complexes". In: *Scientific reports* 6 (2016), p. 21223.

[194] Tom Ronan, Zhijie Qi, and Kristen M Naegle. "Avoiding common pitfalls when clustering biological data". In: *Sci. Signal.* 9.432 (2016), re6–re6.

[195] Ludmila I Kuncheva and Dmitry P Vetrov. "Evaluation of stability of k-means cluster ensembles with respect to random initialization". In: *IEEE transactions on pattern analysis and machine intelligence* 28.11 (2006), pp. 1798–1808.

[196] Danila Vella, Allan Tucker, and Riccardo Bellazzi. *Stability analysis of MTopGO for module identification in PPI networks*. Tech. rep. PeerJ Preprints, 2017.

[197] Douglas G Altman. *Practical statistics for medical research*. CRC press, 1990.

[198] Stephen Swift et al. "Consensus clustering and functional interpretation of gene-expression data". In: *Genome biology* 5.11 (2004), R94.

[199] Santo Fortunato and Marc Barthélemy. "Resolution limit in community detection". In: *Proceedings of the National Academy of Sciences* 104.1 (2007), pp. 36–41.

[200] Xiaoqing Peng et al. "Protein–protein interactions: detection, reliability assessment and applications". In: *Briefings in bioinformatics* (2016).

[201] Christian Von Mering et al. "Comparative assessment of large-scale data sets of protein-protein interactions". In: *Nature* 417.6887 (2002), p. 399.

[202] Kirill Tarassov et al. "An in vivo map of the yeast protein interactome". In: *Science* 320.5882 (2008), pp. 1465–1470.