Universidade de Aveiro Departamento de
2019 Eletrónica, Telecomunicações e Informática

Fernanda Maria dos Reis Brito e Rodrigues Correia

Previsão e Análise da Estrutura e Dinâmica de Redes Biológicas

Prediction and Analysis of Biological Networks Structure and Dynamics

Universidade de Aveiro Departamento de
2019 Eletrónica, Telecomunicações e Informática

Fernanda Maria dos Reis Brito e Rodrigues Correia

Previsão e Análise da Estrutura e Dinâmica de Redes Biológicas

Prediction and Analysis of Biological Networks Structure and Dynamics

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Informática, realizada sob a orientação científica do Professor Doutor José Luís Guimarães Oliveira, Professor Associado com Agregação do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro e do Professor Doutor Joel Perdiz Arrais, Professor Auxiliar do Departamento de Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

I dedicate this work to my parents and my children Renato and Miguel

o júri / the jury

presidente / president

António Manuel Rosa Pereira Caetano
Professor Catedrático da Universidade de Aveiro (por delegação do Reitor da Universidade de Aveiro)

vogais / examiners committee

Miguel Francisco Almeida Pereira Rocha
Professor Associado com Agregação do Departamento de Informática da Universidade da Minho

Sara Alexandra Cordeiro Madeira
Professora Associada do Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa

Rui Carlos Camacho de Sousa Ferreira da Silva
Professor Associado do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto

Sérgio Guilherme Aleixo de Matos
Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

José Luís Guimarães Oliveira
Professor Associado com Agregação do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro (orientador)

acknowledgements

I would like to thank my supervisors, Professor José Luís Oliveira and Professor Joel Arrais, for all the support and supervision during all these years, which were fundamental for the execution of this work. I would also like to thank Professor Sérgio Matos for the good research environment, together with the sharing of some scientific discussions, and Professor Carlos Costa also for the contribution to a good research environment. Finally I would like to thank the University of Aveiro for receving me during this research.

Palavras-chave

redes biológicas, redes de interação de proteinas, topologia de redes, dinâmica de redes, biologia computacional

Resumo

O conhecimento crescente sobre os processos biológicos que regem a dinâmica dos organismos vivos tem potenciado uma melhor compreensão da origem de muitas doenças, assim como a identificação de potenciais alvos terapêuticos. Os sistemas biológicos podem ser modelados através de redes biológicas, permitindo aplicar e explorar métodos da teoria de grafos na sua investigação e caracterização. Este trabalho teve como principal motivação a inferência de padrões e de regras que estão subjacentes à organização de redes biológicas.

Através da integração de diferentes tipos de dados, como a expressão de genes, interação entre proteínas e outros conceitos biomédicos, foram desenvolvidos métodos computacionais, para que possam ser usados na previsão e no estudo de doenças.

Como primeira contribuição, foi proposto um método de caracterização de um subsistema do interactoma de proteínas humano através das propriedades topológicas das redes que o modelam. Como segunda contribuição, foi utilizado um método não supervisionado que utiliza critérios biológicos e topologia de redes para, através de redes de co-expressão, melhorar a compreensão dos mecanismos genéticos e dos fatores de risco de uma doença. Como terceira contribuição, foi desenvolvida uma metodologia para remover ruído (*denoise*) em redes de proteínas, para obter modelos mais precisos, utilizando a topologia das redes. Como quarta contribuição, propôs-se uma metodologia supervisionada para modelar a dinâmica do interactoma de proteínas, usando exclusivamente a topologia das redes de interação de proteínas que fazem parte do modelo dinâmico do sistema.

As metodologias propostas contribuem para a criação de modelos biológicos, estáticos e dinâmicos, mais precisos, através da identificação e uso de padrões topológicos das redes de interação de proteínas, que podem ser usados na previsão e no estudo doenças.

Keywords

biological networks, protein interaction networks, network topology, network modeling, computational biology

Abstract

Increasing knowledge about the biological processes that govern the dynamics of living organisms has fostered a better understanding of the origin of many diseases as well as the identification of potential therapeutic targets. Biological systems can be modeled through biological networks, allowing to apply and explore methods of graph theory in their investigation and characterization. This work had as main motivation the inference of patterns and rules that underlie the organization of biological networks.

Through the integration of different types of data, such as gene expression, interaction between proteins and other biomedical concepts, computational methods have been developed so that they can be used to predict and study diseases.

The first contribution, was the characterization a subsystem of the human protein interactome through the topological properties of the networks that model it. As a second contribution, an unsupervised method using biological criteria and network topology was used to improve the understanding of the genetic mechanisms and risk factors of a disease through co-expression networks. As a third contribution, a methodology was developed to remove noise (denoise) in protein networks, to obtain more accurate models, using the network topology. As a fourth contribution, a supervised methodology was proposed to model the protein interactome dynamics, using exclusively the topology of protein interactions networks that are part of the dynamic model of the system.

The proposed methodologies contribute to the creation of more precise, static and dynamic biological models through the identification and use of topological patterns of protein interaction networks, which can be used to predict and study diseases.

List of contents

List of figures

List of tables

List of acronyms

A	adenine
AUC	area under the ROC curve
BETW	betweenness
BNL	Bayesian network learning
C	cytosine
CA	classification accuracy
CC	clustering coefficient
CDF	cumulative distribution function
CFM	collaborative filtering method
CRC	colorectal cancer
CS	combined score
DAVID	database for annotation, visualization, and integrated discovery
DNA	deoxyribonucleic acid
DT	decision tree
EPR	expression profile reliability
F1	F1-score
FN	false negative
FP	false positive
FPR	false positive rate
FS	features selection
G	guanine
GO	Gene Ontology
GRN	gene regulatory networks
GS	gene significance
GSEA	gene set enrichment analysis

HNSCC	head and neck squamous cell carcinoma
HRG	hierarchical random graph
HTS	high-throughput screening
HUPO-PSI	HUPO proteomics standards initiative
IGS	intrinsic geometry structure
I-Sim	inter-neighbourhood similarity
IMEx	International Molecular Exchange
JC	Jaccard
KEGG	Kyoto encyclopaedia of genes and genomes
KNN	k-nearest neighbours
KS	Kolmogorov Smirnov
MDS	multi-dimensional scaling-based
MIMIx	minimal information about a molecular interaction
MLE	maximum likelihood estimation
MM	module membership
mRNA	messenger RNA
MS	mass spectrometry
NB	naive Bayes
NC	neighborhood clustering
NGIS	next generation interaction sequencing
NGS	next generation sequencing
NN	neural network
NSCLC	non-small cell lung cancer
OM	organization measurement
OMIM	on-line mendelian inheritance in man database
OM THR	OM threshold
PCA	principal component analysis
PDF	probability density function
PLM	protein localization method
PPI	protein-protein interactions
PVM	paralogous verification method
RBF	radial basis function
RF	random forest
RFD	random forest decision

RNA	ribonucleic acid
RNA-Seq	RNA sequencing
ROC	receiver operating characteristic
SCLC	small cell lung cancer
SNPs	single nucleotide polymorphisms
SPM	structural perturbation method
SSN	sample series networks
STRING	search tool for the retrieval of interacting genes/proteins
SVD	singular value decomposition
SVM	support vector machines
T	thymine
TAP	tandem affinity purification
TF	transcription factors
TN	true negative
TP	true positive
TPR	true positive rate
tRNA	transfer RNA
U	uracil
UniProt	universal protein resource
WGCNA	weighted gene co-expression network analysis
Y2Y	yeast two-hybrid

Chapter 1

Introduction

Many real world systems, like the World Wide Web, on-line social communities, citations in scientific literature, communication networks and biological systems, have been studied, aiming a better understanding of their underlying mechanisms (e.g. elements, characteristics, behaviors, relationships). These systems rely on interconnected components and the study of these interactions contributes to the understanding of their processes and dynamics.

Networks are currently used to represent real word systems, because they can model the interactions between the entities of these systems. Graph mathematical theory is one of the methodologies that can be used to analyze the inherent topology of the networks. When real world systems are modeled through this approach, the complex networks that are obtained present topological properties that, typically, are not present in regular or random networks. The study of these complex networks allows the identification of patterns and signatures that can be related to relevant processes existent in these systems [1, 2].

Systems biology aims the integration of different types of biological data, obtained through the biological knowledge, and using experimental, like high throughput, and computational methods. This integration allows the study of static and dynamic behavior of the components of a cell or an organism, and the comparison between different species and states of the same species, to understand the various processes of these systems. The investigation of the complex processes underlying biological systems has been contributing to a deeper insight of the functioning of the living organisms and to the investigation of diseases and therapeutic targets, which have clinical implications and have been contributing to better health-care services and to the advance of other areas in the domain of bioengineering.

Most of the phenotype characteristics and the origin of many diseases can be studied through the genetic information of living systems. The use of biological networks to represent the different relationships between the bio-entities of these systems, allows the study of their structure and dynamics under various perspectives. Several network models have been proposed, being the two most well-known the small world networks, where most nodes can be reached from every other node by a small number steps, and scale-free networks, where their degree distribution follows a power-law. Small world networks tend to be very heterogeneous, containing some nodes with very high degree, i.e. connected to many neighbors, called hubs [3] and the majority of nodes with few connections.

In biological systems, most molecular interactions networks are disassortative, because their hubs have the tendency to link to nodes with fewer interaction partners rather than to other hubs [4]. These characteristics provide a high resistance to failures at random nodes [3], and so these networks are more robust to perturbations, like damage by mutation or viral infection than other network architectures.

1.1 Motivation

The study of biological systems involves the acquisition of biological data, which can be obtained through various experimental and computational techniques. Experimental techniques can be time-consuming and inaccurate because of the limitations of the technologies involved, so computational methods are required to assist in obtaining the missing information, more accurately and faster.

Despite the huge amount of biological data being produced, it is the capacity to keep such data in a structured way that will allow the study of biological systems represented by those data. For this, several models have been proposed to allow the study of the processes and mechanisms inherent to the interactions of the involved bio-entities.

Graph theory, from mathematics, can be used to model objects and the relationships between objects, providing a formal representation of networks. Biological systems can be modeled through this theory, allowing to study and understand them. Biological networks have an associated topology, where one can look for specific patterns or signatures that may be associated with several types of biological mechanisms. The complete knowledge of all these mechanisms and the understanding of their changes according to several parameters is a stepping stone to discover new therapeutic targets, to minimize the spread of diseases, and to cure diseases.

The quantification of several topological network descriptors, allow the characterization of these networks, in a static and dynamic way, and their comparison. The topological study of networks will allow the approximation of imprecise network models to more precise models of the real systems, making them better models, where the processes and mechanisms of these modeled organisms could be identified, better known and studied. In biological networks some motifs have been associated with optimized biological functions [5].

The noise associated to these biological network models, due to the not yet fully known processes of the biological systems and the inaccuracy of the used technologies, should be removed, to find the missing information and to eliminate the incorrect one. Thus these models have to be evaluated and corrected to allow their generalization, in order to be representative of the not completely known real biological networks [6].

Several researchers have been contributing to the understanding of biological functions of genes and proteins, of biological pathways and of the organization of cells, using the computation analysis in biological complex networks models that are representative of real biological systems. To have a more complete understanding of these systems, besides the static topological properties of these biological complex networks it is necessary to study their dynamics, because the interactions associated to biological processes can vary according to several factors associated to space, time and context.

The reverse engineering process or the inference of biological networks aims to construct network models from the observed data that could simulate the several states of biological occurrences.

Prediction methods in biological complex networks can combine different data sources, like protein-protein interactions (PPI), messenger RNA (mRNA) expression data, and other biological information. The prediction of missing nodes and edges in biological complex networks contributes to the finding of unknown bio-entities and interactions and to the assessment of the network reliability [7]. Node prediction in a complex network is challenging and is usually associated on finding important nodes related to critical transitions of biological systems, like nodes that bridge several network modules. Edge prediction in complex networks is associated on finding new interactions and new patterns, and is affected by several factors as data coverage, as well as the structure and dynamics of the network.

PPI networks and gene co-expression networks are examples of complex biological networks that are being studied by researchers. New information about cellular processes have been obtained through the study of the relationship between the similarity of the expression pattern of genes and the interaction of the proteins encoded by them. The identification of protein complexes and function modules from protein-protein interactions networks is important for the understanding of cellular organization and to predict protein functions [8]. A protein complex is defined [7] as a physical aggregation of several proteins via molecular interaction (binding) with each other at the same location and time and a functional module is defined [7] as a number of proteins that interact with each other to control or perform a particular cellular function and that do not necessarily interact at the same time and location. Gene Ontology (GO) [5] has been used to define functional modules. Also, gene expression data have been used to provide more accurate dynamic parameters in the protein-protein interaction networks representative of the biological systems.

The topological study of biological networks uses descriptors and quantifies them. These values allow to characterize topologically these networks both from the local and the global point of view. Understand the influence of the topology of biological networks in the identification of biological processes associated with diseases, such as cancer is still a subject not fully studied, but that has deserved the attention of researchers.

This study focus on using the networks topology to predict and analyze biological networks structure and dynamics. Obtaining better models using the topological characterization of the biological networks will allow the identification of patterns that can be used to denoise these network models that were created from the incomplete known biological data, obtained from the system to study, that will be a better representation of real biological processes and mechanisms. This knowledge will enable new findings related to biology and health science.

1.2 Aims

Biological networks, as a model to represent biological data information, can characterize relationships between bio-entities and have in their structure an embedded topology. Each biological network is a representation of a specific system under study, so the comparison of networks topologies can highlight similarities and dissimilarities of the represented data and dynamics of the systems modeled by those networks.

The domain of this research is enclosed in the study, exploration and analysis of life sciences and biological network inference fields.

The aim of this thesis is to investigate the hypothesis, that the integration of multiple sources of biological knowledge will disclosure patterns and rules in biological networks that allow the prediction of their structure and dynamics, which can be applied on the study of diseases.

Following this, some objectives were defined:

1) To explore the use of topological properties to characterize biological networks;

2) To propose a topology-based methodology to denoise protein interactions networks;

3) To investigate how to capture the dynamics of biological systems using a networks topology-based methodology;

4) To explore the use of networks topology-based models applied to the study of diseases.

In this thesis, the achievement of aim 1) is described, mainly, in the first part of Chapter 4, where the topological properties of a human proteome subsystem are studied using a network-based approach, but is transversal to all of the research. Aim 2) is accomplished through Chapter 5, where a new network topology-based denoising methodology and a new topological measure are proposed, being applied to protein interactions data. Aim 3) is mapped in Chapter 6, where a new methodology is proposed, to capture the dynamics of a biological system using a networks topology-based approach. Finally, aim 4) is reached in the second part of the Chapter 4 and in Chapter 6, where cancer diseases are studied.

1.3 Thesis Structure

This thesis is divided in seven chapters.

Chapter 1 is the present chapter that contextualizes the work behind this research. Starts with an introduction, the motivation and objectives, and ends with the structure of this document.

Chapter 2 gives an introduction to graph and network theory, with a description of a broad range of the topological properties of networks. It is followed by an introduction to network graphlets and motifs, and an introduction to power-law distributions. It finishes with a brief description of three classic network models.

Chapter 3 introduces the principles of molecular biology, followed with the properties of biological networks. Then different types of biological networks are mentioned, along with some platforms and databases containing respective data.

Chapter 4, in the first part, describes the research about the quantitative characterization of the networks, obtained from protein interactions of a subsystem of the human proteome, and, in the second part, describes the research about the use of co-expression networks to study a cancer disease using a network-based approach.

Chapter 5 describes a new network topology-based methodology proposed to denoise protein interactions networks.

Chapter 6 describes a supervised inference methodology, that uses a new approach to capture the dynamics of a biological system to predict cancer diseases. This approach uses the topology of a dynamic set of PPI.

Finally, Chapter 7 states the final remarks and possible future directions.

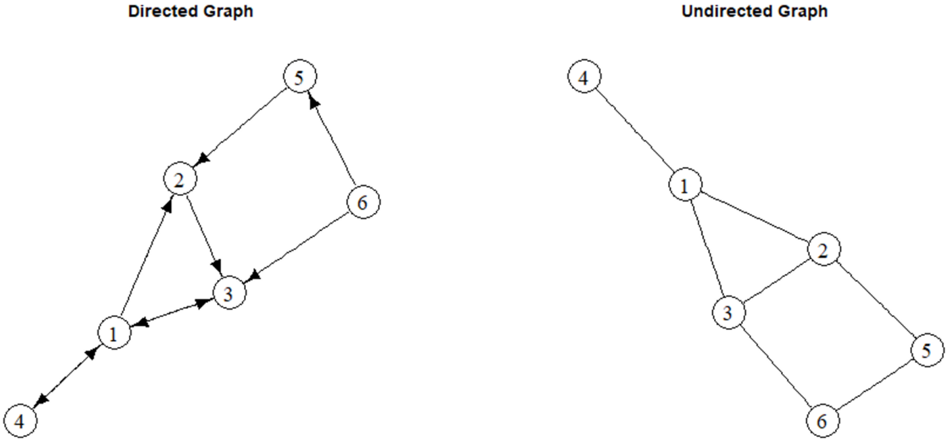
Chapter 2

Graph and Network Theory

This chapter includes a review of the main concepts of graph and network theory, from their mathematical representation, to the description of several topological properties. The most commonly used topological properties and other more specific topological properties, used in this research to characterize networks, are described. The following section introduces the concept of graphlets and presents three network topological similarities measures that use graphelets and compares this concept with the concept of motifs. The comparison of topological patterns or signatures in networks can be used to determine properties of some nodes, based on known properties of other nodes in the network, or to identify groups of nodes with topological similarity, that have meaning. Some applications in biological networks include the prediction of protein functions [9], the identification of cancer genes [10], and the discovery of pathways underlying certain biological processes or protein degradation [11]. It follows, a section with the formal definition and mathematical formulation of power-law distributions, a common concept, present in several real networks. Finally, this chapter includes a description of three complex network models, the Erdös-Rényi random network model, that reproduces well the small world property, the Wattz Strogatz model, that simulates a small world with high clustering coefficient network, and the Barabási–Albert model, that produces a network with a power-law distribution.

2.1 Graph and Network Representation

Graphs are a mathematical abstraction that can be used to model relationships between entities. A graph is composed by nodes (points or vertices), which are interconnected through edges (links, lines or arcs). Formally, a directed graph G is defined as an ordered triple G = (V, E, F) , where f is a function that maps each element in E (the set of edges) to an ordered pair of vertices in V (the set of nodes). An edge (i, j) $\in$ E has a direction from i to j and is called a direct arc or edge (see Figure 2.1). An undirected graph G can be defined as a pair G = (V, E) , where V is a set of vertices representing the nodes and E is a set of edges representing the connections, between the nodes i and j, defined as E = {{i, j}| i, j ∈ V}. A multi-edge connection consists of two or more edges that have the same endpoints. In directed and undirected graphs, it is usual to represent edges by (i, j) with the assumption that, in directed graphs, the edge (i, j) is different from the edge (j, i) and in undirected graphs, (i, j) and (j, i), are the same edge, since edges have no direction.



$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Figure 2.1: A directed graph and an undirected graph and respective adjacency matrices.

The number of nodes of a graph G is denoted by N, and the number of edges of a graph G is denoted by L. If G = (V, E) is a graph, then G_1 = (V_1, E_1) *is* called a subgraph if V_1 ⊆ V and E_1 ⊆ E and each edge in E_1 links vertices in V_1.

A walk can be defined as a pass through a specific sequence of nodes (i_1, i_2, ···, i_M) such that {(i_1, i_2),(i_2, i_3),···,(i_{M-1}, i_M)} ⊆ E. A trail is a walk where no edge can be repeated. A path is a trail where the first and the last nodes may be the same. A cycle is a walk (i_1, i_2, ···, i_M) where $i_1 = i_M$ with no other nodes repeated and M > 3. A graph is called cyclic, if it contains a cycle, or acyclic in not. If (i, j) is an edge in a graph G between nodes i and j, we say that the vertex i is adjacent to the vertex j. An undirected graph is connected if one can go from any node to any other node by following a sequence of edges. A directed graph is strongly connected if there is one directed path from any node to any other node [4, 12].

Given a graph G = (V, E) the adjacency matrix representation consists of a |V| × |V| =

N × N matrix A = [a$_{ij}$], such that a$_{ij}$= 1 if (i, j) ∈ E or a$_{ij}$ = 0 otherwise.

$$A = \begin{bmatrix} \mathrm{a}_{11} & \cdots & \mathrm{a}_{1\mathrm{N}} \\ \vdots & \ddots & \vdots \\ \mathrm{a}_{\mathrm{N}1} & \cdots & \mathrm{a}_{\mathrm{NN}} \end{bmatrix} \tag{2.1}$$

For a weighted graph a$_{ij}$= w$_{ij}$, if (i, j) ∈ E or a$_{ij}$ = 0 otherwise. For undirected graphs the matrix is symmetric because a$_{ij}$= a$_{ji}$.

A graph G = (V, E) can also be represented by an adjacency list, an array of elements, where for each (i, j) ∈ E, j ∈ V belongs to the list of i.

2.2 Topological Properties of the Networks

For the characterization of networks various topological properties can be applied. Some of these properties are more often used than others and are therefore better known. This section describes the topological properties generally used, and other specific topological properties of networks.

2.2.1 Topological Properties Generally Used

The topological properties generally used can be categorized as local and global. The equation of local properties are listed in Table 2.1 and the equations of global properties in Table 2.2.

In an undirected graph, the node degree deg (i) is the number of direct connections or edges the node has to other nodes.

The average degree or total connectivity, ⟨k⟩, of an undirected network G, is the average of the degree of the nodes of G.

If a network is directed, then each node has two different degrees, the in-degree $k_{in}(i)$, which is the number of incoming edges to node i, and the out-degree $k_{out}(i)$, which is the number of outgoing edges from node i.

An independent set in a graph is a subset of the vertices such that no pair of vertices from the two independent sets have an edge in the graph. A clique in an undirected graph G is a subgraph G$'$ which is complete (the degree equals to N-1). The size of a clique is the number of vertices it contains.

A cluster is a subset of vertices that contains edges connecting these vertices, and form a distinct group. The global clustering coefficient, C , that gives an indication of what is the clustering of the whole network, is the number of closed triplets (or 3 × n^{o} of triangles) over the total number of triplets (both open and closed). This measure can be applied to both undirected and directed networks [4].

The local clustering coefficient, C_i for a node i, is a notion of how connected the neighbors of the node i are (*cliquiness*). It measures the ratio of the number of edges between the neighbors of i and the total possible number of such edges. It takes values as $0 \leq \mathrm{C_i} \leq 1$.

The network average clustering coefficient, $\mathrm{C_{avg}}$, was first defined by Watts and Strogatz, as a way to characterize the overall tendency of nodes to form clusters or groups [4, 13, 14] and is the average of the local clustering coefficients of all nodes of the network.

The closer the average clustering coefficient is to 1, the more likely it is for the network to form clusters, showing a modular structure. It is higher than the average clustering coefficient of the random networks, where the average clustering coefficient C_{rand} can be obtained from the same properties of the considered non-random networks [15].

The length of a path is the number of edges forming it. There may be multiple paths connecting two given nodes. The shortest path length, $\mathrm{d\,(i,j)}$, between two nodes, $\mathrm{d\,(i,j)}$ $\mathrm{i,j \in V}$ is also called the distance between $\mathrm{i,j \in V}$.

The network diameter, denoted by d, is the largest distance between two nodes of G. It is a representation of the navigability of the networks [14, 16].

The average shortest path length, $\langle \mathrm{d} \rangle$, between i and any other node, is also known as the characteristic path length. It is defined to be the average value of $\mathrm{d\,(i,j)}$ taken over all pairs of distinct nodes, $\mathrm{i,j \in V}$ which are connected by at least one path. If i is an isolated node, the value is zero. It gives the expected distance between two connected nodes.

Two well know methods for calculating the shortest paths are Dijkstra's greedy algorithm and the Floyd's dynamic algorithm. The first one has a running time complexity of $\mathrm{O}\left(\mathrm{N}^2\right)$ and gives the shortest path between a source vertex i and all other vertices in the network. The second has running time complexity $\mathrm{O}\left(\mathrm{N}^3\right)$ and calculates an all-against-all matrix containing the distances of every node in the network to every other node.

The eccentricity of a node i in a connected network G is the maximum distance between i and any other node in the network. The maximum eccentricity is the network diameter. The minimum eccentricity is the network radius. Radiality is a node centrality index computed by subtracting the average shortest path length of a node i from the diameter of the connected component plus 1 (a number between 0 and 1) [16, 17].

A normalized version of the average number of neighbors $\langle \mathrm{k} \rangle$ is the total connectivity of a network or the density of a network, dens, that shows how sparse or dense a network is [13].

The density (dens) is a value between 0 and 1, and shows how densely the network is populated with edges. A network which contains no edges and solely isolated nodes has a density of 0. A sparse network is a graph where $L \sim \mathcal{O}(N)$. A network is dense if$L \sim \mathcal{O}(N^2)$.

A complete network is a network in which every pair of distinct vertices is connected by an edge. $\mathrm{N\,(N-1)}$ is the maximum number of edges a direct network can have and $\mathrm{N\,(N-1)/2}$ is the maximum number of edges an undirected network can have (with self-loops and duplicated edges ignored).

The network heterogeneity, h, reflects the tendency of a network to contain hub nodes (nodes with a degree greater than the average degree of the network) [17].

Centrality measures can give an insight of the importance of some nodes [4, 18]. One of these measures is the network centralization, cent.

Table 2.1: Summary of general local network descriptors.

Descriptors	Equation
Node degree of node *i* in undirected G	$\mathrm{deg}\,(\mathrm{i}) = \mathrm{k_i} = \sum_{\mathrm{j}\neq \mathrm{i}} \mathrm{a_{ij}} \qquad (2.2)$ k_i is the number of the neighbors of node i A = $[\mathrm{a}_{ij}]$ is the undirected network symmetric adjacency matrix.
Local clustering coefficient of node *i* in undirected G	$\mathrm{C_i} = \frac{2\sum_{\mathrm{l}\neq \mathrm{i}}\sum_{\mathrm{m}\neq \mathrm{i,l}} \mathrm{a_{il}a_{lm}a_{mi}}}{\left(\sum_{\mathrm{l}\neq \mathrm{j}} \mathrm{a_{il}}\right)^2 - \sum_{\mathrm{l}\neq \mathrm{i}} \mathrm{a_{il}}^2} = \frac{2\mathrm{L_k}}{\mathrm{k_i}\,(\mathrm{k_i}-1)} \qquad (2.3)$ k_i is the degree of i in an undirected graph G L_k is the number of edges existent between the *k* neighbors of *i* in *G*
Local clustering coefficient of node *i* in directed *G*	$\mathrm{C_i} = \frac{\sum_{\mathrm{l}\neq \mathrm{i}}\sum_{\mathrm{m}\neq \mathrm{i,l}} \mathrm{a_{il}a_{lm}a_{mi}}}{\left(\sum_{\mathrm{l}\neq \mathrm{j}} \mathrm{a_{il}}\right)^2 - \sum_{\mathrm{l}\neq \mathrm{i}} \mathrm{a_{il}}^2} = \frac{\mathrm{L_k}}{\mathrm{k_i}\,(\mathrm{k_i}-1)} \qquad (2.4)$ k_i is the degree of *i* in an undirected graph *G* L_k is the number of edges existent between the k neighbors of *i* in *G*
Average neighbor degree of node *i* in *G*	$k_{nn}(i) \quad = \quad \frac{1}{k_i}\sum_{j=1}^{N} a_{ij}k_j \qquad (2.5)$
Topological coefficient of node *i* in *G*	$\mathrm{T_i} = \frac{\mathrm{J\,(i,j)}}{\mathrm{k_i}} \qquad (2.6)$ $\mathrm{J\,(i,j)}$ is the number of neighbors shared between the nodes i and j, plus 1if there is a direct link between i and j $\mathrm{k_i}$ is the number of neighbors of i

Continued on next page

Table 2.1 – Summary of general local network descriptors (cont.).

Descriptors	Equation
Closeness centrality of node *i* in *G*	$$\mathrm{C_{clo}\left(i\right)=\frac{1}{\sum_{j\in V} d\left(i,j\right)}} \quad (2.7)$$ $\mathrm{d\left(i,j\right)}$ is the distance between the nodes i and j
Eigenvector centrality of node *i* in *G*	$$\mathrm{C_{eiv}(i) = \frac{1}{-}\sum_{k\in V} a_{ki}C_{eiv}(k)} \quad (2.8)$$
Betweenness centrality of node *w* in *G*	$$\mathrm{C_b\left(w\right)=\sum_{i\neq j\neq w\in V}\frac{\sigma_{ij}\left(w\right)}{\sigma_{ij}}} \quad (2.9)$$ σ_{ij} is the total number of shortest paths between i and j $\sigma_{\mathrm{ij}}\left(\mathrm{w}\right)$ the total number of shortest paths from i to j that pass through w $\mathrm{i,j,w\in V}$ are all distinct
Eccentricity centrality of node *i* in *G*	$$\mathrm{C_{ecc}(i) = \frac{1}{\max d\left(i,j\right)}} \quad (2.10)$$ $\mathrm{d\left(i,j\right)}$ is the shortest path between nodes i and j
Subgraph centrality of node *i* in *G*	$$\mathrm{C_{sg}(i) = \sum_{k=1}^{\infty}\frac{\left(A^k\right)_{ii}}{k!}} \quad (2.11)$$ A the adjacency matrix of G
Matching index between node *i* and node *j* in *G*	$$\mathrm{M_{ij}=\frac{\sum common\_neighbors}{\sum total\_number\_of\_neighbors}=\frac{\sum_{k,l}^{N} a_{ik}a_{jl}}{k_i+k_j-\sum_{k,l}^{N} a_{ik}a_{jl}}} \quad (2.12)$$

Continued on next page

Table 2.1 – Summary of general local network descriptors (cont.).

Descriptors	Equation
Assortativity coefficient between node *i* and node *j* in *G*	$$\mathrm{rs} = \frac{\mathrm{L}^{-1}\sum_{\mathrm{i,j}} \mathrm{k_i k_j} - \left[\mathrm{L}^{-1}\sum_{\mathrm{i,j}} \frac{1}{2}\left(\mathrm{k_i}+\mathrm{k_j}\right)\right]^2}{\mathrm{L}^{-1}\sum_{\mathrm{i,j}} \frac{1}{2}\left(\mathrm{k_i}^2+\mathrm{k_j}^2\right) - \left[\mathrm{L}^{-1}\sum_{\mathrm{i,j}} \frac{1}{2}\left(\mathrm{k_i}+\mathrm{k_j}\right)\right]^2} \qquad (2.13)$$ $(\mathrm{i,j}) = 1, \cdots, \mathrm{L}$ L is the number of edges $\mathrm{k_i}$ and $\mathrm{k_j}$ the degrees of the vertices at either ends of edge $(\mathrm{i,j})$

Networks with a topology in the form of a star have centralization close to 1, and decentralized networks, which are networks whether the nodes have on average the same connectivity, are characterized by having centralization close to 0.

The topological coefficient of a node i, $\mathrm{T_i}$, is a relative measure showing the extent to which a node shares neighbors with other nodes. Nodes that have one or no neighbors are assigned a topological coefficient of 0 (zero).

Closeness centrality, $\mathrm{C_{clo}}$, is a measure that can identify important nodes that can communicate quickly with other nodes of the network to spread information.

The betweenness centrality, $\mathrm{C_b}$, finds nodes that are intermediate between neighbors [19]. This measure favors nodes that join communities (dense subnetworks), rather than nodes that lie inside a community. Nodes with high betweenness lie on larger number of shortest paths of the network. The betweenness centrality of each node is a number between 0 and 1.

The eigenvector centrality, $\mathrm{C_{eiv}(i)}$, of node i from a graph $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ has a higher value if it is linked to by other important nodes.

The eccentricity centrality, $\mathrm{C_{ecc}}$, is the measure that shows how easily accessible a node is from other nodes.

The subgraph centrality, $\mathrm{C_{sg}}$, is the measure that ranks nodes according to the number of subgraphs of the overall network the node belongs, with more weight given to small subgraphs.

Two vertices that are functionally similar do not always have to be connected. The similarity of two nodes can be calculated by a matching index $\mathrm{M_{ij}}$, based on the number of common neighbors shared by nodes i and j. It is often used to cluster different components according to some property [4].

A network is called assortative if the vertices with high degree (hubs) have the tendency to connect with other vertices that also have high degree of connectivity (other hubs), like in social networks. If the vertices with higher degree have the tendency to connect with other vertices with low degree then the network is called disassortative. Biological networks are disassortative.

The assortativity coefficient, rs, is the measure of how assortative or disassortative a network is and is equivalent to the Pearson's correlation coefficient (PCC) of the degrees at either ends of an edge. It is defined as the covariance of the two nodes divided by the product of their standard deviations. The range of the rs values is between +1 and −1 and if rs < 0 the network is disassortative and if rs > 0 the network is assortative.

Another way to correlate degrees [4] is to calculate the average neighbor degree, $\mathrm{k_{nn}(i)}$. Usually is compared to the average neighbor degree of all nodes of degree k, $\mathrm{k_{nn\_random}(k)}$ for a random network.

Table 2.2: Summary of general global network descriptors.

Descriptors	Equation
Average degree of G	$$< k >= \frac{2 * L}{N} \qquad (2.14)$$ $2 * \mathrm{L} = \sum_{\mathrm{i} \in \mathrm{V}} \mathrm{k_i}$ L, N is the number of nodes and edges respectively of an undirected G
Global clustering coefficient of G	$$\mathrm{C} = \frac{3 \times \text{number of triangles}}{\text{number of connected triples of nodes}} = \\ = \frac{\text{number of closed triplets}}{\text{number of connected triples of nodes}} \qquad (2.15)$$
Average clustering coefficient of G	$$\mathrm{C_{avg}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{C_i} \qquad (2.16)$$
Average clustering coefficient of a random G	$$\mathrm{C_{rand}} = \frac{1}{\mathrm{N}} \frac{\left(\mathrm{k}^2 - \mathrm{k}\right)^2}{\mathrm{k}^3} \qquad (2.17)$$
Average neighbour degre of all nodes of degree k of a random G	$$\mathrm{k_{nn\_random}}\left(\mathrm{k}\right) = \frac{\mathrm{k}^2}{\mathrm{k}} \qquad (2.18)$$

Continued on next page

Table 2.2 – Summary of general global network descriptors (cont.).

Descriptors	Equation
Network diameter of *G*	$$\mathrm{d} = \frac{1}{\mathrm{N}}\left(\max_{\mathrm{i,j}} \mathrm{d}\left(\mathrm{i,j}\right)\right) \tag{2.19}$$
Characteristic path length of *G*	$$\langle \mathrm{d} \rangle = \frac{2}{\mathrm{N}\left(\mathrm{N}-1\right)} \sum_{\mathrm{i}=1}^{\mathrm{N}} \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{d}\left(\mathrm{i,j}\right) \tag{2.20}$$
Density of an directed *G*	$$\mathrm{dens} = \frac{\mathrm{L}}{\mathrm{N}\left(\mathrm{N}-1\right)} \tag{2.21}$$
Density of an undirected *G*	$$\mathrm{dens} = \frac{\sum_{\mathrm{i}} \sum_{\mathrm{j}\neq\mathrm{i}} \mathrm{a_{ij}}}{\mathrm{N}\left(\mathrm{N}-1\right)} = \frac{\mathrm{k}}{\mathrm{N}-1} = \frac{2\mathrm{L}}{\mathrm{N}\left(\mathrm{N}-1\right)} \tag{2.22}$$
Heterogeneity of *G*	$$\mathrm{h} = \frac{\sqrt{\mathrm{variance}\left(\mathrm{k}\right)}}{\mathrm{k}} = \frac{\sqrt{\mathrm{k}^2 - \mathrm{k}^2}}{\mathrm{k}} \tag{2.23}$$
Centralization of *G*	$$\mathrm{cent} = \frac{\mathrm{N}}{\mathrm{N}-2}\left(\frac{\max\left(\mathrm{k}\right)}{\mathrm{N}-1} - \mathrm{Density}\right) \approx \frac{\max\left(\mathrm{k}\right)}{\mathrm{N}} - \mathrm{density} \tag{2.24}$$

2.2.2 Specific Network Properties

There are other relevant topological network descriptors that can be used on the analysis of networks, for example biological networks, like gene and protein-protein interaction networks. We can divide these descriptors in three categories [19], namely: 1) descriptors based on distances in a graph; 2) descriptors based on other graph invariants; 3) and more recent graph complexity measures [20]. A complete list of these descriptors formulation is presented in Table 2.3, Table 2.4 and Table 2.5.

The first class uses node distances to describe the networks structure. The Wiener Index, W (G), was introduced by Wiener in 1947, to study the correlations between the boiling points of paraffin and their molecular structure [21, 22].

Table 2.3: Summary of distance-based network descriptors.

Descriptors	Equation	Ref
Wiener	$$\mathrm{W}(\mathrm{G}) = \frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{d}(\mathrm{i},\mathrm{j}) \qquad (2.25)$$ d(i, j) is the shortest distances between i, j∈V	D1.1
Hararay	$$\mathrm{H}(\mathrm{G}) = \frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}} (\mathrm{d}(\mathrm{i},\mathrm{j}))^{-1},\ \mathrm{i}\neq\mathrm{j} \qquad (2.26)$$	D1.2
Balaban J	$$\mathrm{J}(\mathrm{G}) = \frac{\mathrm{L}}{\mu+1}\sum_{(\mathrm{i},\mathrm{j})\in\mathrm{E}} [\mathrm{DS}_\mathrm{i}\mathrm{DS}_\mathrm{j}]^{-\frac{1}{2}} \qquad (2.27)$$ DS_i is the distance sum (row sum) of i (the sum of the distances of node i to the other nodes) $\mu = \|\mathrm{E}\| + 1 - \mathrm{N}$ is the cyclomatic number for one connected component	D1.3
Dobrynin mean distance deviation	$$\Delta\mathrm{D}(\mathrm{G}) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}\in\mathrm{V}} \Delta\mathrm{D}(\mathrm{i}) \qquad (2.28)$$ $\Delta\mathrm{D}(\mathrm{i}) = \|\mathrm{D}(\mathrm{i}) - \mathrm{D}_{\mathrm{av}}(\mathrm{G})\|$is the Dobrynin distance vertex deviation (from average) $\mathrm{D}(\mathrm{i}) = \sum_{\mathrm{i}\in\mathrm{V}} \mathrm{d}(\mathrm{i},\ \mathrm{j})$ is the Dobrynin vertex centrality	D1.16
Dobrynin average distance of graph vertices	$$\mathrm{D}_{\mathrm{av}}(\mathrm{G}) = \frac{2\mathrm{D}(\mathrm{G})}{\mathrm{N}} \qquad (2.29)$$	D1.15
Compacteness	$$\mathrm{C}(\mathrm{G}) = \frac{4\mathrm{W}(\mathrm{G})}{\mathrm{N}(\mathrm{N}-1)} \qquad (2.30)$$	D1.5

Continued on next page

Table 2.3 – Summary of distance-based network descriptors (cont.).

Descriptors	Equation	Ref
Product of Row Sum	$$\mathrm{PRS}\,(\mathrm{G}) = \prod_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{d}\,(\mathrm{i},\mathrm{j}) = \prod_{\mathrm{i}=1}^{\mathrm{N}} \mu\,(\mathrm{i}) or \\ \log\,(\mathrm{PRS}\,(\mathrm{G})) = \log(\prod_{\mathrm{i}=1}^{\mathrm{N}} \mu(\mathrm{i})) \quad (2.31)$$	D1.6
Hyper-distance-path	$$\mathrm{D_P}\,(\mathrm{G}) = \frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}} \mathrm{d}(\mathrm{i},\mathrm{j}) + \frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{N}} \binom{\mathrm{d}(\mathrm{i},\mathrm{j})}{2} \quad (2.32)$$	D1.7
Dobrynin eccentricity	$$\mathrm{e}\,(\mathrm{G}) = \sum_{\mathrm{i}\in\mathrm{V}} \mathrm{e}\,(\mathrm{i}) \quad (2.33)$$ $\mathrm{e}\,(\mathrm{i}) = \max_{\mathrm{j}\in\mathrm{V}} \mathrm{d}\,(\mathrm{i},\mathrm{j})$ is the Dobrynin vertex eccentricity	D1.8
Dobrynin average vertex eccentricity of a graph G	$$\mathrm{e_{av}}\,(\mathrm{G}) = \frac{\mathrm{e}\,(\mathrm{G})}{\mathrm{N}} \quad (2.34)$$	D1.9
Dobrynin eccentric	$$\Delta\mathrm{G} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}\in\mathrm{V}} \Delta\mathrm{e}\,(\mathrm{i}) \quad (2.35)$$ $\Delta\mathrm{e}\,(\mathrm{i}) = \lvert\mathrm{e}\,(\mathrm{i}) - \mathrm{e_{av}}\,(\mathrm{G})\rvert$is the Dobrynin vertex centrality	D1.10
Dobrynin graph integration or distance of a graph	$$\mathrm{D}(\mathrm{G}) = \frac{1}{2}\sum_{\mathrm{i}\in\mathrm{V}} \mathrm{D}\,(\mathrm{i}) \quad (2.36)$$ $\mathrm{D}(\mathrm{i}) = \sum_{\mathrm{i}\in\mathrm{V}} \mathrm{d}\,(\mathrm{i},\mathrm{j})$is the Dobrynin vertex centrality	D1.11
Dobrynin unipolarity or minimal distance	$$\mathrm{D}^*\,(\mathrm{G}) = \min_{\mathrm{i}\in\mathrm{V}} \mathrm{D}\,(\mathrm{i}) \quad (2.37)$$	D1.12

Continued on next page

Table 2.3 – Summary of distance-based network descriptors (cont.).

Descriptors	Equation	Ref
Dobrynin variation	$$\mathrm{var}\left(G\right) = \max_{i \in V} \Delta D^{*}\left(i\right) \tag{2.38}$$ $\Delta D^{*}\left(i\right) = D\left(i\right) - D^{*}\left(G\right)$ is the Dobrynin distance vertex deviation (from its minimum)	D1.13
Dobrynin centralization	$$\Delta G^{*} = \sum_{i \in V} \Delta D^{*}\left(i\right) \tag{2.39}$$	D1.14

Table 2.4: Summary of other invariants-based network descriptors.

Descriptors	Equation	Ref
Total adjacency	$$A\left(G\right) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} \tag{2.40}$$	D2.1
Zagreb 1	$$Z_1\left(G\right) = \sum_{i \in V} k_i \tag{2.41}$$	D2.2
Zagreb 2	$$Z_2\left(G\right) = \sum_{(i,j) \in E}^{N} k_i k_j \tag{2.42}$$	D2.3
Modified Zagreb	$$MZI\left(G\right) = \sum_{(i,j) \in E}^{N} \frac{1}{k_i k_j} \tag{2.43}$$	D2.4
Augmented Zagreb	$$AZI\left(G\right) = \int_{(i,j) \in E}^{N} \left(\frac{k_i k_j}{k_i + k_j - 2}\right)^{3} \tag{2.44}$$	D2.5

Continued on next page

Table 2.4 – Summary of other invariants-based network descriptors (cont.).

Descriptors	Equation	Ref
Variable Zagreb	$$\mathrm{VZI}\left(\mathrm{G}\right)=\sum_{(\mathrm{i,j})\in \mathrm{E}}^{\mathrm{N}}\frac{\mathrm{k_i}+\mathrm{k_j}-2}{\mathrm{k_i k_j}} \qquad (2.45)$$	D2.6
Randic connectivity	$$\mathrm{R}\left(\mathrm{G}\right)=\sum_{(\mathrm{i,j})\in \mathrm{E}}^{\mathrm{N}}\left[\mathrm{k_i k_j}\right]^{-\frac{1}{2}} \qquad (2.46)$$	D2.7
Complexity *B*	$$\mathrm{B}\left(\mathrm{G}\right)=\sum_{\mathrm{i}=1}^{\mathrm{N}}\frac{\mathrm{k_i}}{\mu\left(\mathrm{i}\right)} \qquad (2.47)$$	D2.8
Normalized edge complexity	$$\mathrm{E_N}\left(\mathrm{G}\right)=\frac{\mathrm{A(G)}}{\mathrm{N}^2} \qquad (2.48)$$	D2.9
Atom-bond connectivity	$$\mathrm{ABC}\left(\mathrm{G}\right)=\sum_{(\mathrm{i,j})\in \mathrm{E}}^{\mathrm{N}}\sqrt{\frac{\mathrm{k_i}+\mathrm{k_j}-2}{\mathrm{k_i k_j}}} \qquad (2.49)$$	D2.10
Geometric-arithmetic 1	$$\mathrm{GA1}\left(\mathrm{G}\right)=\sum_{(\mathrm{i,j})\in \mathrm{E}}^{\mathrm{N}}\frac{\sqrt{\mathrm{k_i k_j}}}{\frac{1}{2}\left(\mathrm{k_i}+\mathrm{k_j}\right)} \qquad (2.50)$$	D2.11
Geometric-arithmetic 2	$$\mathrm{GA2}\left(\mathrm{G}\right)=\sum_{(\mathrm{i,j})\in \mathrm{E}}^{\mathrm{N}}\frac{\sqrt{\mathrm{n_i n_j}}}{\frac{1}{2}\left(\mathrm{n_i}+\mathrm{n_j}\right)} \qquad (2.51)$$ $\mathrm{n_i}=\left\|\mathrm{x}\in\mathrm{V}:\mathrm{d}\left(\mathrm{x,i}\right)<\mathrm{d}\left(\mathrm{x,j}\right)\right\|$ $\mathrm{n_j}=\left\|\mathrm{x}\in\mathrm{V}:\mathrm{d}\left(\mathrm{x,j}\right)<\mathrm{d}\left(\mathrm{x,i}\right)\right\|$	D2.12

Continued on next page

Table 2.4 – Summary of other invariants-based network descriptors (cont.).

Descriptors	Equation	Ref
Geometric-arithmetic 3	$$\mathrm{GA3}\left(G\right)=\sum_{(i,j)\in E}^{N}\frac{\sqrt{m_i m_j}}{\frac{1}{2}\left(m_i+m_j\right)} \qquad (2.52)$$ $m_i= \left\|\{f\in E : d\left(f,i\right) < d\left(f,j\right)\{\right\|$ $m_j= \left\|\{f\in E : d\left(x,j\right) < d\left(x,i\right)\}\right\|$ $d\left(f,v\right)=\min\left\{d\left(x,v\right),d\left(y,v\right)\right\}$ is the distance between an edge $f=\{x,y\}$ and a vertex v	D2.13
Narumi-Katayama	$$\mathrm{NK}=\sum_{i=1}^{N}k_i \qquad (2.53)$$	D2.14

Other examples of distance measures are the Hararay index H(G), the Balaban J index $J\left(G\right)$ that allows measuring the average distance sum connectivity. The $J\left(G\right)$ uses the cyclomatic number, also known as circuit rank, that is the minimum number of edges that must be removed from the graph to break all its cycles. In this category can also be included the measures like the compactness, $C\left(G\right)$, the product of row sum, $PRS\left(G\right)$, the hyper-distance-path [25], $D_P\left(G\right)$ and the Skorobogatov and Dobrynin descriptors.

The descriptors based on other graph invariants use characteristics other than distances, such as degree, number of nodes, number of edges, and others. Examples are the index of total adjacency A(G), the Zagreb indices, or the geometric-arithmetic indices (see Table 2.4).

Finally, some more recent graph complexity measures have been tested [20], which are based in product and entropy measures. The objective is to have complexity measures to differentiate between graphs with the same number of nodes and edges. Higher values of these measures indicate important modular structures. Many real networks are small world, so they have short path lengths (high efficiency) with the number of links not too high (cost function). So, these measures can be used to define the efficiency complexity of a graph. In sparse networks, when links are added, efficiency increases faster than cost and an optimum is found in a graph with medium number of links.

The Graph index complexity $Cr\left(G\right)$ is one of those measurements. Another complexity measure is the Medium articulation $MAg\left(G\right)$ (see Table 2.5). A clique has $\frac{N(N-1)}{2}$ edges and has the highest redundancy R_{clique} and the lowest mutual information I_{clique}. A path has the lowest redundancy R_{path} and the highest mutual information I_{path}. Considering this, the product $(R\left(G\right)-R_{\mathrm{path}}\left(G\right))(I\left(G\right)-I_{\mathrm{clique}}(G))$ is zero in extreme cases (clique and path) and positive otherwise. $\mathrm{MA_R}\left(\mathrm{G}\right)$ and $\mathrm{MA_I}\left(\mathrm{G}\right)$ have values between 0 and 1. The product $\mathrm{MAg}\left(\mathrm{G}\right)$ is used, because $\mathrm{MA_R}\left(\mathrm{G}\right)$ discriminates worse than $\mathrm{MA_R}\left(\mathrm{G}\right)$ between graphs with the same L and N and because $\mathrm{MA_I}\left(\mathrm{G}\right)$ has a too high value for very sparse graphs.

Table 2.5: Summary of other more recent network descriptors.

Descriptors	Equation	Ref
Medium articulation	$\mathrm{MAg}\left(\mathrm{G}\right)=\mathrm{MA_R}\left(\mathrm{G}\right).\mathrm{MA_I}\left(\mathrm{G}\right)$ (2.54) $\mathrm{MA_R}\left(\mathrm{G}\right)=4\left(\frac{\mathrm{R(G)-R_{path}(G)}}{\mathrm{R_{clique}(G)-R_{path}(G)}}\right)\left(1-\frac{\mathrm{R(G)-R_{path}(G)}}{\mathrm{R_{clique}(G)-R_{path}(G)}}\right)$is the redundancy $\mathrm{R}\left(\mathrm{G}\right)=\frac{1}{\mathrm{L}}\sum_{\mathrm{i,j>i}}\log\left(\mathrm{k_ik_j}\right)$ $\mathrm{R_{clique}}\left(\mathrm{G}\right)=2\log\left(\mathrm{N}-1\right)$ $\mathrm{R_{path}}\left(\mathrm{G}\right)=2\frac{\mathrm{N-2}}{\mathrm{N-1}}\log 2$ $\mathrm{MA_I}\left(\mathrm{G}\right)=4\left(\frac{\mathrm{I(G)-I_{clique}(G)}}{\mathrm{I_{path}(G)-I_{clique}(G)}}\right)\left(1-\frac{\mathrm{I(G)-I_{clique}(G)}}{\mathrm{I_{path}(G)-I_{clique}(G)}}\right)$ is the mutual information $\mathrm{I}\left(\mathrm{G}\right)=\frac{1}{\mathrm{L}}\sum_{\mathrm{i,j>i}}\log\left(\frac{2\mathrm{L}}{\mathrm{k_ik_j}}\right)$ $\mathrm{I_{clique}}\left(\mathrm{G}\right)=\log\frac{\mathrm{N}}{\mathrm{N-1}}$ $\mathrm{I_{path}}\left(\mathrm{G}\right)=\log\left(\mathrm{N}-1\right)-\frac{\mathrm{N-3}}{\mathrm{N-1}}\log 2$	D3.1
Efficiency	$\mathrm{Ce}\left(\mathrm{G}\right)=4\left(\frac{\mathrm{E}\left(\mathrm{G}\right)-\mathrm{E_{path}(G)}}{1-\mathrm{E_{path}(G)}}\right)\left(1-\frac{\mathrm{E}\left(\mathrm{G}\right)-\mathrm{E_{path}(G)}}{1-\mathrm{E_{path}(G)}}\right)$ (2.55) $\mathrm{E}\left(\mathrm{G}\right)=\frac{2}{\mathrm{N(N-1)}}\sum_{\mathrm{i}}\sum_{\mathrm{j>i}}\frac{1}{\mathrm{d(i,j)}}$ $\mathrm{E_{path}}\left(\mathrm{G}\right)=\frac{2}{\mathrm{N(N-1)}}\sum_{\mathrm{i=1}}\mathrm{N}-1\frac{\mathrm{N-i}}{\mathrm{i}}$	D3.2
Graph complexity	$\mathrm{Cr}\left(\mathrm{G}\right)=4\mathrm{c_r}\left(1-\mathrm{c_r}\right)$ (2.56) $\mathrm{c_r}=\frac{\mathrm{r}-2\cos\frac{\pi}{\mathrm{N+1}}}{\mathrm{N}-1-2\cos\frac{\pi}{\mathrm{N+1}}}$ r is the largest of the real eigenvalues of G	D3.3

Continued on next page

Table 2.5 – Summary of other more recent network descriptors (cont.).

Descriptors	Equation	Ref
Offdiagonal	$\mathrm{OdC}(\mathrm{G}) = -\frac{1}{\log(\mathrm{N}-1)} \sum_{\mathrm{n}=0}^{\mathrm{k_{max}}-1} \tilde{\mathrm{a}}_{\mathrm{n}} \log \tilde{\mathrm{a}}_{\mathrm{n}} \qquad (2.57)$ $\tilde{\mathrm{a}}_{\mathrm{n}} = \frac{\mathrm{a_n}}{\sum_{\mathrm{n}=0}^{\mathrm{k_{max}}-1} \mathrm{a_n}}$ $\mathrm{a_n} = \sum_{\mathrm{i}=1}^{\mathrm{k_{max}}-\mathrm{N}} \mathrm{c}_{\mathrm{k_1},\mathrm{k_1}+\mathrm{N}}$ $\mathrm{k_{max}}$ is the maximum degree of all nodes in G $\mathrm{c_{k_1k_2}}$ is the number of all neighbours with degree $\mathrm{k_2} \geq \mathrm{k_1}$ of all nodes with degree $\mathrm{k_1}$	D3.4
Spanning tree sensitivity	$\mathrm{STS}(\mathrm{G}) = \frac{\mathrm{H}(\{\mathrm{S_{ij}}\})}{\mathrm{logm_{cu}}} \qquad (2.58)$ $\mathrm{H}(\{\mathrm{S_{ij}}\}) = -\sum_{\mathrm{l}} \mathrm{a_{llog a_l}}$ is the entropy of the list $\left\{\mathrm{S_{ij}^1}, \mathrm{S_{ij}^2}, \ldots, \mathrm{S_{ij}^k}\right\}$ for all different $\mathrm{s_{ij}}$ for all $\mathrm{k} \leq \mathrm{L}$ $\mathrm{S_{ij}} = \mathrm{s_{ij}} - (\min\{\mathrm{s_{ij}}\} - 1)$ because the distribution of the entropy of the list are not equally distributed $\mathrm{s_{ij}} > 0$ is the sensitivity and is the number of spanning trees in the graph minus the number of spanning trees of the subgraph with the edge (i, j) deleted $\mathrm{a_l} = \frac{\mathrm{S_{ij}^l}}{\sum_{\mathrm{r}}^{\mathrm{k}} \mathrm{S_{ij}^r}}$ $\mathrm{m_{cu}} = \mathrm{N}^{1.68} - 10$ is a normalization factor, is an upper bound for the number of edges of the most complex graph for a given number of nodes	D3.5
Spanning tree sensitivity differences	$\mathrm{STSD}(\mathrm{G}) = \frac{\mathrm{H}(\mathrm{Ld})}{\mathrm{logm_{cu}}} \qquad (2.59)$ $\mathrm{H}(\mathrm{Ld})$ is the entropy defined similarly to $\mathrm{H}(\{\mathrm{S_{ij}}\})$ $\mathrm{Ld} = \mathrm{Ld_1}, \mathrm{Ld_2}, \ldots, \mathrm{Ld_d}$ contains the $\mathrm{d} \leq \mathrm{k} - 1$ different entries of L $\mathrm{L} = \left\{\mathrm{S_{ij}^2} - \mathrm{S_{ij}^1}, \ldots, \mathrm{S_{ij}^k} - \mathrm{S_{ij}^{k-1}}\right\}$ is a new list constructed from the next list $\left\{\mathrm{S_{ij}^1} < \mathrm{S_{ij}^2} < \ldots, < \mathrm{S_{ij}^k}\right\}$ is an ordered list of $\mathrm{k} \leq \mathrm{L}$ different sensitivities	D3.6

The off-diagonal complexity, OdC(G), measures the diversity of values of the complex graph G in the node-node link correlation matrix $[c_{k_1 k_2}]$ and is high if the nodes of a given degree of G is similar to the degree of its neighbors.

The Spanning tree sensitivity, STS(G), is another complexity descriptor, with $0 \leq STS < 1$, which measures the number of spanning trees in the graph minus the number of deleted spanning trees of a subgraph. In simple graphs, like clique or star, all links play the same role and have the same sensitivity. All links essential to keep the graph connected have the same sensitivity. Trees have always zero spanning tree sensitivity. Similarly, there is also the descriptor spanning tree sensitivity differences, STSD(G). Comparing two graphs with the same number of different S_{ij}, according to these two last measures, the more complex graph is the one that has a more homogeneous distribution.

2.3 Graphlets and Motifs

Graphlets and motifs are two different concepts based on subgraphs of a network. Graphlets are small connected non-isomorphic induced subgraphs [23, 24], so they must contain all of the edges between its nodes, that are present in the large network. Motifs are partial subgraphs and can contain only some of them, but need to be over represented in the network compared to the randomized versions of the same network [25].

Graphlets nodes are differentiated by their topological equivalence. Pržulj defined 73 non-equivalent types of nodes, designated by orbits, and organized in 30 subgraphs [23, 24]. Three network topological similarities measures were defined using graphlets: 1) the relative graphlet frequency distance (RGF-distance) between two networks, $D(G,H)$ [26]; 2) the graphlet degree distribution agreement (GDD-agreement) between two networks G and H, the $A_{arith}(G,H)$ or the $A_{geo}(G,H)$ [22]; and 3) the graphlet degree vector (GDV) or the graphlet degree signature similarity $S(u,v)$ between two nodes u and v [9].

The RGF-distance $D(G,H)$ between two networks G and H compares the 3-5 node graphlets relative frequencies F_i, for $i = 1, \ldots 29$ orbits.

$$D(G,H) = \sum_{i=1}^{29} |F_i(G) - F_i(H)| \tag{2.60}$$

Different graphlets can have a large amplitude of frequencies, so to avoid that the most frequent graphlet influences more the distance, it is applied the logarithm to minimize this amplitude.

$$F_i(G) = -\log(N_i(G)/T(G)) \tag{2.61}$$

$$\mathrm{T}\left(\mathrm{G}\right)=\sum_{\mathrm{i}=1}^{29}\mathrm{N_i}\left(\mathrm{G}\right) \tag{2.62}$$

where $\mathrm{N_i}\left(\mathrm{G}\right)$ is the number of graphlets of type i, i$\in\{1,\ldots,29\}$ in the network G, and $\mathrm{T}\left(\mathrm{G}\right)$ is the total number of graphlets of G.

The GDD-agreement [24] between two networks G and H is a generalization of degree distribution. It measures for each k, the number of nodes touching k of each of the 30 graphlets $\mathrm{G_0},\mathrm{G_1},\ldots,\mathrm{G_{29}}$, in each of the 73 orbits.

The GDD-agreement between two networks G and H is the arithmetic or the geometric average of the $\mathrm{j^{th}}$ GDD-agreements over all j orbits [24]

$$\mathrm{A_{arith}}\left(\mathrm{G},\mathrm{H}\right)=\frac{1}{73}\sum_{\mathrm{j}=0}^{72}\mathrm{A^j}(\mathrm{G},\mathrm{H}) \tag{2.63}$$

$$\mathrm{A_{geo}}\left(\mathrm{G},\mathrm{H}\right)=\left(\prod_{\mathrm{j}=0}^{72}\mathrm{A^j}(\mathrm{G},\mathrm{H})\right)^{\frac{1}{73}} \tag{2.64}$$

where

$$\mathrm{A^j}\left(\mathrm{G},\mathrm{H}\right)=1-\mathrm{D^j}\left(\mathrm{G},\mathrm{H}\right) \tag{2.65}$$

The distance $\mathrm{D^j}\left(\mathrm{G},\mathrm{H}\right)$ is the Euclidean distance of the two normalized $\mathrm{j^{th}}$ distributions vectors of G and H and is between 0 and 1, where 0 means that G and H have similar $\mathrm{j^{th}}$ GDD.

$$\mathrm{D^j}\left(\mathrm{G},\mathrm{H}\right)=\frac{1}{\sqrt{2}}\left(\sum_{\mathrm{k}=1}^{\infty}\left[\mathrm{N_G^j}\left(\mathrm{k}\right)-\mathrm{N_H^j}\left(\mathrm{k}\right)\right]^2\right)^{\frac{1}{2}} \tag{2.66}$$

$\mathrm{N_G^j}\left(\mathrm{k}\right)$ is the normalization of the distribution $\mathrm{S_G^j}\left(\mathrm{k}\right)$ of the graph G to force the distributions to have a total area under the curve of 1, before they are compared

$$\mathrm{N_G^j}\left(\mathrm{k}\right)=\frac{\mathrm{S_G^j}(\mathrm{k})}{\mathrm{T_G^j}} \tag{2.67}$$

where $\mathrm{S_G^j}\left(\mathrm{k}\right)$ are the scaled $\mathrm{d_G^j}\left(\mathrm{k}\right)$, which are the number of nodes in G touching the corresponding graphlet at orbit j, k times.

$$\mathrm{S_G^j}\left(\mathrm{k}\right)=\frac{\mathrm{d_G^j}(\mathrm{k})}{\mathrm{k}} \tag{2.68}$$

The graphlet degree vector GDV, or graphlet degree signature, counts the number of graphlets each node touches a particular orbit, for all graphlets on 2 to 5 nodes [9]. The resulting vector of 73 coordinates is the signature of a node that characterizes the topology of its neighbourhood to a distance of 4. The graphlet degree signatures similarity, $\mathrm{S(u,v)}$, between nodes u and v gives an indication of the local topological similarity of u and v and can be calculated by:

$$\mathrm{S(u,v)} = 1 - \mathrm{D(u,v)} \tag{2.69}$$

where $\mathrm{D\,(u,v)}$ is the total distance between nodes u and v and is in the interval [0, 1). Distance 0 means that signatures of nodes u and v are identical.

$$\mathrm{D\,(u,v)} = \frac{\sum_{\mathrm{i=0}}^{72} \mathrm{D_i(u,v)}}{\sum_{\mathrm{i=0}}^{72} \mathrm{w_i(u,v)}} \tag{2.70}$$

$\mathrm{D_i\,(u,v)}$ is the distance between the $\mathrm{i^{th}}$ orbits of nodes u and v and $\mathrm{u_i}$ denotes the $\mathrm{i^{th}}$ coordinate of the signature vector of the node u belonging to the graph G, i.e. $\mathrm{u_i}$ is the number of times node u is touched by an orbit i in G.

$$\mathrm{D_i\,(u,v)} = \mathrm{w_i} \times \frac{|\mathrm{log u_i{+}1{-}log v_i{+}1}\ \ |}{\mathrm{log\,(max\,\{u_i, v_i\}\ {+}2)}} \tag{2.71}$$

Motifs need to be over represented in the network compared to the randomized versions of the same network, while graphlets do not have that need.

Motifs are subgraphs that occur significantly more frequently in the real network, as compared to the random networks, with the same number of nodes, edges, and degree distribution of the real network. Identifying motifs is a way to uncover topological patterns in complex networks and reflects functional properties of the network [27]. They are network specific, but there are families of networks that have the same series of motifs.

To characterize such families, it can be used a vector whose $\mathrm{i^{th}}$ entry quantifies the importance of the $\mathrm{i^{th}}$ motif with respect to the other motifs of the network [28]. For each subgraph i, the statistical significance is described by the z-score:

$$Z_i = \frac{Nreal_i - < Nrand_i >}{std(Nrand_i)} \tag{2.72}$$

where $Nreal_i$ is the number of times a subgraph of type i appears in the network, $< Nrand_i >$ is the mean of its appearances in the randomized network ensemble and $std(Nrand_i)$ is the standard deviation of its appearances in the randomized network ensemble.

The significance profile vector (SP) is the vector of z-scores normalized to length 1:

$$SP_i = \frac{Z_i}{\sqrt{\sum_i {Z_i}^2}} \tag{2.73}$$

The normalization emphasizes the relative significance of subgraphs, rather than the absolute significance, which is important for comparison of networks of different sizes [28].

2.4 Power-law Distributions

Mathematically, a quantity x obeys a power-law if it satisfies the probability distribution $\mathrm{p}\left(\mathrm{x}\right) \propto \mathrm{x}^{-\alpha}$, where α is a constant parameter of the distribution, known as the exponent or scaling parameter [29, 30].

There are continuous power-law distributions, with real values, and discrete power-law distributions, where the quantity of interest can take only a discrete set of values, normally positive integers.

A continuous power-law distribution is described by a probability density function (PDF) $\mathrm{p}\left(\mathrm{x}\right)$. There must be some lower bound $\mathrm{x_{min}}$ to the power-law behavior.

In the discrete case and in the case of integer values, x can take only a discrete set of integer values with a probability distribution $\mathrm{p}\left(\mathrm{x}\right)$. This distribution diverges at zero, so there must be a lower bound $\mathrm{x_{min}}> 0$ for the power-law behaviour.

The mathematical expressions of the continuous and discrete power-law distributions can be seen in Table 2.6.

The complementary cumulative distribution function (CDF) of a power-law distributed variable, $\mathrm{P}\left(\mathrm{x}\right)$ is, for the continuous and discrete cases, defined as $\mathrm{P}\left(\mathrm{x}\right) = \mathrm{Pr}\left(\mathrm{X}\geq\mathrm{x}\right)$ and their mathematical expressions can also be seen in Table 2.6.

Discrete power-laws can be approximated by its continuous equivalent, considering that the values of x were generated from a continuous power-law rounded to the nearest integer.

The scaling parameter α and the lower-bound of the scaling area $\mathrm{x_{min}}$ can be estimated, from the fitting of power-law forms to empirical distributions. Taking the logarithm of both sides of power-law equation, we have a straight line [29].

$$\ln(\mathrm{p}\left(\mathrm{x}\right)) = \alpha\ \ln(\mathrm{x})\ \ + \mathrm{k} \tag{2.74}$$

The CDF follows a power-law with an exponent $\alpha{-}1$.

Assuming that data follow a power-law for $\mathrm{x}\geq\mathrm{x_{min}}$ the maximum likelihood estimation (MLE) of the scaling parameter, $\widehat{\alpha}$ and the respective standard error σ, can be calculated (see Table 2.6).

To quantify the distance between two probability distributions, for non-normal data, it can be used the Kolmogorov Smirnov (KS) statistic. It is calculated by the maximum distance $\mathrm{D_{max}}$ between the CDFs of the data and of the fitted model, as

$$\mathrm{D_{max}} = \max_{\mathrm{x}\geq\mathrm{x_{min}}} \left|\mathrm{S}\left(\mathrm{x}\right) - \mathrm{P}\left(\mathrm{x}\right)\right| \tag{2.75}$$

where $\mathrm{S}\left(\mathrm{x}\right)$ is the CDF of the data for the observations with value at least $\mathrm{x_{min}}$, and $\mathrm{P}\left(\mathrm{x}\right)$ is the CDF for the power-law model that best fits the data in the region $\mathrm{x}\geq\mathrm{x_{min}}$. The estimate $\widehat{\mathrm{x}}_{\mathrm{min}}$ is then the value of $\mathrm{x_{min}}$ that minimizes $\mathrm{D_{max}}$.

The KS statistic is insensitive to differences between distributions at the extreme limits of the range of x. In these limits the CDFs tend to zero and one. To avoid this problem, $\mathrm{D_{max}}$ ca be re-weighted as

$$\mathrm{D_{max}}= \max_{\mathrm{x}\geq \mathrm{x_{min}}} \frac{|\mathrm{S}\left(\mathrm{x}\right) - \mathrm{P}\left(\mathrm{x}\right)|}{\sqrt{\mathrm{P}\left(\mathrm{x}\right)\left(1 - \mathrm{P}\left(\mathrm{x}\right)\right)}} \tag{2.76}$$

Table 2.6: Power-law distributions density functions, parameters and error estimation for continuous and discrete cases.

Power-law	Continuous	Discrete
PDF	$\begin{aligned} \mathrm{p}\left(\mathrm{x}\right)\mathrm{dx} &= \\ = \mathrm{P_r}\left(\mathrm{x}\leq \mathrm{X}\leq \mathrm{x+dx}\right) &= \\ &= \mathrm{Cx}^{-\alpha}\mathrm{dx} \end{aligned} \quad (2.77)$ X is the observed value C is a normalization constant $\mathrm{p}\left(\mathrm{x}\right) = \frac{\alpha - 1}{\mathrm{x_{min}}}\left(\frac{\mathrm{x}}{\mathrm{x_{min}}}\right)^{-\alpha}$ $\mathrm{x_{min}}$ is the lower bound to the power-law behaviour For $\alpha > 1$, $\mathrm{C} = \left(\alpha - 1\right)\mathrm{x}_{\mathrm{min}}^{\alpha - 1}$	$\mathrm{p}\left(\mathrm{x}\right) = \mathrm{Pr}\left(\mathrm{X} = \mathrm{x}\right) = \mathrm{Cx}^{-\alpha} \quad (2.78)$ x can take only a discrete set of integer values $\mathrm{p}\left(\mathrm{x}\right) = \frac{\mathrm{x}^{-\alpha}}{\zeta(\alpha,\mathrm{x_{min}})}$ $\mathrm{C} = \frac{1}{\zeta(\alpha,\mathrm{x_{min}})}$ $\mathrm{x_{min}} > 0$ is the lower bound to the power-law behaviour $\zeta\left(\alpha, \mathrm{x_{min}}\right) = \sum_{\mathrm{n}=0}^{\infty} \mathrm{n+x_{min}}^{-\alpha}$ is the Hurwitz zeta function
CDF	$\mathrm{P}\left(\mathrm{x}\right) = \left(\frac{\mathrm{x}}{\mathrm{x_{min}}}\right)^{-\left(\alpha - 1\right)} \quad (2.79)$	$\mathrm{P}\left(\mathrm{x}\right) = \frac{\zeta\left(\alpha, \mathrm{x}\right)}{\zeta\left(\alpha, \mathrm{x_{min}}\right)} \quad (2.80)$

Continued on next page

Table 2.6 – Summary of other more recent network descriptors (cont.).

Power-law	Continuous	Discrete
MLE of α	$\widehat{\alpha}= 1 + \mathrm{n}\left[\sum_{\mathrm{i}=1}^{\mathrm{n}} \ln\frac{\mathrm{x_i}}{\mathrm{x_{min}}}\right]^{-1}$ (2.81) $\mathrm{x_i}$, $\mathrm{i} = 1\cdots\mathrm{n}$ are the observed values of i such that $\mathrm{x_i}\geq\mathrm{x_{min}}$	$\frac{\zeta'(\widehat{\alpha}, \mathrm{x_{min}})}{\zeta(\widehat{\alpha}, \mathrm{x_{min}})}= -\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\ln\mathrm{x_i}$ (2.82) or $\mathcal{L}(\alpha) = = -\mathrm{n}\ln(\zeta(\alpha,\mathrm{x_{min}})) -\alpha\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{x_i}$ (2.83) $\mathrm{x_{min}}> 0$ is the lower bound for the power-law behaviour $\widehat{\alpha} = 1+ +n\left[\sum_{\mathrm{i}=1}^{\mathrm{n}}\ln\frac{\mathrm{x_i}}{\mathrm{x_{min}}-\frac{1}{2}}\right]^{-1}$ (2.84) approximated by its continuous equivalent with $\mathrm{x_i}$ rounded to the nearest integer
Standard error on $\widehat{\alpha}$	$\sigma=\frac{\widehat{\alpha}-1}{\sqrt{\mathrm{n}}}+\mathrm{O}(1/\mathrm{n})$ (2.85) higher-order correction is positive	$\sigma=\frac{1}{\sqrt{\mathrm{n}\left[\frac{\zeta''(\widehat{\alpha},\mathrm{x_{min}})}{\zeta(\widehat{\alpha},\mathrm{x_{min}})}-\left(\frac{\zeta'(\widehat{\alpha},\mathrm{x_{min}})}{\zeta(\widehat{\alpha},\mathrm{x_{min}})}\right)^{2}\right]}}$ (2.86)

The bootstrap method can be used to determine the uncertainty of $\widehat{\mathrm{x}}_{\mathrm{min}}$. $\mathrm{x_{min}}$ and α can be estimated from a synthetic data set, generated with a similar distribution to the original, by drawing a new sequence of points $\mathrm{x_i}$, $\mathrm{i} = 1,\cdots,\mathrm{n}$, uniformly at random from the original data, with n the number of measurements. Then the standard deviation of these estimates over a large number of repetitions (1000) of this process can be derived from the original estimated parameters.

Finally, the fact that the PDF or CDF distribution of data, in a log-log plot, is approximately straight, is not a sufficient condition of being a power-law. A goodness-of-fit test, which generates a p-value, should be used to know if the hypothesis is acceptable, given the data. In this test, it is calculated the distance between the distribution of the empirical data and the hypothesized model that will be compared with distance measurements for comparable synthetic data sets drawn from the same model. The p-value is defined as the fraction of the synthetic distances that is larger than the empirical distance and, if the p-value is large (close to 1), then the difference between the empirical data and the model can be attributed to statistical fluctuations alone; if it is small (usually $p < 0.1$), the model is not a plausible fit to the data, and the power-law hypothesis can be rejected. Even in this case there may be other distributions that match the data equally well or better. So, when n is small, the goodness-of-fit test can be used with other distributions to compare p-values [29, 30].

There are other methods, besides the KS test, which can compare two distributions, like the likelihood ratio test, fully Bayesian approaches, cross-validation, or minimum description length (MDL).

2.5 Network Models

To study biological systems using network based approaches, there are various network models that can be used.

Real networks have an inherent structure different from random networks where there is no defined structure. Random networks are often used in the comparison with real networks. [4, 31]. Random networks are also called Erdös-Rényi networks, because these two mathematicians gave a high contribute to understand the properties of these networks. They have low clustering and are characterized as small world, where the characteristic path length follows $<d> \sim \log(N)$. One other model, proposed by Watts and Strogatz, generates networks with small world properties and high clustering.

Scale-free networks are networks where their degree distribution follows a power-law, which is scale invariant, i.e., inversely proportional to a degree exponent α. Several power-law distributions have been found in the network representations of different domains, like in physics, biology, social sciences and economic systems. A scale-free network can be constructed by progressively adding nodes to an existing network and introducing links to existing nodes with preferential attachment, so that the probability of linking to a given node i is proportional to the number of existing links k_i that the node has. The Barabási-Albert network model generates scale-free networks with high clustering, but without modularity, since the clustering coefficient does not depend on nodes degree.

Hierarchical networks have highly clustered areas, where communication is done by few nodes highly connected, called hubs. They are scale-free networks with a high clustering that, whose distribution follow a power law of degree -1, C(k) $\sim k^{-1}$.

Next sections give a brief description of the Erdös-Rényi model, the Watts and Strogatz model and the Barabási-Albert models.

2.5.1 Erdös-Rényi

The Erdös-Rényi model has the properties of a random graph. A random network can be defined by a G(N, L) model, where N labeled nodes are connected with L random links or by a G(N, p) model, where each pair of N labeled nodes is connected with probability p [32]. In Figure 2.2 is shown an example of a random network generated from the Erdös-Rényi model, with 15 nodes 0.2 of probability of edge creation (on left) and the log-log plot of its degree distribution (on right).

This network model is obtained by taking a number of N vertices and connecting nodes by selecting undirected edges E from the $N(N-1)/2$ possible edges randomly (excluding multiple and self-edges). The probability of two random vertices to be connected is given by

$$p = \frac{2L}{N(N-1)} \tag{2.87}$$

The degree distribution of a random network is a binomial distribution, with the average degree of the network, <k>, equals to, p (N-1). Since real networks are sparse, with N >><k>, when N→∞ (large networks) their degree distribution is well approximated by a Poisson distribution, which is a distribution that has only one parameter, <k>. So, in a random network the probability of a node to have degree k or the degree distribution of a random network is given by

$$p\,(k) \approx e^{-\langle k\rangle} \frac{\langle k\rangle^{k}}{k!} \tag{2.88}$$

The Erdös-Rényi random network has low heterogeneity, so it is considered homogeneous, with most vertices with similar degree, distributed around the <k> and its clustering coefficient is C_{rand}= p= <k>/N, meaning that the probability that two verties with a common neighbor are connected equals the probability that any pair of randomly chosen vertices are connected. The average clustering coefficient of a random network can also be calculated with the same properties of the correspondent non-random networks, for comparison. In a random network, the local clustering coefficient is independent of the node's degree and depends on the system size as 1/N.

An Erdös-Rényi random network has low clustering, and the average distance between two nodes, or average path length, depends on log(N), $\langle d\rangle$ =log(N)/log($\langle k\rangle$), the small world property, where <d>∼log(N).

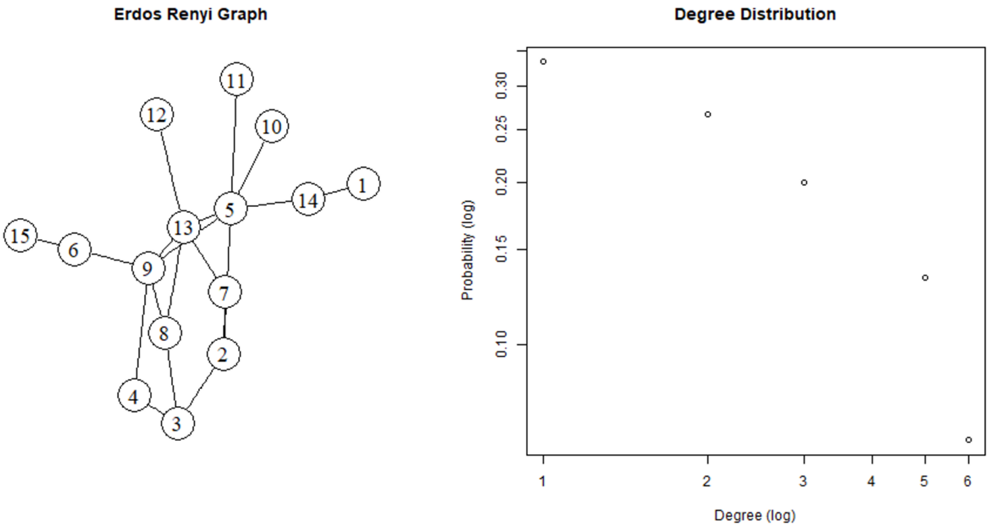


Figure 2.2: A random network, generated from the Erdös-Rényi model, with 15 nodes, 0.2 of probability of edge creation and respective log-log plot of degree distribution.

2.5.2 Watts and Strogatz

Most real-world networks are not homogeneous, because the number of connections between each node varies greatly. The Watts and Strogatz model (Figure 2.3) was introduced to describe networks that reproduces both, small world like in random networks and a higher clustering coefficient than random networks [14].

In this model, the probability of a node to have degree k is a Poisson like distribution, the average path length is constant and the characteristic path length is $\langle \mathrm{d} \rangle = \log(\mathrm{N})$.

Figure 2.3 shows an example of a random network generated from the Watts-Strogatz model with 15 nodes, 5 nearest neighbors in ring topology, 0.2 of probability of rewiring each edge (on left) and the log-log plot of its degree distribution (on right).

2.5.3 Barabási-Albert

A random network has comparable degrees and the average degree <k> is the scale of the network. A scale-free network does not have a scale, has a highly heterogeneous and its degree distribution follows a power-law. Many real networks show a degree distribution that significantly deviates from Poisson distribution, having a power-law tail. A scale-free network has a few highly connected nodes, called hubs and a lot of nodes with a small degree.

In real networks, the number of nodes continually grows thanks to the addition of new nodes and nodes prefer to link to the nodes with more neighbors.

The Barabási-Albert network model is another network model that has a scale-free degree distribution, but which is based in growth, which means that the number of the vertices of

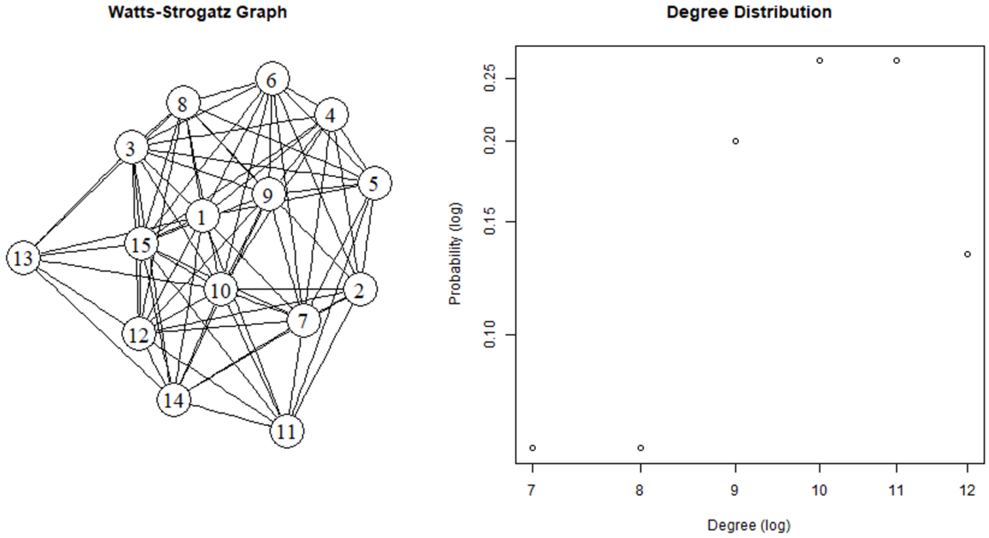


Figure 2.3: A small world network, generated from the Watts-Strogatz model with 15 nodes, 5 nearest neighbors in ring topology and 0.2 of probability of rewiring each edge and respective log-log plot of degree distribution.

the network is not constant as in previous models and preferential attachment.

Next figure (figure 2.4) shows an example network, which is a scale-free network, generated from the Barabási-Albert model, with 15 nodes (on left) and the log-log plot of its degree distribution (on right).

Preferential attachment means that new edges are not randomly introduced and that the probability of a vertex i receiving a new edge depends on its degree $\mathrm{k_i}$. Preferential attachment applies the concept “rich get richer”, where new nodes attach preferentially to nodes that are already well connected.

Initially, it starts with a small number of nodes $\mathrm{N_0}$. At each step, a new node is added and gets linked to the existing network. The degree distribution of a network generated by the Barabási-Albert model follows a power-law distribution of α= 3

$$p(k) \quad \sim \quad k^{-3} \tag{2.89}$$

The network characteristic path length that is created by the Barabási–Albert model (see Figure 2.4) is shorter than in random networks, following

$$\langle \mathrm{d} \rangle \sim \log \log \mathrm{(N)} \tag{2.90}$$

which characterizes ultra-small world networks.

The clustering coefficient in the Barabási–Albert network model, decreases with the

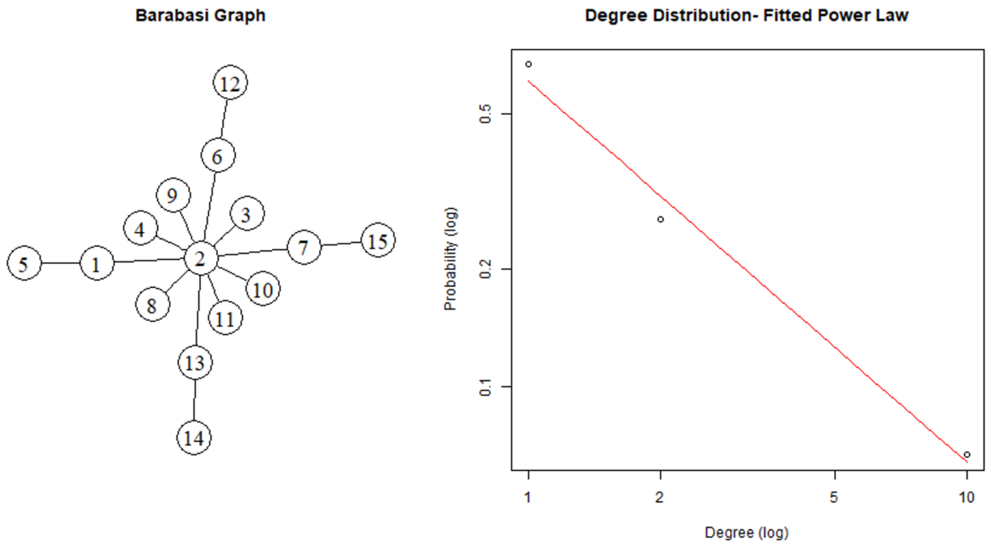


Figure 2.4: A scale-free network, generated from the Barabási-Albert model, with 15 nodes and respective log-log plot of degree distribution.

network size, $\mathrm{C} \sim \mathrm{N}^{-0.75}$, , distinct from small world networks, where clustering coefficient is independent of the network size. They are independent of the node degree k, distinct from hierarchical networks, where clustering is a function of the node degree, following a power-law $C(k) = k^{-1}$ [16].

There are several extensions to the Barabasi-Albert model to create models for different degree exponents, different cluster coefficients, different degree correlations, different evolutionary growth processes and using rewiring the edges according to some rule. A survey of statistical network models can be found in [33, 34].

2.6 Summary

This chapter started with the mathematical definition of a network. Then, general and specific topological properties used to characterize networks were defined. To compare the topological structure of real networks and to distinguish them, graphlets and motifs can be used to find meaningful patterns. Also, since many real networks follow a power-law, the definition of power-law in the continuous and discrete case was provided, as well as how to calculate its parameters, and how to evaluate if a PDF or a CDF follow a power law. The last section provides a brief introduction of three well known models, the Erdös-Rényi random model, the Watts and Strogatz model and the Barabasi-Albert model.

Chapter 3

Biological Networks

The comprehension of biological processes is an important step to understand the functioning of organisms at a system level. Several scientific areas, such as biology, medicine, mathematics and engineering, contribute to the study of these processes, complementing each other, through the knowledge acquired from different perspectives. Biological processes can be viewed as systems where multiple biological entities interact. One way of representing these systems is through networks, allowing the construction of models that represent the inherent structure and dynamics of their biological processes.

This chapter begins with a brief introduction to molecular biology, followed by a description of the properties of biological networks. A summary of various types of biological networks is presented, with a greater focus on two types of biological networks, protein interaction networks and gene co-expression networks, which were the most used in the developed research.

3.1 Principles of Molecular Biology

The basis of biological inheritance is the deoxyribonucleic acid (DNA) replication. DNA is made of a double helix of two complementary strands, where each strand of DNA is a chain of four types of nucleotides that correspond to four nucleobases, adenine (A), cytosine (C), guanine (G) and thymine (T). During replication, these strands separate, and each one serves as a template to originate two double helix of DNA.

The central dogma of molecular biology, stated by Francis Crick [35], describes the mechanism of protein synthesis, the flow of genetic information from DNA to mRNA, the ribonucleic acid (RNA) transcription, and from mRNA to the protein, the translation. RNA is transcribed in the nucleus of the cell and then transported to the cytoplasm and translated by the ribosome in eukaryotic organisms.

During the transcription, a mRNA chain is generated with one strand of the DNA double helix as a template, and the information in a section of DNA is transferred to a piece of mRNA, helped by the RNA polymerase and transcription factors. The RNA structure is very similar to the DNA structure but in RNA the nucleotide uracil (U) replaces the nucleotide T in DNA. The first product of transcription is different in prokaryotic and eukaryotic cells. Eukaryotic cells need post-transcriptional modification to produce the final mRNA. The primary transcript has further processing to finish translation. Splicing is a modification in which introns are removed and exons are joined, and alternative splicing is where proteins translated from alternatively spliced mRNA contain differences in their amino acid sequence and, often, in their biological functions, which contributes to the diversity of proteins produced by a single mRNA.

The synthesis of proteins from RNA is known as translation. Translation uses an mRNA sequence as a template to guide the synthesis of a chain of amino acids to form the protein. Translation has four phases: activation, initiation, elongation and termination. The mRNA goes to a ribosome, where it is translated. The mRNA is read by the ribosome as triplet codons (nucleotide triplets), usually beginning with an AUG, or the methionine codon to produce a specific amino acid chain, a polypeptide. Complexes of initiation and elongation factors bring aminoacylated transfer RNA (tRNA) into the ribosome mRNA complex, matching the codon in the mRNA to the anti-codon on the tRNA. As the amino acids are linked into the growing peptide chain, they begin folding, which requires other proteins called chaperone proteins. Translation ends with a UAA, UGA, or UAG stop codon. The polypeptide chain is then released from the ribosome as a mature protein.

Besides these processes, there are other, more specific, processes in living organisms, like reverse transcription and RNA replication. Reverse transcription is the transfer of information from RNA to new DNA, that was verified to contribute, for example, to the propagation of retroviruses, like the human immunodeficiency virus (HIV), to the genetic diversity in eukaryotes via retrotransposons, and to the replication of telomeres.

Many viruses replicate using RNA replication, which copies RNA to another RNA, using enzymes called RNA replicases.

All these processes are schematically represented in Figure 3.1. This figure represents the transfer of genetic information through the transcription, translation, reverse transcription and RNA replication processes. A more detailed description can be found in [36].

Living systems biological components have complex interactions and its study requires the integration of experimental and computational research. These dynamic systems exhibit properties, such as nonlinear dynamics and emergent behavior that are difficult to be inferred studying their components in an isolated form.

Systems biology can be defined as the understanding of structure and dynamics of the biological systems through the quantitative and qualitative analysis of their models. One example of these models are the networks of interactions between bio-entities, constructed with powerful prediction capabilities through the integration of experiments, with the use of mathematical methods and computation models and with the contribution of the knowledge obtained from several other scientific areas. Systems biology needs to integrate aspects like

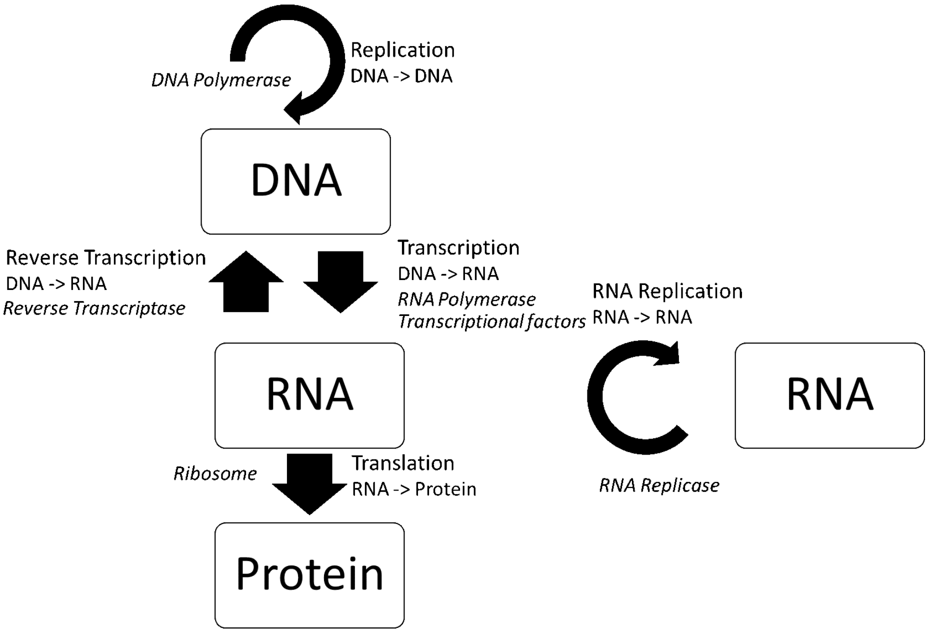


Figure 3.1: Transfer of genetic information in living organisms.

complexity, hierarchical structured levels of observation, geometrical relationships, non-linear dynamics, network modeling and the influence of biophysical constraints in order to find organizing principles that explain the evolution of systems in space and time [37, 38].

Some authors divide the systems biology in two approaches [37, 39]: pragmatic and theoretic. Pragmatic relies principally on high-throughput technologies and on massive data integration through mathematical modeling. Theoretical recognizes that complex physiological and adaptive phenomena take place at biological levels of organization higher than the sub-cellular ones.

Several aspects of the biological systems, such as robustness, the structure and the dynamics of their topological models, and how to apply this knowledge to drug discovery, have been investigated by several researchers [39, 40].

Biological complex systems may exhibit the property of self-organization, being the original system adaptable to the changes of the environment. They can be modeled through networks that allow the integration of information at different levels, to model the dynamics of their biological processes to be studied. Biological networks give a mathematical representation of the interactions of biological systems, which allows the use of computational methodologies and technologies.

3.2 Properties of Biological Networks

The interactions between the elements of a cell, in particular PPI are responsible for the biological functions of the living species. Proteomic studies are manually performed by researchers and require highly specialized knowledge, using several techniques, in different biological contexts, often with high error rates. Modeling the interactions between the bio-entities helps the study and understanding of the functional relationships existent between them.

The interactions between the biological systems can be represented by networks and topological properties can be used to characterize them. The comparison between networks can be done by using similarities of their topologies, allowing the generalization of their properties to real biological networks, or by their dissimilarities, to highlight topological differences that can help to identify relevant patterns that can be associated to biological processes of real networks.

Biological networks have distinguishing properties which are important to note and many share the following properties:

a) Biological networks are finite and sparse, $L \ll L_{max}$, and have more highly connected nodes than a random network. The degree of the biggest hub is proportional to N. New nodes prefer to link to highly connected nodes, while in random networks they link randomly [4, 16].

b) The degree distribution of many biological networks follows a power law, $p_k \sim k^{-\alpha}$ so they are scale-free. In random networks, the degree distribution is independent of the network size and is approximated by a Poisson distribution [2, 16, 29, 30].

c) Many biological networks are approximately scale-free networks with exponent $2 < \alpha < 3$ and the average path length $\langle d \rangle$ is proportional to $\log \quad \log(N)$ meaning that the network is an ultra-small world, where hubs reduce the path length linking to a large number of small-degree nodes. In random networks, $\langle d \rangle$ is proportional to $\log(N)$, meaning that they are small worlds [13, 41].

d) They have a much higher clustering coefficient than expected for a random network of similar size N and L. If $C(k)$ is measured by averaging the local clustering coefficient of all nodes with the same degree k, the clustering coefficient decreases with k. High degree nodes tend to have a smaller clustering coefficient than low degree nodes, so decreases with k and is largely independent of the network size N. They have some degree of hierarchy, where sparsely connected nodes belong to highly clustered areas. Communication between the different highly clustered neighborhoods is done through few hubs. The hierarchical modularity has a scaling parameter of the clustering coefficient, which follows a power-law $C(k) \sim k^{-1}$, while in random networks $C(k)$ is independent of k [2, 16, 29].

e) They are disassortative, meaning that hubs have the tendency to connect with small degree nodes, due to degree correlation [16].

Power-laws appear in a diverse range of natural and man-made phenomena and are not well characterized by their typical or average values. Most biological networks approximate a scale-free topology, but not all biological networks are scale-free, for example the transcription regulatory networks of S. cerevisiae and Escherichia coli is the combination of a scale-free network and an exponential network [16].

The two fundamental processes that are related with the development of real biological networks are the growth process, where, through time, new nodes join the system, and the preferential attachment, where nodes prefer to connect to nodes that already have many links. Duplicated genes produce identical proteins that interact with the same protein partners. Highly connected proteins have a higher probability of having a link to a duplicated protein, and therefore a higher probability of gaining new links [16, 42–45].

The Barabasi-Albert model describes scale-free degree distributions networks and it is used as a model to biological networks. The concept behind this model is to reveal information about the dynamics of the network, especially from an evolutionary perspective: growth and preferential attachment.

Power-law is rarely seen in its pure form in real systems because several processes affect the topology of real networks influencing the degree distribution, so there must be careful to false rejection of power-law hypothesis. They can have a low-degree saturation, k_{sat}, with a signature that is a flattened p_k for $k < k_{sat}$ .This indicates that there are fewer small degree nodes than expected for a pure power law. They can also have a high-degree cut-off, k_{cut}, that appears as a rapid drop in p_k for $k > k_{cut}$, indicating that we have fewer high-degree nodes than expected in a pure power law. This also limits the size of the largest hub, making it smaller than predicted. The presence of such cut-offs doesn't mean that the network is not scale-free, but could mean that additional phenomena take place in the system, that need to be understood.

The network diameter and the clustering coefficient play an important role in comparing a model with real systems. The distances in a scale-free model like Barabasi-Albert are smaller than the distances in a random graph of similar size and the clustering coefficient decays slower than expected for a random network, indicating that the obtained network is locally more clustered.

Biological networks are disassortative networks, similar to technological networks that are also disassortative, but opposed to social networks that are assortative [4, 16]. In a random network, the average degree of a node's neighbor k_{nn_random} is independent of node's degree k, and depends only on global properties $\left\langle k^2 \right\rangle$ and $\langle k \rangle$. In an assortative network, k_{nn} increases with k and, in a disassortative network decreases with k. So, an approximate degree correlation function, assuming that $\mathrm{K_{nm}}$ (k) follows a power law, is $\mathrm{K_{nm}}\,(\mathrm{k}) = \mathrm{ak}^{\mu}$, where the correlation exponent μ is negative, if the network is disassortative, it is positive if it is assortative and it is zero if it is a random network. Besides the correlation exponent μ , another measure can be used, the degree correlation coefficient r (Pearson correlation coefficient), which is also negative for disassortative networks. In the second measure it is assumed the linear dependence of r of $\mathrm{k_{nm}}\,(\mathrm{k}) \sim \mathrm{rk}$ with slope r. Examples of biological disassortative networks are the PPI and metabolic networks. Their disassortative characteristic is due to the fact that these networks are scale-free, obtaining similar values, when $\mathrm{k_{nm}}$ (k) is compared with the correspondent value in a degree preserving randomization network, without multi-links and self-links.

In a large scale-free network, another important characteristic is the fact that random removal of some nodes does not damage these networks, since hubs are much less than nodes with low degree. So, scale-free networks are robust against random failures, also because biological mechanisms are redundant, but are very fragile against attacks to their hubs.

Robustness is an intrinsic property of cellular networks that enables them to maintain their functions after the occurrence of random perturbations. Biological networks both in health and in disease are robust [46]. Drug action often fails due to the robustness of biological complex networks and drugs side-effects often indicate a point of fragility of the affected networks [47–49]. Robustness analysis was used with efficacy, to reveal main drug targets and describe drug actions [50] to combat genetic diseases, for example killing cancer cells [43, 44]. Considering the network dynamics, central nodes, such as hubs, or inter-modular overlaps and bridges were shown to act as efficient mediators of perturbations [51, 52].

3.3 Types of Biological Networks

Networks have been successfully used to model several components of biological processes. Bio-entities like genes, protein and metabolites are functionally linked and different experimental techniques are used to find them, like double mutant synthetic lethality to find genetic interactions or transcriptome expression profiling to find gene co-expression. According to the bio-entities they relate, and according to the interaction represented by them, biological networks can be divided in several categories: 1) PPI; 2) Gene co-expression networks; 3) Gene regulatory networks (GRN); 4) Signal transduction networks; 5) Metabolic and biochemical networks.

3.3.1 Protein-Protein Interaction Networks

Proteins interact with each other and the mapping of protein interactions contributes to understand the complex molecular relationships in living systems. These PPI depend on the type of the cell, its cycle phase and state, development stage, environmental conditions, protein modifications, presence of cofactors, and presence of other binding partners [53]. PPI can define most cellular processes and the characterization of PPI networks contributes for the understanding of the mechanisms of the biological processes in a cell [4, 13, 16, 53]. Close proteins homologs frequently interact in the same way, existing some conservation in the interaction patterns between similar proteins and domains [54].

The interactome is the complete PPI (physical) map of a living organism and there are efficient large-scale technologies to identify them. Functional and physical interactions between proteins are different concepts.

PPI can be modeled as complex networks, where proteins are represented by nodes, and interactions represented by edges. Biological networks are generally sparsely connected, which is considered an evolutionary advantage for preserving robustness to random failures, and tend to be heterogeneous, with few nodes highly connected (hubs) and many nodes with few connections [2, 16, 29, 30, 55]. The study of the topological properties of complex networks allows the understanding of their structures and the highlighting of some similarities, like small world properties [14], power-law degree distributions to distinguish from random networks, high average clustering coefficient showing modularity [15, 55–57], and clustering degree distribution to identify hierarchies in their organization [16]. Some researchers argue that not all PPI networks follow a power-law [29, 30, 57] and in most cases, only the tail of the degree distribution follows a power-law, existing a value $\mathrm{k_{min}}$ for which the power-law is observed.

There are different experimental techniques to detect protein interactions, individually or screening interactions on a genomic scale. Some techniques enable screening of a large number of proteins in a cell and others monitor and characterize specific biochemical and physic–chemical properties of a protein complex [4, 54].

Some well-known large-scale and high-throughput techniques that can detect proteins interactions in living systems are the pull down assays, tandem affinity purification (TAP), mass spectrometry (MS), phage display, yeast two-hybrid (Y2Y), micro-arrays and, more recently, next generation interaction sequencing (NGIS) - Y2H. They originated the construction of large-scale maps of protein interaction networks and various datasets for different organisms [4, 54, 58, 59].

Experimental approaches have limitations, like the low interaction coverage and experimental biases to certain protein types and cellular localizations. These limitations can be surpassed by the use of computational methods for predicting protein interactions [13, 54, 60]. Computational methods can be applied to various types of prediction problems and can be useful to choose potential targets for experimental screening or for validating experimental data. Several methods use both, experimental and computational methods.

There are computation methods to identify interactions that can be based on the genomic context and can be explored through three inference methods: Domain fusion, conserved neighborhood, and phylogenetic profiles. The domain fusion method, or Rosetta Stone method, infers PPI from protein sequences in different genomes. This method is based on the observation that some interacting proteins have homologs in other genomes, which are fused into one protein chain, called Rosetta Stone protein. Domain fusion method has the least coverage compared to other genomic context methods. The conserved neighborhood method is based on the hypothesis that, if genes that encode two proteins are neighbors on the chromosome in several genomes, the corresponding proteins are likely to be functionally linked. This method has two requirements, one is to identify orthologous in another genome and the other is to find those orthologous that are adjacent on the chromosome. It has a better coverage than the previous one and it focus on operons, where genes are transcribed with a common orientation (co-directionally). The phylogenetic profiles method identifies functional linkages between proteins based on the hypothesis that if they have similar phylogenetic profiles they tend to be functionally linked. Genes with similar phylogenetic profiles essentially produce similar phenotypes [16, 23].

Some of the computational methods are based on co-evolution to predict interactions. Co-evolution can be defined as the joint evolution of ecologically interacting species, suggesting the existence of mutual selective pressure on two or more species. Interacting proteins very often co-evolve and the changes in one protein, leading to the loss of function or interaction, should be compensated by the correlated changes in another protein. The orthologous of co-evolving proteins also tend to interact, being possible to infer unknown interactions in other genomes [60, 61]. Correlated mutations in multiple sequence alignments can be used to identify functional interactions between proteins. An interaction index can be calculated based on the correlation values, to detect the presence of a distinctive number of compensatory mutations in corresponding proteins of different species that will indicate the co-adaptation of interacting proteins.

Homology-based inference of PPI works better within species than across species, being accurate only if there is a high level of sequence identity [62]. Other group of computation methods is based on co-expression of genes [13, 60]. Other methods, used to identify PPI, mine data from the information of experimental protein associations. This information can be obtained from literature, through classification-based approach [63], or through the identification of abstracts about PPI from literature [64].

Several classification methods have been used to predict PPI. Those methods use several data sources to train a classifier to differentiate between positive examples of truly interacting protein pairs from the negative examples of non-interacting pairs. Each protein or protein pair can be encoded as a feature vector, where features may represent a particular information source of protein interactions, or evidence coming from various experimental methods. Random forest decision (RFD) and support vector machines (SVM) rank as top classifiers [60].

Each network gives a static view of the PPI, but PPI in a living system are complex dynamic interactions, so it should be considered the dynamics and strength of the interactions to make more accurate predictions [13].

Several validation methods of PPI data have been proposed [57], like the expression profile reliability (EPR) method based on the fact that interacting proteins are co-expressed; the paralogous verification method (PVM) based on the fact that if two proteins interact, their paralogs probably interact; and protein localization method (PLM) that assumes that true positives are interacting proteins localized in the same cellular part having a common cellular function.

The knowledge of molecular and function properties of individual proteins, obtained by researchers in various areas of biology, is stored in protein databases like UniProt[1] [65]. These databases are manually curated by researchers. The molecular function of proteins is not yet fully determined and predicting protein function is still a research area in computational biology. Experimental and computational techniques have been developed to infer interactions and protein functions from PPI networks [4] that hold information on how different proteins operate together to enable the biological processes within the cell. The degree distribution of several PPI networks is approximately scale-free and there are always proteins with a high degree of connectivity that appear to be of high biological significance and being the very important for the survival of the cell. It has been shown that these networks are highly dynamic [3, 43].

There are several biological databases containing PPI data, which can be used to construct biological network. In Table 3-1 is presented a list of the main databases, some more general and others specific of an organism. A comparison of these PPI databases and repositories can be found in [53] and [13].

One of the databases is the search tool for the retrieval of interacting genes/proteins (STRING) [66]. Here, PPI can be resultant from high-throughput experimental data, from the mining of databases and literature, from the analysis of co-expressed genes and from computational predictions, including those based on genomic context analysis. STRING covers around one thousand organisms, from bacteria and archaea to humans. Interactions are benchmarked independently and then a combined score is calculated, where a higher confidence means that more than one type of information supports a given interaction [13, 63].

To improve data quality of molecular interactions and curated molecular interactions, the International Molecular Exchange (IMEx) Consortium of molecular interaction database providers [67], founded by DIP, IntAct, and MINT, together with the HUPO proteomics standards initiative (HUPO-PSI) [2], defined the minimal information about a molecular interaction (MIMIx) standard.

[1] http://www.uniprot.org

[2] http://psidev.info/

3.3.2 Gene Co-expression Networks

Gene co-expression networks, also named transcript-transcript association networks, are an example of correlation networks. They can be used to analyze data of biological systems obtained from DNA micro-array or RNA sequencing (RNA-Seq) technologies. A gene co-expression network captures information on the correlation of gene expression in different biological conditions and is a weighted undirected network, where the nodes are genes, the edges are pairs of genes that have significantly similar expression patterns and the edge weights represents the strength of correlation of a pair of genes.

Table 3.1: List of PPI data repositories and databases by alphabetic order.

Acronym	Name	Link (accessed at 03/10/18)
BioGRID	Biological General Repository for Interaction Datasets	http://thebiogrid.org/
DIP	Database of Interacting Proteins	http://dip.doe-mbi.ucla.edu/dip/
DroID	Drosophila Interaction Database	http://www.droidb.org/
HPID	Human Protein Interaction Database	http://wilab.inha.ac.kr/hpid/
HPRD	Human Protein Reference Database	http://www.hprd.org
IntAct	Molecular Interaction Database	http://www.ebi.ac.uk/intact/
MINT	Molecular Interaction database	http://mint.bio.uniroma2.it/
MIPS	Mammalian Protein-Protein Interaction Database	http://mips.helmholtz-muenchen.de/proj/ppi/
STITCH	Search Tool for Interacting Chemicals	http://stitch.embl.de/
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	http://string-db.org/

The Pearson correlation coefficient is frequently used as a co-expression measure and this coefficient, after a threshold applied, can be used to construct a gene co-expression network, also designated as relevance network. The creation of a correlation matrix requires demanding computation resources and so the network analysis is often restricted to a subset of genes. The adjacency matrix represents the connection strength between each pair of nodes [44]. Two of the databases, where gene expression data can be retrieved are the ArrayExpress [68] and

Expression Atlas [69]. More repositories and databases can be seen in Table 3.2.

Table 3.2: List of gene expression data repositories or databases by alphabetic order.

Acronym	Name	Link (accessed at 03/10/18)
ArrayExpress	ArrayExpress Archive of Functional Genomics Data	https://www.ebi.ac.uk/arrayexpress/
COXPRESdb	coexpression database	http://coxpresdb.jp/
CSB.DB	A Comprehensive Systems-Biology Database	http://www.csbdb.de/index.html
Expression Atlas	Expression Atlas	https://www.ebi.ac.uk/gxa/home
Genevestigator	Genevestigator	https://genevestigator.com
GEO	Gene Expression Omnibus	https://www.ncbi.nlm.nih.gov/geo

Gene expression profiles across samples can be highly correlated and may correspond to protein complexes or pathways [6]. Gene co-expression networks defined as weighted correlation networks preserve the continuous nature of the co-expression information [70]. The analysis of weighted gene co-expression networks [44] have been used to identify co-expressed modules that may correspond to pathways and intra-modular hub genes representative of respective modules [17, 70].

Techniques such as micro-array experiments, like DNA micro-arrays or next generation sequencing (NGS), like RNA-Seq evaluate a large number of genomic sequences (genes), under multiple conditions (samples) [71, 72]. Gene expression usually refers to the amount of messenger RNA that corresponds to a gene.

A gene expression micro-array can measure the expression level (mRNA abundance) of thousands of genes under multiple conditions. Gene co-expression networks constructed from gene expression micro-arrays data capture the relationships between transcripts, using correlation analysis to build the correlation matrix, which is converted to an adjacency matrix representing the co-expression network. Each gene corresponds to a node and two genes are connected by an edge if their expression values are highly correlated [73, 77]. The normalized micro-array expression data can be represented by a $\mathrm{m} \times \mathrm{n}$ dimensional matrix whose i^{th} column x_{i} is a numeric vector (representing for example gene expression levels) corresponding to the i^{th} gene (or probe), with m components corresponding to m sample measurements.

The NGS technique determines the DNA or cDNA sequence and it does not require a prior knowledge of the genome or normalization methods. When applied to cDNA is called RNA-Seq and can be applied to gene expression profiling between samples.

The adjacency matrix of a correlation network is constructed based on pairwise correlations between numeric vectors. These numeric vectors represent observed quantity measurements of variables. The relationship between a pair of numeric vectors can be measured by a correlation coefficient and besides the Pearson correlation coefficient, there are other correlation coefficients that can be used, like the Spearman correlation or the bi-weight mid correlation coefficient.

The Pearson correlation is defined as

$$\mathrm{cor}\,(\mathrm{x},\mathrm{y}) = \frac{\mathrm{cov}(\mathrm{x},\mathrm{y})}{\sqrt{\mathrm{var}\,(\mathrm{x})\,\mathrm{var}(\mathrm{y})}} \tag{3.1}$$

where the covariance between vectors $\mathrm{x} = (\mathrm{x_u})$ and $\mathrm{y} = (\mathrm{y_u})$, for $\mathrm{u} = 1, \ldots, \mathrm{m}$ observations, is defined as follows

$$\mathrm{cov}\,(\mathrm{x},\mathrm{y}) = \frac{\sum_{\mathrm{u}} (\mathrm{x_u} - \mathrm{mean}\,(\mathrm{x})\,)(\mathrm{y_u} - \mathrm{mean}\,(\mathrm{y})\,)}{\mathrm{m} - 1} \tag{3.2}$$

The variance of x is $\mathrm{var}\,(\mathrm{x}) = \mathrm{cov}(\mathrm{x},\mathrm{x})$.

Pearson correlation detects only linear correlations. If x and y are two vectors and if $\mathrm{y} = \mathrm{ax} + \mathrm{b}$, the correlation $\mathrm{cor}\,(\mathrm{x},\mathrm{y})$, between vectors x and y, has values in the interval $-1 \leq \mathrm{cor}(\mathrm{x},\mathrm{y}) \leq 1$ and has the sign of a. $\mathrm{cor}\,(\mathrm{x},\mathrm{y}) = 1$ in the case of a perfect direct correlation, $\mathrm{cor}\,(\mathrm{x},\mathrm{y}) = -1$ in the case of perfect decreasing correlation and $\mathrm{cor}\,(\mathrm{x},\mathrm{y}) = 0$, if they are independent.

The Pearson correlation is sensitive to outlying observations, but the Spearman correlation is more robust to outliers. The Spearman correlation does not require a linear relationship and its value is the Pearson correlation of the ranks of the two quantitative vectors x and y

$$\mathrm{SpearmanCorr}\,(\mathrm{x},\mathrm{y}) = \mathrm{cor}\,(\mathrm{rank}\,(\mathrm{x})\,, \mathrm{rank}\,(\mathrm{y})) \tag{3.3}$$

The Spearman correlation detects monotonic relationships, linear or not linear.

Another alternative is the bi-weight mid-correlation, which is based on the median. Bi-weight mid-correlation has the relative high power of the Pearson correlation and the relative robustness of the Spearman correlation to outliers [73, 74].

3.3.3 Gene Regulatory Networks

GRN contain information concerning the control of gene expressions in cells, through many variables. Gene expression includes the transcription of the gene to mRNA and the translation to protein and eventual post-translational modification. In these networks nodes are genes and edges are their regulators. These networks are bipartite (two types of nodes, genes and regulators) and direct (regulators control genes). One example of regulators are the transcription factors (TF) and other are the microRNAs. TF are proteins they can work as activators or inhibitors when controlling the transcription process and are also encoded by genes and regulated.

The expression of one gene can be controlled by the gene product, protein, or by another gene. These networks have specific motifs and patterns in their topology. GRNs are usually sparsely connected and they are very sensitive and flexible to evolution. The number of regulators $\mathrm{N_{reg}}$ grows faster than the number of genes $\mathrm{N_{tot}}$ they regulate and it has been shown that $\frac{\mathrm{N_{reg}}}{\mathrm{N_{tot}}}$ is proportional to N for prokaryotes and is proportional to $\mathrm{N}^{0.3}$ for eukaryotes, where N is the network size. They have a modular topology and follow power-law distributions [4].

GRN data have been also collected in several databases, such as JASPAR [80], and TRANSFAC [81] (see Table 3.3, for a more extensive list).

3.3.4 Signal Transduction Networks

Signal transduction networks can be modeled by multi-edged directed graphs to represent a series of interactions between different bio-entities such as proteins, chemicals or macromolecules, to analyze signal transmission from the outside to the inside of the cell, or within the cell. Through these networks, one wants to understand how a cell senses its environment and reacts to it. Several proteins interact via activation and inhibition to convert an external signal into a physiological response

Table 3.3: List of GRN repositories or databases by type of interactions and by alphabetic order.

Acronym	Name	Link (accessed at 03/10/18)
protein-DNA		
BCI	B-Cell Interactome	https://systemsbiology.columbia.edu/b-cell-interactome
JASPAR	JASPAR	http://jaspar.genereg.net/
TRANSFAC	TRANSFAC	http://www.gene-regulation.com/pub/databases.html
post-translational modification		
KinomeXplorer	KinomeXplorer	http://kinomexplorer.info/
PHOSIDA	PHOsphorylation SIte DAtabase	http://141.61.102.18/phosida/index.aspx
Phospho-ELM	Phospho-ELM	http://phospho.elm.eu.org/about.html
PSP	PhosphoSitePlus	https://www.phosphosite.org/homeAction.action

Similarly to GRN, these networks also exhibit common topological patterns and motifs, important for biological functionality. The nodes with the highest centralities in such networks correspond to domains involved in signal transduction and cell-cell contacts. Signal transduction networks are sparse and they follow scale-free distributions [4, 75].

A database that stores information about signal transduction pathways is TRANSPATH [76], a database of mammalian signal transduction and metabolic pathways (see Table 3.4).

Table 3.4: Signal transduction database.

Acronym	Name	Link (accessed at 03/10/18)
TRANSPATH	TRANSPATH	http://genexplain.com/transpath/

3.3.5 Metabolic and Biochemical Networks

Metabolic and biochemical networks allow studying and modeling the metabolic pathways existent in the cells of the various organisms, containing the information about biochemical events, like the chemical reactions of the metabolism and their respective regulatory interactions and how they are correlated. Metabolic pathways are a series of chemical reactions occurring within a cell at different time points and catalyzed by enzymes that modify the metabolites. For adequate functioning, enzymes can be dependent on other cofactors such as vitamins.

Metabolic networks are scale-free, with a small world structure when considering the topology based on its metabolites. The probability that a substrate participates as input in k metabolic reactions follows the power-law distribution p(k) $=\mathrm{k}^{-\alpha_{\mathrm{in}}}$, $\alpha_{\mathrm{in}}\approx 2.2$, whereas the probability of a substrate to be produced by k metabolic reactions equals similarly to p(k) $=\mathrm{k}^{-\alpha_{\mathrm{out}}}$, $\alpha_{\mathrm{out}}\approx 2.2$. Metabolic networks are extremely heterogeneous, robust and vary between organisms. These networks apparently preserve the network diameter even in distantly related organisms and can form hierarchical structures where specific patterns and motifs are over-represented [4].

There are also several databases holding information about metabolic and biochemical networks (see Table 3.5), where the Kyoto encyclopaedia of genes and genomes (KEGG) [77, 78] is one of the most well-known.

Table 3.5: List of metabolic and biochemical repositories or databases by alphabetic order.

Acronym	Name	Link (accessed at 03/10/18)
BIGG	Biochemical Genetic and Genomic	http://bigg.ucsd.edu/
BioCyc	BioCyc Database Collection	http://biocyc.org/
EcoCyc	EcoCyc E. Coli Database	http://ecocyc.org/
KEGG	Kyoto Encyclopedia of Genes and Genomes	https://www.genome.jp/kegg/pathway.html
metaTIGER	Metabolic Gene Evolution Resource	http://www.bioinformatics.leeds.ac.uk/metatiger/

3.4 Summary

This chapter began with an introduction to the principles of molecular biology, with an emphasis on biological networks and their properties. The last sections described several types of biological networks as well the main data repositories with a greater focus on the PPI networks and co-expression used in this research.

Chapter 4

Networks Topology-based Methods Applied to Biological Data

With the evolution of computational methods and tools, biological data has increased enormously and is at disposal in many databases, being necessary to extract meaningful biological knowledge from the information contained in those data.

The inference of biological networks from heterogeneous biological data from different datasets and using the theory of graphs and statistics, allow to study to what extent the structure of these networks is related to biological and pathological processes and their dynamics. Unsupervised methods, using clustering techniques, and supervised methods, are used to study biological networks either globally using global topological properties or locally using graphlets, motifs and modules.

The analysis of biological networks using a system approach, allows to identify patterns that may be associated with biological processes that can be used in the diagnosis and prediction of diseases, including cancer and in the establishment of relationships between genotypes and phenotypes. The contributions of systems biology are addressed in the review [79].

Biological networks allow the representation of different levels of abstraction, depending on the represented interactions. Their study and integration allow a better understanding of the system structure and its mechanisms, like those related to robustness, to the management of functions faults tolerance, caused by changes or loss of functions derived by diseases or the environment changes. This knowledge is very important for the development of resistance therapeutic agents [80], since about 73% of the genes are considered not essential and disease genes tend to correspond to PPI non-hubs [81, 82].

Physiological and pathological processes have the participation of functional modules of proteins that interact in a stable or transient way to perform structural or functional actions associated to these processes. Highly co-expressed genes are more likely to be co-regulated, playing biologically important and regulatory roles in processes related to diseases. Interacting proteins are more likely to be encoded by genes with similar expression profiles, so genes expression can be used to understand the organization and dynamics of PPI networks, relying on the hypothesis that genes with shared interactions in the networks tend to share common functions.

This chapter describes two studies that use the network topology to characterize biological networks. One of the studies is the characterization of the human oral proteome network, which had not yet been performed, through its topological properties, and the other study consists of the use of co-expression networks as models to be able to associate patient risk factors with modules of genes from these networks for the head and neck squamous cell carcinoma (HNSCC) disease.

4.1 Introduction

Several researchers evaluated the topological properties of different kinds of networks: Newman [2, 29] evaluated the topological properties of twenty-seven datasets from different areas, like social, biological and technological; Colliza et al. [15] evaluated the topological properties of three distinct PPI networks of S. cerevisiae; Liu et al. [83] evaluated the topological properties of classical music from Bach, Mozart, Chopin, and Chinese pop music; Clauset et al. [30] evaluated twenty-four datasets from different areas, like physics, earth sciences, computer and information sciences, two of them being the PPI network of *S. cerevisiae* and the metabolic network of the *bacterium E. coli.*

Biological networks are generally sparsely connected, which is considered an evolutionary advantage for preserving robustness to random failures, and tend to be heterogeneous, with few nodes highly connected (hubs) and many nodes with few connections [2, 4, 16, 29, 30]. These networks tend to be robust against random perturbations, but the removal of hubs often leads to system failure [3].

Studying the topology of biological networks using topological properties, which are quantitatively measurable, allows to detect and compare biological processes, including those contained in diseases, to try find associated topological signatures.

Network measures unsupervised methods, using clustering, and supervised methods, using techniques of machine learning when there are data whose class is known, or networks statistics can be used to describe the topological properties of a single network, allowing their characterization and allowing to compare different networks and identify interrelated bio-entities with biological significance. Networks concepts can be divided in three categories, general, intra-modular and inter-modular, all according to their application, either to the whole network without reference to modules, to a module, or to describe relationships between modules, respectively. Modular network concepts can be used to define meta-networks, which are networks whose nodes are modules [44, 84, 85]. Meta-networks can contribute to complexity reduction in terms of data, which allows better manageability and easier interpretation of the extracted knowledge.

So, the characterization of biological networks using topological properties allows understanding their structures and highlighting some similarities. Some topological properties generally used are (see Section 2.2.1), the network diameter, d , and the shortest path length that may indicate small-world properties of the analyzed network [14], the power-law degree distribution existent in non-random networks, differently from random networks [16], the high average clustering coefficient, C_{avg}, since biological networks have a significantly higher average clustering coefficient compared to random networks, that indicates modularity [4, 15, 56, 86], and the clustering degree distribution that identifies hierarchies in networks organization [16].

The degree distribution of most biological networks approximates a power-law, so they are named scale-free [2, 16, 29, 30] and in most cases only the tail of the distribution follows a power-law, existing a value x_{min} for which the power-law is observed [29, 30]. However, this is still a controversial subject, as some researchers defend that some PPI networks do not follow a power-law [57].

Several methods can be applied to detect and characterize power-law distributions. One of them is the least-squares fitting of the distribution by a straight line in a log-log plot, which is not considered very accurate. More accurate methods are the maximum-likelihood fitting methods with goodness-of-fit tests, based on the KS statistic to obtain the slope of the fitted line, the value of x_{min} and a p-value. Usually, a value of $p \leq 0.1$ is considered for ruling out the power-law hypothesis [29, 30].

Biological data has a high degree of complexity and the use of computational methods is necessary. Traditional computational approaches to extract evidences from data took a lot of time and most of the times led to inconclusive results. Using networks to represent the interactions of bio-entities allowed the use of graph theory to extract all the knowledge to represented inherent biological processes.

PPI networks can be constructed using datasets generated by several experimental and computation approaches, and co-expression networks datasets can be generated by high-throughput gene expression profiling technologies like micro-arrays [87] or RNA-Seq [72], techniques that evaluate a large number of genes under multiple conditions (samples), like multiple disease states, through time or for different individuals with the same condition.

PPI can be modeled as complex networks, where proteins are represented by nodes, and interactions represented by edges or links. The human interactome is the network formed by all human protein-protein interactions and still is a complex and not yet completely known system. Networks topological properties quantity measures have been recognized by their contribution to describe and understand their structures, functional relationships and evolution. Their analysis are also known to help in having a better comprehension of the diseases mechanisms and in the identification of drug targets.

Proteins with high betweenness centralities in PPI networks are called “bottlenecks” and usually have essential functional and dynamic properties [4, 18]. For example, the eigenvector centrality measurement has been used for efficient page ranking on the web and in biology has been used to identify synthetic genetic interactions, gene-disease associations or network hubs. Also, in biological networks, proteins or other bio-entities with low eccentricities usually have a marginal functional role in the system.

The study of gene co-expression networks from mRNA gene expression data, helps to extract structural and functional features that can be used to better understand the data. Proteins do not work isolated and are not always active and protein interactions are often encoded by co-expressed genes. Also shared genes/proteins interactions tend to be functionally related. Genes that are co-expressed are more likely to encode interacting proteins and studying co-expressed patterns in co-expressed networks can contribute to the understanding of cellular processes behind those biological systems.

In this chapter, two case studies are presented. The first one is the topological analysis of the human oral proteome network, where proteins were obtained from proteomic studies done by researchers. This dataset was studied for the first time under a systemic view using networks to model their interactions in [88]. The second one is the study of the HNSCC, where the weighted gene co-expression networks analysis weighted gene co-expression network analysis (WGCNA) methodology was used to identify molecular mechanisms associated with HNSCC and to find the contribution of several risk factors, like alcohol use and age in this type of carcinoma [70].

In the first study it can be seen that the human oral proteome network is a subsystem of the whole human proteome with very defined topological characteristics and that its study contributes to a better understanding of this subsystem that was still quite unknown from the point of view of its topological behavior. In the second study, computational techniques were used to associate, modules of co-expressed genes with risk factors associated with several patients, using the meta-information relationship of these modules, for a particular disease, the HNSCC.

Systemic study of biological networks, building different network models, contributes to the understanding, through differentiated computational techniques, for the understanding of real systems under study.

4.2 Quantitative Characterization of Protein Networks of the Oral Cavity

The human oral proteome is a subsystem of the human proteome complex system and their proteins were obtained from proteomic studies done by researchers. Here is described the exploratory study of human oral PPI networks done, using different confidence scores and obtained from different prediction methods. This study includes the analysis of relevant topological properties of the human oral PPI network dataset, the comparison to respective random networks, the analysis of their degree distribution, that is supposed to follow a power-law distribution, and the evaluation of that assumption.

The results of this study, about the characterization of the oral protein network, is presented in the paper [88] and this study aimed to better understand the organization and network topology of the human oral proteome. Several PPI networks were built with several confidence scores and different prediction methods and various network topological measurements were used. They were also compared with random networks of the same size.

4.2.1 Data Sets

The human oral proteome dataset was obtained from proteomic studies done by researchers [89]. To obtain PPI networks induced by the oral proteins it was used the STRING database [90] that, given several distinct types and sources of PPI information, provides an integration and evaluation service. Interactions in STRING are provided with a confidence score and are obtained from different prediction methods like Experiments, Co-occurrence, Co-expression, Databases, Neighborhood, Gene Fusion and Text Mining [63].

Using the human oral proteome dataset, several networks were constructed representing the entire set of PPI for different confidence scores ($\geq$ 100, $\geq$ 200, $\cdots$ $\geq$ 900) and for different prediction methods (Experiments, Co-occurrence, Co-expression, Databases and Neighborhood). These networks were constructed as undirected, unweighted and with no self-edges.

4.2.2 Topological Properties

An undirected graph G can be defined as a pair $\mathrm{G = (V, E)}$, where V is a set of vertices representing the nodes and E is a set of edges representing the connections between the nodes i and j. The number of nodes of a graph G is denoted by N and the number of edges of a graph is denoted by L. Given a graph $\mathrm{G = (V, E)}$ the adjacency matrix representation consists of a $\mathrm{N \times N}$ matrix, $\mathrm{A = [a_{ij}]}$, such that $\mathrm{a_{ij}}$= 1 if $\mathrm{(i, j) \in E}$ or $\mathrm{a_{ij}}$= 0 otherwise or for an weighted network $\mathrm{a_{ij}{=}w_{ij}}$ if $\mathrm{(i, j) \in E}$ or $\mathrm{a_{ij}}$= 0 otherwise. For undirected graphs the matrix is symmetric [2, 4].

The following topological properties (see Section 2.2.1) were used to study the human oral proteome networks.

The average number of neighbors, denoted by $\langle \mathrm{k} \rangle$, which indicates the average connectivity of a node in the network, where the degree $\mathrm{k_i}$ of a node i in an undirected graph represents the number of neighbors of the node i [2, 4, 17, 91].

The average clustering coefficient, $\mathrm{C_{avg}}$, of the whole network, being the average of the local clustering coefficient of the nodes i in G with $0 \leq \mathrm{C_i} \leq 1$ [2]. For random networks with the same properties of the considered datasets, it was considered C_{rand} [15, 56, 86].

The network diameter, d , which is the largest distance between two nodes. The average shortest path length or characteristic path length,<d>, that gives the expected distance between two connected nodes. The eccentricity is the maximum non-infinite length of a shortest path between i and another node in the network. The maximum node eccentricity is the diameter. The network radius, r, is the minimum among the non-zero eccentricities of the nodes in the network. A normalized version of the average number of neighbors $\langle k \rangle$, is the density, dens, of a network, which varies between 0 and 1 . Networks with a star-like topology have centralization, cent, close to 1, whereas decentralized networks are characterized by having centralization close to 0. The network heterogeneity, h, reflects the tendency of a network to contain hub nodes [2].

In scale-free networks the degree distribution is a power-law distribution, being inversely proportional to a degree exponent α. The value of α determines many properties of the system. For smaller values of α, the role of the hubs in the network becomes more important. For $\alpha > 3$, hubs are not relevant, while for $2 < \alpha < 3$, there is a hierarchy of hubs, with the most connected hubs in contact with a small fraction of all nodes [13, 41]. The small scaling parameter typically lies in the range $2 < \alpha < 3$, although there are exceptions.

A network that presents a power-law distribution is also an evidence of being a small-world network having a characteristic path length similar to that of random networks, but have a much higher clustering coefficient than that of random networks. Scale-free networks are highly robust against random node failures, but are sensitive to the failure of hubs [2, 43].

In practice power-law applies only for values greater than some minimum x_{min} and only the tail of the distribution follows a power-law. If x represents the quantity whose distribution we are interested in, a power-law distribution is described by a probability density $p(x)$ in continuous and discrete case (see Table 2-6) [29, 30]. The fitting of power law forms to empirical distributions give some estimate of the slope α and the lower-bound x_{min}.

Another method of plotting the data is to calculate a CDF, $P(x)$, which is defined for the continuous and discrete cases by $P(x)$ that also follows a power law, but with a different exponent $\alpha - 1$ [29].

Using the least-squares linear regression on the logarithm of the histogram to extract the slope α generates systematic errors [30]. The method of maximum likelihood for fitting power-law distributions to observed data gives accurate parameter estimates in the limit of large sample size, $\widehat{\alpha}$, which is given for the continuous and discrete cases.

The estimate $\widehat{x}_{min}$ is the value of x_{min} that minimizes the maximum distance between the CDFs of the data and fitted model, using the KS statistic. To quantify the uncertainty in estimation of x_{min} it was used the bootstrap method (see Section 2.4) [30].

Since being roughly straight on a log-log plot is a necessary, but not sufficient condition for power-law behavior, it was used a goodness-of-fit test, which generates a p-value that quantifies the plausibility of the power-law behavior and it is considered that the power law can be ruled out if $p \leq 0.1$. To generate the synthetic data it was used the semi-parametric approach [30].

4.2.3 Results and Conclusion

The topological properties of the networks computed included, the number of nodes and edges, the connected components, the network diameter, radius, density, centralization and heterogeneity, the C_{avg} and the corresponding average clustering coefficient for a random network (C_{rand}) [15, 56, 86], the characteristic path length, and the distributions of node degrees and clustering degrees [91, 92]. Studied networks degree distributions were fitted to the power-law model and the corresponding p-value was measured using maximum-likelihood fitting methods with goodness-of-fit tests based on the KS statistic [30].

The topological properties measured for each studied PPI networks, with different confidence scores and different prediction methods, are listed in Table 4.1 and are shown in Table 4.3 and in Table 4.4.

The largest component represents almost the whole network in all networks except for the Co-occurrence network. Comparing C_{avg} with C_{rand}, between the studied networks and correspondent random networks, indicates modularity (see figure 4.1). It was observed (see Table 4.3 and Table 4.4) that with the increase of the confidence score, the networks size, average degree, density and centralization decreases, but the diameter, radius, characteristic path length and heterogeneity increases. Considering different prediction methods to obtain the PPI networks, the diameter varies from 7 (Neighborhood) to 21 (Co-occurrence) and the characteristic path length varies from 2.80 (Neighborhood) to 7.31 (Co-occurrence).

Figure 4.2 shows the cumulative degree distribution of the studied networks. The basic parameters calculated for the analysis of the degree distributions of the studied networks are described in Table 4.2.

For the PPI network of the *yeast Saccharomyces cerevisiae*, Clauset [30] obtained the value of $\alpha = 3.1 \pm 0.3$ and $p = 0.31$.

The studied datasets had values (see Table 4.5) from $\alpha = 1.53$ for confidence score ≥ 900 to $\alpha = 3.5$ for confidence score ≥ 100 and ≥ 200 and from $\alpha = 1.77$ for the Co-occurrence dataset to $\alpha = 2.57$ for the Neighborhood dataset. The p-values show that power-law distribution model is consistent for the networks with confidence score ≥ 100, ≥ 700 and ≥ 800 and for all the networks of the prediction methods except the Co-occurrence network.

So, in this study, the main topological properties of human oral PPI networks using different confidence scores and different prediction methods were evaluated [88]. The node degree distributions were fitted to the power-law model and the corresponding p-values were calculated, using maximum-likelihood fitting methods and goodness-of-fit tests based on the KS statistic.

Table 4.1: Topological properties measured.

Name	Description
N	The number of proteins of the largest component
%N (LC)	The % of proteins of the largest component regarding the total number of proteins for each confidence score or prediction method
L	The number of interactions of the largest component
%L (LC)	The % of interactions of the largest component regarding the total number of interactions for each confidence score or prediction method
C_{avg}	C_{avg}
C_{rand}	C_{rand}
d	The diameter of the network
r	The radius of the network
<d>	The characteristic path length
<k>	The average degree of the network
dens	The network density
cent	The network centralization
h	The network heterogeneity

Table 4.2: Parameters for the analysis of degree distributions.

Name	Description
n	The size of the dataset
<x>	The average of the observed values
x_{max}	The maximum of the observed values
x_{min}	The minimum of the observed values where the distribution follows a power-law
$x_{min\ err}$	The error of x_{min}
α	The slope of the fitted power-law
α_{err}	The slope error
p	The p-value

Table 4.3: Topological properties of the studied PPI networks with different confidence scores and prediction methods (part 1).

Description	N	%N (LC)	L	%L (LC)	C_{avg}	C_{rand}
CS≥100	3052	100.00	200841	100.00	0.302	0.001
CS≥200	3031	100.00	152817	100.00	0.275	0.001
CS≥300	2993	100.00	90092	100.00	0.237	0.003
CS≥400	2962	100.00	61944	100.00	0.237	0.013
CS≥500	2916	99.93	47345	100.00	0.257	0.024
CS≥600	2836	99.75	35846	99.99	0.282	0.042
CS≥700	2577	99.12	26965	99.94	0.329	0.062
CS≥800	2359	99.74	21202	99.91	0.348	0.085
CS≥900	1840	93.59	10944	98.84	0.342	0.251
Experiments(LC)	2271	98.65	17159	99.87	0.183	0.136
Co-expression(LC)	1860	96.77	51007	99.81	0.467	0.002
Co-occurrence(LC)	393	53.47	1668	70.23	0.377	0.244
Databases(LC)	1341	91.22	12573	98.26	0.505	0.073
Neighborhood(LC)	466	99.57	5030	99.98	0.340	0.013

Table 4.4: Topological properties of the studied PPI networks with different confidence scores and prediction methods (part 2).

Description	d	r	<d>	<k>	dens	cent	h
CS≥100	6	3	2.23	131.61	0.04	0.27	0.91
CS≥200	6	3	2.23	100.18	0.03	0.23	0.94
CS≥300	6	4	2.59	60.20	0.02	0.13	0.90
CS≥400	7	4	2.84	41.83	0.01	0.01	0.96
CS≥500	7	4	3.07	31.47	0.01	0.01	1.04
CS≥600	9	5	3.35	25.28	0.01	0.09	1.11
CS≥700	9	5	3.81	20.93	0.01	0.08	1.16
CS≥800	10	5	3.91	17.98	0.01	0.07	1.19
CS≥900	14	8	5.05	11.90	0.01	0.06	1.30
Experiments(LC)	9	5	3.61	15.11	0.01	0.15	1.49
Co-expression(LC)	13	7	3.49	54.85	0.03	0.18	1.28
Co-occurrence(LC)	21	11	7.31	8.49	0.02	0.08	0.86
Databases(LC)	16	8	4.84	18.75	0.01	0.07	1.05
Neighborhood(LC)	7	4	2.80	21.59	0.05	0.17	1.05

Many real networks have been found to have approximately scale-free topologies with the associated topological properties presented and this study, made with protein-protein interaction networks obtained from the human oral proteome, showed it [88].

From Figure 4.3 it can be observed that clustering degree distributions are not independent of the degree, decreasing with it. This gives evidence of some hierarchical modularity [16].

We can conclude that most of the studied networks generate scale-free networks with

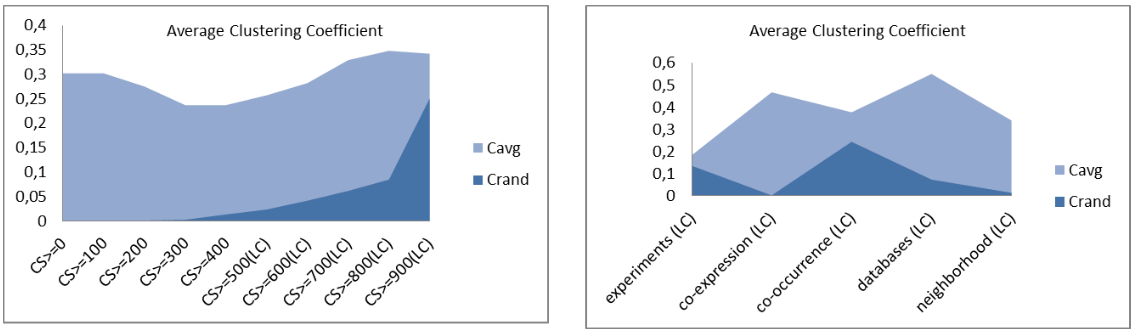


Figure 4.1: C_{avg} of the studied PPI networks comparison with the corresponding C_{rand} with different confidence scores and different prediction methods.

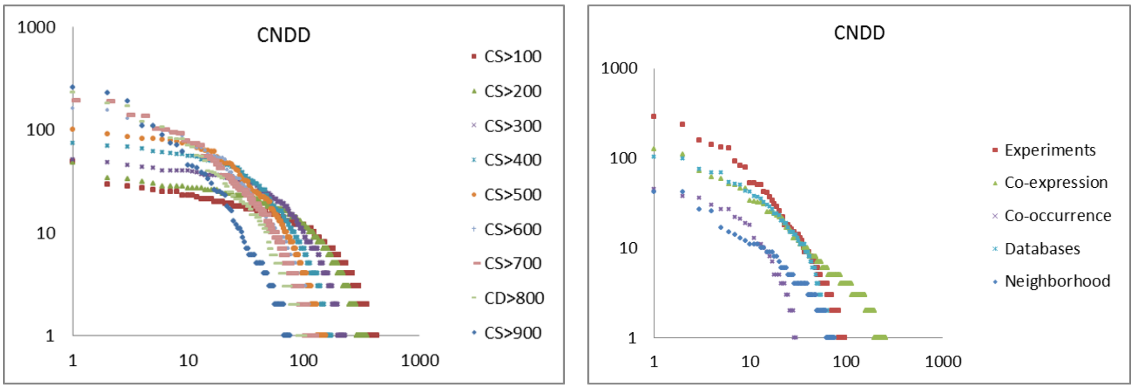


Figure 4.2: Cumulative node degree distribution with logarithmic binning of the studied PPI networks with different confidence scores and different prediction methods.

high degree of modularity and with some hierarchical organization. The small diameter also indicates small world properties.

While exploratory, this study aims to contribute to a better understanding of the human oral biology as a subsystem of the human biological system not yet totally known. These topological properties can be used to find new interactions and detect false positives in order to get better models of the biological systems studied.

Table 4.5: Basic parameters of the degree distribution of the studied PPI networks.

Description	n	<x>	x_{max}	x_{min}	$x_{min\ err}$	α	α_{err}	p
CS≥100	439	7.11	49	11	0.75	3.50	0.10	0.180
CS≥200	352	8.61	48	15	4.04	3.50	0.57	0.006
CS≥200	352	8.61	48	15	4.04	3.50	0.57	0.006
CS≥300	230	13.01	51	5	7.06	1.81	0.68	0.000
CS≥400	175	16.93	73	4	8.02	1.69	0.53	0.000
CS≥500	156	18.69	101	17	10.38	2.17	0.51	0.000
CS≥600	133	21.32	159	7	12.05	1.76	0.45	0.004
CS≥700	117	22.03	192	15	7.94	2.05	0.32	0.298
CS≥800	103	22.90	231	12	6.32	1.95	0.25	0.350
CS≥900	76	24.24	255	2	8.97	1.53	0.30	0.056
Experiments(LC)	97	23.41	290	9	5.16	1.78	0.20	0.200
Co-expression(LC)	256	72.70	126	3	0.64	2.06	0.10	0.667
Co-occurrence(LC)	30	13.10	45	4	6.45	1.77	0.72	0.021
Databases(LC)	75	17.88	103	11	8.73	2.08	0.52	0.108
Neighborhood(LC)	75	12.21	42	8	2.48	2.57	0.47	0.601

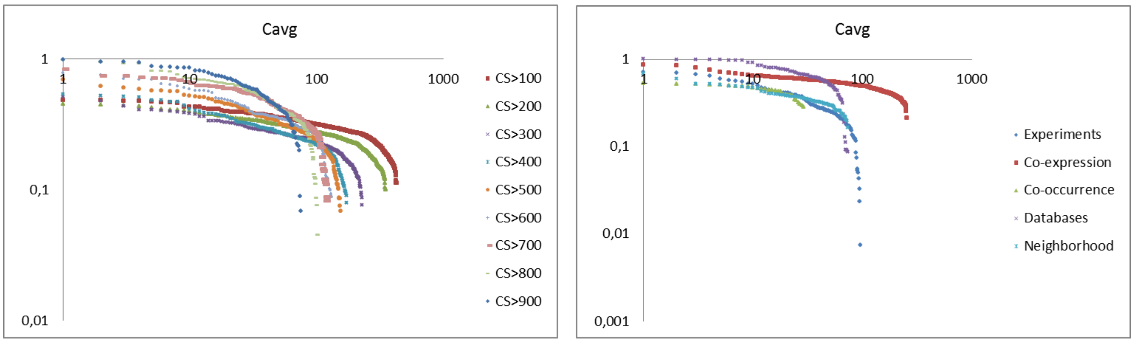


Figure 4.3: Cumulative clustering degree distributions with logarithmic binning of the studied PPI networks, with different confidence scores and different prediction methods.

4.3 Analysis of Genes Co-Expression Networks

This section is based on the work published in [70] and consisted in the analysis of a genes co-expression network, where the WGCNA method is applied to HNSCC data. Topological properties from weighted gene co-expression correlation networks built from the HNSCC genes data were used to relate traits with groups of genes that could influence this disease.

Weighted gene co-expression correlation networks have been applied in several medical areas [70, 85, 93].

HNSCC is the sixth most common cancer worldwide, affecting 600,000 new patients each year [94]. Several risk factors such as smoking habits, alcohol use, and human papillomavirus infection have already been documented as having a very high correlation with this type of cancer [95, 96]. Despite that, still lacks a full comprehension of genomic processes that are associated with HNSCC and more importantly the individual contribution of each of these factors, when crossed with epidemiological characteristics and the existence of other risk factors associated with this disease.

4.3.1 Data Sets

This research aims to contribute to reveal the molecular mechanisms associated with HNSCC and the contribution of other risk factors besides smoking habits and alcohol use, like differentiation, sex, age, tumor site and race with a major focus in the age and alcohol use experimental factor types [70]. This work addresses this problem by adjusting the use of gene co-expression networks to analyze a HNSCC dataset.

The HNSCC expression dataset was downloaded from the public micro-array gene expression database ArrayExpress [97, 98], from the investigation E-GEOD-39366 - Molecular Subtypes in Head and Neck Cancer [expression].

A total of 138 tumor arrays were considered from the 163 samples, after removing low-quality and duplicate arrays, and arrays from non-HNSCC samples. Probes produced expression values for 15,595 genes. Experimental factor types considered are: differentiation, alcohol use, sex, age, tumor site, smoking pack years and race with a major focus in age and alcohol use experimental factor types.

To extract the biological meaning of the genes was used the database for annotation, visualization, and integrated discovery (DAVID) bioinformatics resource, which is an integrated biological knowledge-base with a large list of genes and proteins, and a list of gene ontology functional terms. This resource includes also data mining tools [99, 100].

4.3.2 Weighted Gene Co-Expression Analysis Approach

Many biological networks are scale-free, meaning that their degree distribution follows a power-law [16, 41, 43, 88]. Scale-free networks are extremely heterogeneous and this topology is linked to the evolution of biological systems and correspondent growth of the network where preferentially attachment is present. Genes that are co-expressed are more likely to encode interacting proteins and the most used measure of co-expression is the Pearson correlation coefficient, which assumes linearity and after using a threshold value it is used to build co-expression networks, also called by some authors "relevance networks" [101]. In co-expression networks, nodes are connected if they have a significant pairwise expression profile association across some condition.

In this study it was used the WGCNA methodology [44]. The co-expression network construction from gene expression data uses correlation analysis to build the correlation matrix, which is converted to an adjacency matrix representing the co-expression network. Each gene corresponds to a node and two genes are connected by an edge if their expression values are highly correlated.

A co-expression network can be represented by a symmetric adjacency matrix, $\mathrm{A} = [\mathrm{a_{ij}}]$ with values in $[0, 1]$. For weighted networks, the adjacency matrix returns the connection strength between gene pairs and, as gene co-expression similarity measure, can be used the absolute value of the Pearson product moment correlation to relate every pairwise gene–gene relationship.

$$\mathrm{a_{ij}}= |\mathrm{cor}\,(\mathrm{x_i}, \mathrm{x_j})| \tag{4.1}$$

An adjacency function can be used to transform the original network into a new network. For the construction of weighted gene co-expression networks [44], the adjacency matrix is constructed using a "soft" power adjacency function $\mathrm{a_{ij}}$, where for an unsigned network

$$\mathrm{a_{ij}}=|\mathrm{cor}\,(\mathrm{x_i}, \mathrm{x_j})|^{\beta} \tag{4.2}$$

A choice of a power $\beta >\ 1$ is used to emphasize large adjacencies at the expense of low ones. To choose the parameter value β, it is used the scale-free topology criterion, being β the value obtained through the trade-off between the lowest integer such that the resulting network satisfies approximate scale-free topology with the highest mean number of connections.

A clustering method is afterwards used to find network modules, that can be represented by their eigengenes, which will be used to build meta-networks, that are networks of modules. These meta-networks can then be related to external information.

It can be defined the gene significance (GS) based on a micro-array sample risk factor, defining gene significance measure as a function GS that assigns a nonnegative number to each gene. The higher $\mathrm{GS_i}$ the more biologically significant is gene i. Risk factor based gene significance is defined as (the absolute value of) the correlation between the gene and the risk factor.

Considering a significance measure $\mathrm{GS} = (\mathrm{GS_1},\ \ldots, \mathrm{GS_n})$ as a vector with n components, corresponding to a node, $\mathrm{GS_i}$ for node i quantifies the significance or importance with regard to a particular item under consideration. Node significance does not necessarily correspond to statistical significance. If a statistical significance ($\mathrm{p-value}$) is available for each node, then the significance measure based on the p-value can be defined as

$$\mathrm{GS_i} = -\log(\mathrm{p-value_i}) \tag{4.3}$$

The gene significance measure allows to incorporate external gene information into network analysis [44, 93].

A quantitative micro-array sample trait $\mathrm{t} = (\mathrm{t_1}, \mathrm{t_2}, \ldots, \mathrm{t_m})$ can be used to define a trait based gene significance measure. For example, a trait-based node significance measure can be defined as the absolute values of the correlation between the $\mathrm{i^{th}}$ node profile and the sample trait t

$$\mathrm{GS_i} = |\mathrm{cor}\,(\mathrm{x_i}, \mathrm{t})\,| \tag{4.4}$$

where cor is the Pearson correlation. Alternatively a correlation test $\mathrm{p-value}$ or a regression-based $\mathrm{p-value}$ for attesting the statistical significance between $\mathrm{x_i}$ and the sample trait t can be used [93].

A signed significance measure is

$$\mathrm{GS_i} = \mathrm{cor}\,(\mathrm{x_i}, \mathrm{t}) \tag{4.5}$$

The interpretation of gene co-expression relationships depends on the biological context and the identified co-expression modules may also have functional interpretations [17, 44, 84, 85].

The WGCNA approach uses gene co-expression networks to study gene expression data and is defined by the following steps [44, 70]:

1) Definition of a gene co-expression similarity;

2) Definition of a family of adjacency functions;

3) Determination of adjacency functions parameters and of a node dissimilarity measure ;

4) Identification of the network modules using clustering;

5) Association of network concepts and association of these concepts to external gene or sample information.

Genes Co-expression Similarity

To measure the level of concordance between gene expression profiles across experiments is used a measure of similarity between the gene expression profiles.

The $n \times n$ similarity matrix $S = [s_{ij}]$ is transformed into an $n \times n$ adjacency matrix $A = [a_{ij}]$, which encodes the connection strengths between pairs of nodes. A is an undirected and symmetric matrix with non-negative entries and the diagonal elements of A are set to 0. For weighted networks $a_{ij} \in [0,\ 1]$ [44].

If x_i is a vector with m components containing the i^{th} gene expression profiles (where $i = 1, ..., n$) across m micro-arrays, two different measures of co-expression similarity can be used to compare a pair of gene expression profiles x_i and x_j. The first measure $S = [s_{ij}]$ is the absolute value of the Pearson correlation coefficient.

$$s_{ij} = |cor(x_i,\ x_j)| \tag{4.6}$$

The second measure S_{signed} is a linear transformation of the correlation that retains its sign:

$$s_{ij}^{signed} = \frac{1 + cor(x_i, x_j)}{2} \tag{4.7}$$

s_{ij}^{signed} equals 1, 1/2, and 0 if the correlation equals 1, 0 and −1, respectively.

Adjacency Functions Family

To define the adjacency matrix, it is used an adjacency function. This function transforms co-expression similarities into connection strengths. The adjacency function is used to transform the original network into a new network and has certain parameters, which can be determined using different statistical or biological criteria [44].

The co-expression similarities can be transformed into a weighted gene co-expression network using the power transformation, also known as soft-threshold function [44, 84]

$$a_{ij}^{unsigned,\ weighted} = power_{ij}\ (S,\beta) = s_{ij}^{\beta} \tag{4.8}$$

$$a_{ij}^{signed,\ weighted} = power_{ij} (S_{signed},\beta) = s_{signed,ij}^{\beta} \tag{4.9}$$

The weighted adjacency between two genes can be defined as a power $\beta \geq 1$ of the absolute value of the correlation coefficient. This way, strong correlations were privileged to weak correlations to minimize noise and due to the small number of samples compared to the number of genes [44].

The correlation network adjacencies can be unsigned and signed respectively

$$A_{ij}^{unsigned,\ weighted} = (\ |cor\ (x_i, x_j)|\ )^{\beta} \tag{4.10}$$

$$A_{ij}^{signed,\ weighted} = (0.5 + 0.5cor\ (x_i, x_j)\ )^{\beta} \tag{4.11}$$

The use of signed or unsigned networks depend on the application [44, 84].

The resulting adjacency matrix is used to define a distance, i.e., a measure of node dissimilarity, used as input of a clustering method to define network modules. Once the modules have been defined, additional network concepts can be defined, like the intra-modular connectivity and modules and their hub genes can be related to external gene information [44].

The “soft” threshold weighs each connection by a number in $[0, 1]$ and there is empirical evidence that weighted networks can yield more robust results than unweighted networks, being biologically more meaningful to encode gene co-expression.

Adjacency Functions Parameters

Biological networks are robust, random failures tolerant and approximate scale-free networks. These properties can be used to choose the threshold. The linear regression model fitting index R^2 can be used to quantify how well a network satisfies a scale-free topology. However, there is a trade-off between maximizing scale-free topology model fit (R^2) and maintaining a high mean number of connections [44, 102–104].

The choice of the adjacency functions parameters determines the connectivity patterns and the topological properties of the network.

The mean connectivity criterion selects adjacency function parameters, such that the mean connectivity takes a given value

$$\mathrm{mean}\,(\mathrm{k}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\mathrm{k_i}}{\mathrm{n}} \tag{4.12}$$

The mean connectivity of a weighted or unweighted network is a monotonically decreasing function of the adjacency function parameters. The higher the parameter value the lower is the mean connectivity. Very low or very high values of mean connectivity of a network lead to an uninformative network [17, 44, 84].

High values of power adjacency function parameter β lead to low values of the connectivity, density, maximum adjacency ratio and clustering coefficient. For most network the higher the β the larger the heterogeneity and the module separability.

The value of the parameter β is obtained through the trade-off between the lowest integer such that the resulting network satisfies approximate scale-free topology (linear model fitting index R^2 of the regression line between logp(k) and log(k) larger than 0.8) with the highest mean number of connections. The mean connectivity should be high so that the network contains enough information (e.g., high power for detecting modules, clusters of genes and hub genes) [44, 70, 84, 105].

Network Modules

An important aim of co-expression network analysis is to detect subsets of nodes, modules, that are tightly connected to each other, which may encode pathways or protein complexes [17, 44].

Modules can be identified using methods that are based on node dissimilarity measures and clustering methods. Modules can be found using hierarchical clustering, like in this study. Hierarchical clustering is a method of cluster analysis that constructs a hierarchy of clusters. It groups different genes, from the co-expressed networks, together by observing their common properties in a systemic view. This can help to find genes that are co-expressed [106]. In this study, hierarchical clustering takes a dissimilarity measure as input.

There are several dissimilarity measures, being one of them the topological overlap dissimilarity measure since it was found to result in biologically meaningful modules [44, 104].

The topological overlap of two nodes reflects their relative interconnectedness and the topological overlap matrix (TOM) can be defined as

$$\mathrm{TOM_{ij}} = [\mathrm{w_{ij}}] \tag{4.13}$$

TOM is a similarity measure, since it is non-negative and symmetric, and the corresponding dissimilarity measure is

$$\mathrm{d^{w}_{ij}} = 1 - \mathrm{w_{ij}} \tag{4.14}$$

Modules in weighted gene co-expression network are defined as groups of highly correlated genes with high topological overlap. A pair of genes is said to have high topological overlap if they are both strongly connected to the same group of genes. The use of topological overlap is a filter to exclude very weak connections during network construction. The TOM transformation can lead to a more robust network and larger modules [17, 44, 70].

The TOM-based measure of connectivity $\mathrm{w_i}$ is

$$\mathrm{w_i} = \sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{w_{ij}} \tag{4.15}$$

where $\mathrm{w_{ij}}$ is the topological overlap between two nodes i and j. Thus, a node has high TOM-based connectivity if it has high overlap with many other nodes.

A network connectivity measure can be defined with respect to the whole network (whole-network connectivity) or with respect to the genes of a particular module (intra-modular connectivity) [44].

The topological overlap transformation TOM(A) replaces each adjacency $\mathrm{a_{ij}}$ by a normalized count of neighbors shared by the nodes i and j. In an unweighted network, the number of shared neighbors of genes i and j is given by $\sum_{\mathrm{u}\neq \mathrm{i,j}} \mathrm{a_{iu}a_{ju}}$, and for a weighted network with an adjacency matrix A, TOM is defined as

$$\mathrm{TOM_{ij}(A)} = \frac{\sum_{\mathrm{u}} \mathrm{a_{iu}a_{uj}} + \mathrm{a_{ij}}}{\min\left(\mathrm{k_i}, \mathrm{k_j}\right) + 1 - \mathrm{a_{ij}}} \tag{4.16}$$

$\mathrm{TOM_{ij}}$ is a value in $[0,1]$ and $\mathrm{TOM_{ij}} = \mathrm{TOM_{ji}}$ [84]. $\mathrm{TOM_{ij}(A)}$ also satisfies the conditions of an adjacency matrix [17, 44, 84].

The topological overlap based dissimilarity measure is

$$DissTOM_{ij} = 1 - TOM_{ij} \tag{4.17}$$

There are several approaches for defining network modules and one of them is that modules can be defined as clusters that result from using a pairwise node dissimilarity as input of the average linkage hierarchical clustering. Branches in the resulting cluster tree (dendrogram) are the modules, and different branch cutting techniques can be used. For example, using a constant height cut-off value, or using an algorithm for the selection of the height cut-off value. The Dynamic Tree Cutting algorithm adaptively chooses cutting values depending on the shape of the branches. This module detection method allowed to obtain biologically meaningful modules [70, 84].

Modules in gene co-expression networks tend to be approximately factorisable if the corresponding expression profiles are highly correlated. Approximate factorisability is a very strong structural assumption on an adjacency matrix and does not hold for general networks. There is empirical evidence that many clusters (modules) of genes or proteins in real networks are approximately factorisable. In PPI networks, only after replacing the original adjacency matrix by the topological overlap matrix, the resulting modules are approximately factorisable [17].

With intra-modular networks concepts, the topological properties within a module can be measured, and with inter-modular network concepts, the topological properties among modules can be measured [17, 44].

The generalized topological overlap GTOM is a generalization of TOM considering longer ranging relationships between nodes [107].

Co-expression modules may form a biologically meaningful meta-network showing a higher-order organization of the transcriptome. Modules in a meta-network of modules are designated as meta-modules [84].

An eigengene network reduces a gene co-expression network involving thousands of genes to a meta network involving module representatives (one eigengene per module). Eigengene networks can be used for the analysis of one eigengene network or of several eigengene networks. They can be used to compare module relationships across different data sets and for differential eigengene network analysis [84], which is a much smaller network.

The module eigengene [84] summarizes the expression profiles of a module. It is the most highly connected intra-modular hub gene and allows treating modules as single units. To define the module eigengene of a module, it can be used the singular value decomposition (SVD) of the module expression matrix. The gene expression matrix of the $\mathrm{I^{th}}$ module is denoted by

$$X^{(I)} = (x_{il}^{(I)}) \tag{4.18}$$

where the index $\mathrm{i} = 1, 2, ..., \mathrm{n_I}$ corresponds to the module genes and the index $\mathrm{l} = 1,\ 2,\ ...,\ \mathrm{m}$ corresponds to the micro-array samples. We assume that each gene expression profiles $\mathrm{x_i^{(I)}}$, i.e. each row of $\mathrm{X^{(I)}}$, has mean 0 and variance 1. The singular value decomposition of $\mathrm{X^{(I)}}$ is denoted by

$$\mathrm{X^{(I)}{=}UDV^T} \tag{4.19}$$

where the columns of the orthogonal matrices U and V are the left and right singular vectors, respectively. Specifically, $\mathrm{U^{(I)}}$ is a $\mathrm{n^{(I)} \times m}$ matrix with orthonormal columns, $\mathrm{V^{(I)}}$ is an $\mathrm{m \times m}$ orthogonal matrix, and $\mathrm{D^{(I)}}$ is an $\mathrm{m \times m}$ diagonal matrix of singular values $\{|\ \mathrm{d_l^{(I)}}|\}$. The matrices $\mathrm{V^{(I)}}$ and $\mathrm{D^{(I)}}$ are given by [84]

$$\mathrm{V^{(I)}{=}\ (v_1^{(I)}\ v_2^{(I)}\ \cdots v_m^{(I)})} \tag{4.20}$$

$$\mathrm{D^{(I)}{=}\left(\left|d_1^{(I)}\right|,\ \left|d_2^{(I)}\right|,\ \cdots,\ \left|d_m^{(I)}\right|\right)} \tag{4.21}$$

It is assumed that the singular values $\left|\mathrm{d_l^{(I)}}\right|$ are arranged in non-increasing order. The first column of $\mathrm{V^{(I)}}$ is the module eigengene [84].

$$\mathrm{E^{(I)}{=}v_1^{(I)}} \tag{4.22}$$

In a network of modules each node in the network corresponds to a module. If $\mathrm{q_1}$ and $\mathrm{q_2}$ are two modules and $\mathrm{M_{q_1}}$ is the set of $\mathrm{n^{(q_1)}}$ nodes inside the module $\mathrm{q_1}$ the adjacencies between the nodes of the two modules can be represented by an $\mathrm{n^{(q_1)}{\times}n^{(q_2)}}$ dimensional sub-matrix $\mathrm{A^{(q_1,q_2)}}$ of the full adjacency matrix A.

A measure (a number between 0 and 1) of adjacency between the two modules $\mathrm{A^{(q_1,q_2)}}$ can be calculated as

$$\mathrm{A_{q_1,q_2}^{avg}{=}\ mean\left(A^{(q_1,q_2)}\right) {=} \frac{\sum_{i\in M_{q_1}}\sum_{j\in M_{q_2}} A_{ij}}{n^{(q_1)}n^{(q_2)}}} \tag{4.23}$$

$\mathrm{A^{avg}}$ uses average, but other measures like the maximum or the minimum can be used depending on the applications.

This measure can be used to define a network between modules, as

$$\mathrm{A}_{\mathrm{q}_1,\mathrm{q}_2}=\begin{cases} \mathrm{A}^{\mathrm{avg}}_{\mathrm{q}_1,\mathrm{q}_2}\text{if } \mathrm{q}_1\neq\mathrm{q}_2 \\ 1 \text{ if } \mathrm{q}_1=\mathrm{q}_2 \end{cases} \tag{4.24}$$

$\mathrm{A}_{\mathrm{modules}}$ is the $\mathrm{Q}\times\mathrm{Q}$ dimensional symmetric matrix whose $\mathrm{q}_1,\mathrm{q}_2$ element is given by $\mathrm{A}_{\mathrm{q}_1,\mathrm{q}_2}$ (which measures the adjacency between the two modules). The diagonal elements of $\mathrm{A}_{\mathrm{modules}}$ are set to 1. $\mathrm{A}_{\mathrm{modules}}$ can be interpreted as an adjacency matrix between modules, i.e., it represents a weighted network whose nodes are modules.

Measures of adjacency between the two modules q_1 and q_2 can be defined based on $\mathrm{A}^{(\mathrm{q}_1,\mathrm{q}_2)}$, for e.g., $\mathrm{A}^{\mathrm{avg}}_{\mathrm{q}_1,\mathrm{q}_2}=$ mean $\left(\mathrm{A}^{(\mathrm{q}_1,\mathrm{q}_2)}\right)$ or alternatively if it can be used the eigenvector $\mathrm{E}^{(\mathrm{q}_1)}$ and $\mathrm{E}^{(\mathrm{q}_2)}$ of the respective modules to define an eigenvector measure of inter-modular adjacency [84], as

$$\mathrm{A}_{\mathrm{q}_1\mathrm{q}_2}=\left|\mathrm{cor}(\mathrm{E}^{(\mathrm{q}_1)},\mathrm{E}^{(\mathrm{q}_2)})\right|^{\beta} \tag{4.25}$$

$$\mathrm{A}^{\mathrm{ave}}_{\mathrm{q}_1,\mathrm{q}_2}\approx\left|\mathrm{cor}\left(\mathrm{E}^{(\mathrm{q}_1)},\mathrm{E}^{(\mathrm{q}_2)}\right)\right|^{\beta}=\mathrm{A}_{\mathrm{q}_1\mathrm{q}_2} \tag{4.26}$$

An eigenvector network is defined as a correlation network between module eigenvectors. Detecting a high correlation between module eigenvectors may be of biological interest (interactions between pathways) or may mean poorly defined modules that should be merged. The correlations among eigengenes in gene co-expression networks have been used to define several biological eigengene networks [70, 84].

The dissimilarity of two modules q_1 and q_2 can be calculated by

$$\mathrm{diss}\left(\mathrm{q}_1,\mathrm{q}_2\right)=1-\mathrm{cor}\left(\mathrm{E}^{\{(\mathrm{q}_1)\}},\mathrm{E}^{\{(\mathrm{q}_2)\}}\right) \tag{4.27}$$

and the eigengene network can be defined as the signed correlation network

$$\mathrm{A}_{\{\mathrm{q}_1,\mathrm{q}_2\}}=0.5+0.5\ \mathrm{cor}\left(\mathrm{E}^{\{(\mathrm{q}_1)\}},\mathrm{E}^{\{(\mathrm{q}_2)\}}\right) \tag{4.28}$$

The module membership (MM) measures how closely related a particular gene is to eigengenes in co-expression networks, and can be defined as the correlation between the expression profile of the studied gene x_j and the eigengene E_I,

$$\mathrm{k}_{\mathrm{E}_\mathrm{I}(\mathrm{j})}=\mathrm{cor}(\mathrm{x}_\mathrm{j},\mathrm{E}_\mathrm{I}) \tag{4.29}$$

The closer $\mathrm{k}_{\mathrm{E}_\mathrm{I}(\mathrm{j})}$ is to 1 or -1, the stronger the evidence that the j^{th} gene is part of the I^{th} module [77].

Network Concepts and External Information

Because eigengene networks are orders of magnitude smaller than the original gene co-expression networks the dissimilarity based on the topological overlap for finding meta-modules is not used and instead, it can be used the following dissimilarity [84]

$$\mathrm{diss_{ij}}\left(\mathrm{A_{eigen}}\right) = \frac{1-\mathrm{cor}\left(\mathrm{E_I},\mathrm{E_J}\right)}{2} \tag{4.30}$$

Modules can be compared and consensus modules may represent biological pathways shared. Also, a consensus dissimilarity measure that compares topological overlap matrices, can be used as input to hierarchical clustering [84].

Functional enrichment analysis using gene ontology information was used to help the understanding of the biological meaning of the consensus modules. It can also include clinical characteristics information and relate it with the consensus modules obtained [84, 105].

Finally, network concepts for comparing two networks can be used to choose the parameter values of an adjacency function. Parameters can be chosen such that

$$\mathrm{A^{[test]}} \approx \mathrm{A^{[ref]}} \tag{4.31}$$

being $\mathrm{A^{[ref]}}$ and $\mathrm{A^{[test]}}$ two $\mathrm{n} \times \mathrm{n}$ dimensional adjacency matrices. For example, these matrices can define the connectivity patterns among genes before and after a biological experiment, if a reference network $\mathrm{A^{[ref]}}$, based on prior knowledge about the connectivity of the nodes, is available.

If $\mathrm{NC^{[ref]}}$ and $\mathrm{NC^{[test]}}$ represent the values of a network concept in the reference and test network respectively, the differential network concept is defined as:

$$\mathrm{Diff.NC} = \mathrm{NC^{[ref]}} - \mathrm{NC^{[test]}} \tag{4.32}$$

For example, for the scaled connectivity, it can be calculated by

$$\mathrm{Diff.K} = \mathrm{K^{[ref]}} - \mathrm{K^{[test]}} \tag{4.33}$$

Ranking the nodes according to a suitable defined differential network concept (or more than one), allows finding nodes that have differential connectivity patterns across two networks [84, 105].

For measuring the similarity between two networks it can be measured whether $\mathrm{NC^{[ref]}}$ is correlated with $\mathrm{NC^{[test]}}$ across nodes and Pearson correlation can be used

$$\mathrm{cor.NC} = \mathrm{cor}\left(\mathrm{NC^{[ref]}}, \mathrm{NC^{[test]}}\right) \tag{4.34}$$

For example, the similarity of the connectivity correlation between $\mathrm{A^{[ref]}}$ and $\mathrm{A^{[test]}}$ can be calculated by

$$cor.K = cor(K^{[ref]}, K^{[test]}) \tag{4.35}$$

Other network concept correlations can be used. Reference and test networks exhibit similar patterns, if cor has a high value [84, 105].

A sample trait $t = (t_1,\cdots,t_m)$ can also be used to define a trait-based eigenvector significance measure using also a correlation test [70, 84].

4.3.3 Results and Conclusion

The methodology used to analyze expression data was the WGCNA [44] that follows the five steps, described in detail in the previous section. To identify gene modules was used the hierarchic clustering. This method calculates the eigengenes of each module to define a network of modules and the correlation between the eigengenes and the risk factors, identifying modules of genes where those are more expressed and associating these concepts to gene ontology functional terms. The preliminary results described in this paper contribute to reveal the molecular mechanisms associated with HNSCC and the contribution of experimental factors types like differentiation, alcohol use, sex, age, tumor site, smoking pack years and race.

When applying the WGCNA to the HNSCC dataset, scale-free topology criterion was used to choose the power β for the unsigned weighted correlation network and it was chosen $\beta = 5$. The scale-free topology plot of the weighted HNSCC co-expression network constructed with power $\beta = 5$, satisfies a scale-free topology approximately with $R^2 = 0.96$, a value close to 1. Networks whose scale-free topology index R^2 is close to 1 are said to be approximately scale-free.

It was defined a topological overlap matrix and constructed a hierarchical tree (average linkage) to define modules as branches of the tree. Eigengenes for each module were calculated and a network among modules was defined, where each node of the network correspond to a module (see Figure 4.4). It was constructed a hierarchical clustering dendrogram of the eigengenes E and a heat map to visualize the eigengene network defined by the signed correlation network. Modules highly correlated are similar and can be merged.

Multidimensional scaling can be used to visualize pairwise relationships specified by a dissimilarity matrix, where each row of the dissimilarity matrix is a point in a Euclidean space and the Euclidean distances between a pair of points reflect the corresponding pairwise dissimilarity. The input is the TOM dissimilarity and each dot is colored by the corresponding module assignment (see Figure 4.5). Colors from each module are well separated, showing distinct modules.

To identify modules associated with the risk factors and because each eigengene is a summary of the expression profiles of the respective module, eigengenes and risk factors were correlated. Each row corresponds to a module eigengene, and each column to a risk factor. Each cell contains the corresponding correlation and p-value. The table is color-coded by

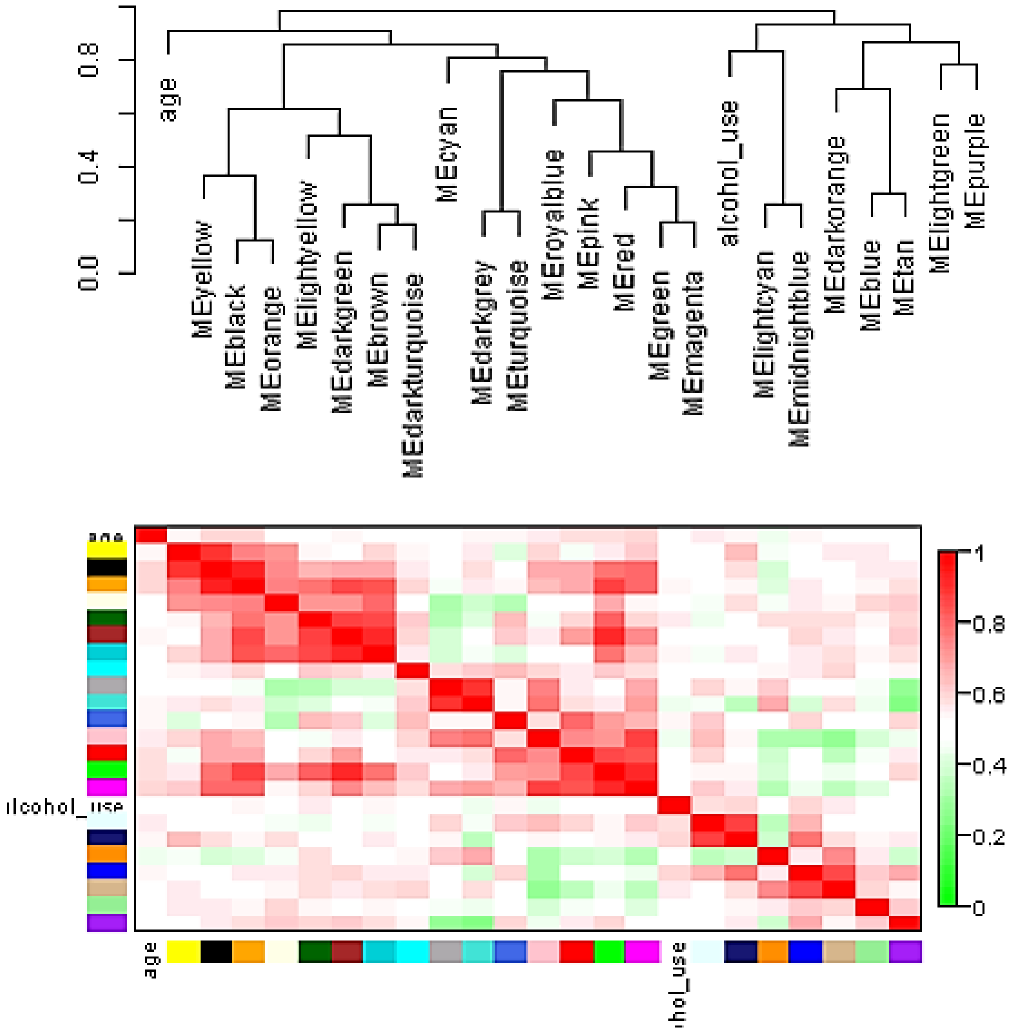


Figure 4.4: Visualization of the eigengene network representing the relationships among the modules and the age and alcohol use.

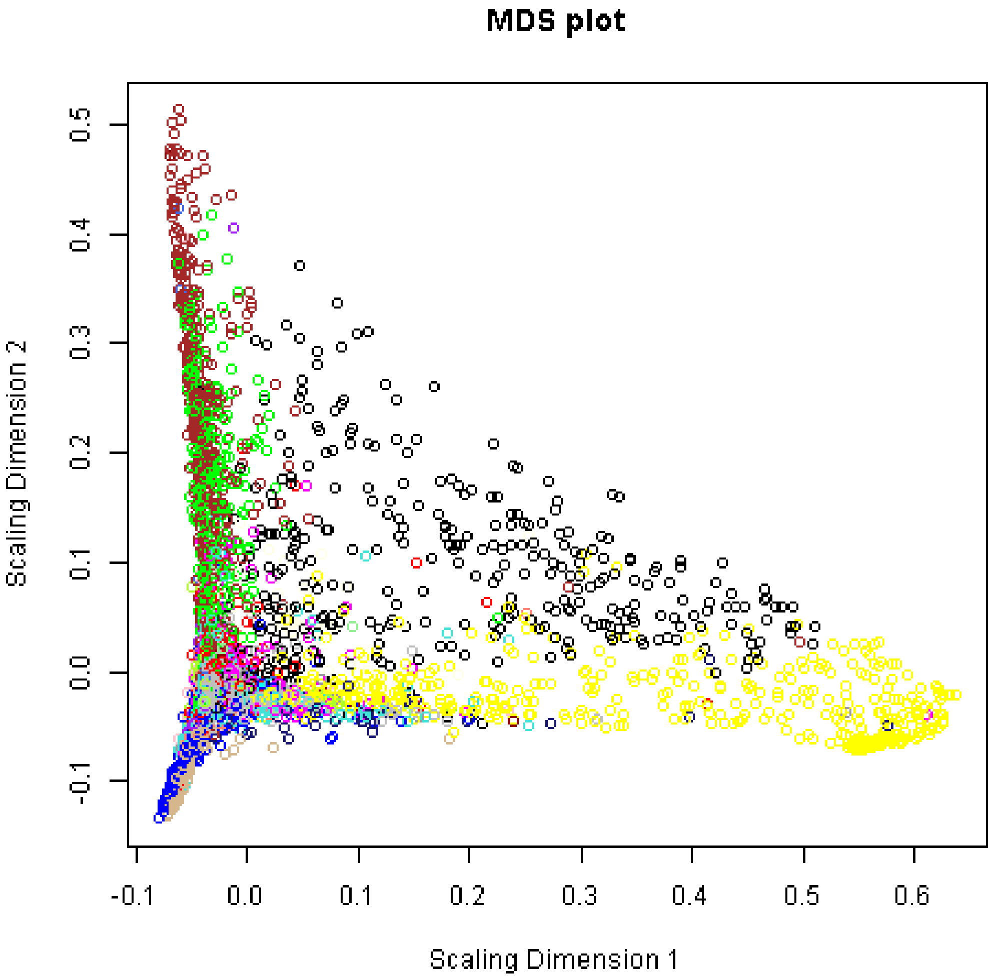


Figure 4.5: Multidimensional scaling.

correlation according to the colour legend. Age is more correlated with the magenta, black, green and light green modules, and alcohol use with blue, light cyan, tan and pink modules (see Figure 4.6). Two different experimental factors were correlated with different modules (different genes) in this type of cancer.

The correlations between age and alcohol use and the respective module eigengenes can be measured using gene significance (GS) and module membership (MM) to identify genes with high significance for age and alcohol use and high module memberships in the identified modules (see Figure 4.6).

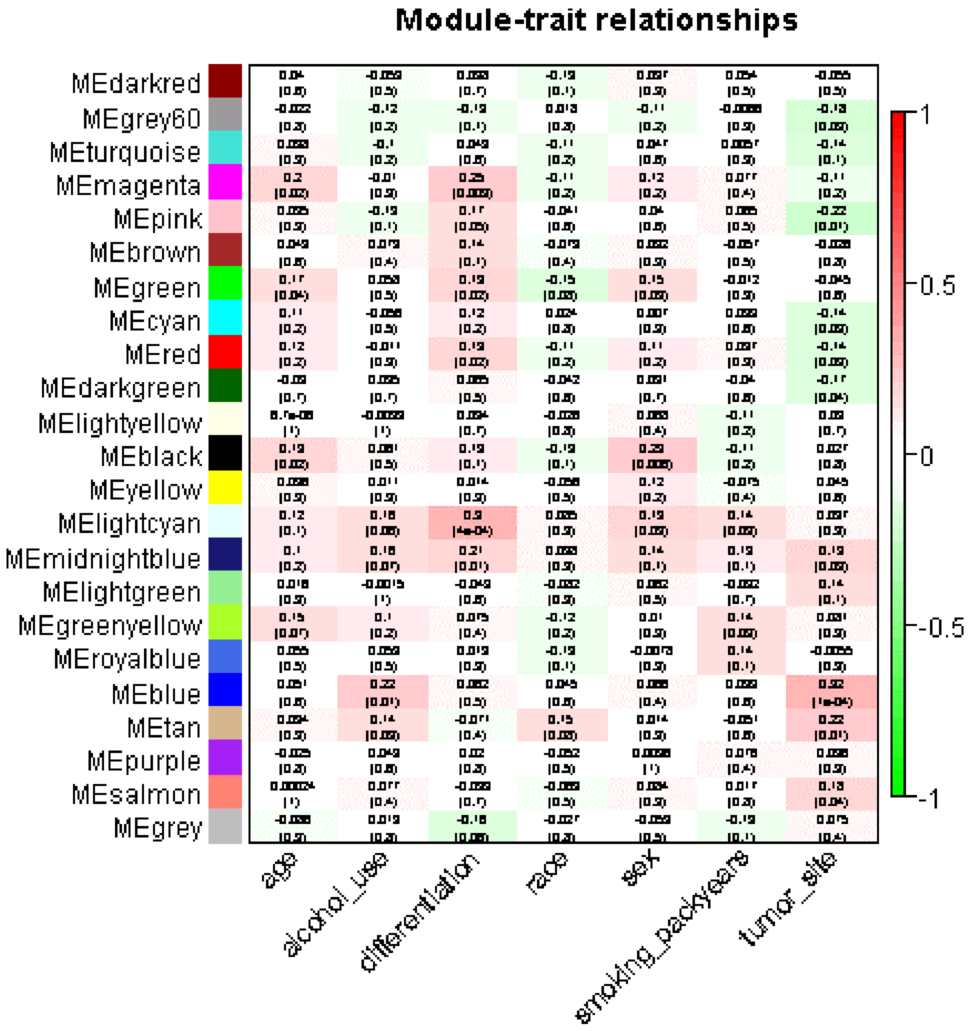


Figure 4.6: Module-risk factor associations.

Gene ontology analysis was performed using DAVID [100] for two of the modules, modules black and green. These two modules were more correlated with age and the results obtained were:

- For the black module, considering the three correlation higher values with the age risk factor, respectively of 0.233; 0.190; 0.16 - tyrosine kinase, non-receptor, 2; peptide YY, 2 (seminal plasmin); and oxytocin;

- For the green module, considering the three correlation higher values with the age risk factor, respectively: 0.218; 0.216; and 0.215) - family with sequence similarity 89, member A; hypothetical protein LOC100134229; and Rap guanine nucleotide exchange factor (GEF).

Results show that gene expression profiles across samples can be highly correlated [44]. Gene co-expression networks were defined as weighted correlation networks, to preserve the continuous nature of the co-expression information, where strong correlations were privileged relatively to weak correlations, to minimize noise and due to the small number of samples compared to the number of genes. The quantitative micro-array sample risk factor was used to define the risk factor-based gene significance measure.

It can be also noticed that this methodology allows the identification of distinct modules, being co-expression modules summaries of interdependencies, through the eigengenes modules.

Correlations between risk factors and HNSCC gene expression data modules were quantified, but some physiological risk factors, like race, showed no correlation with HNSCC. The analysis for this disease was mainly focused in the risk factors age and alcohol use, which were more correlated with different sets of modules from the HNSCC gene expression dataset (see Figure 4.7).

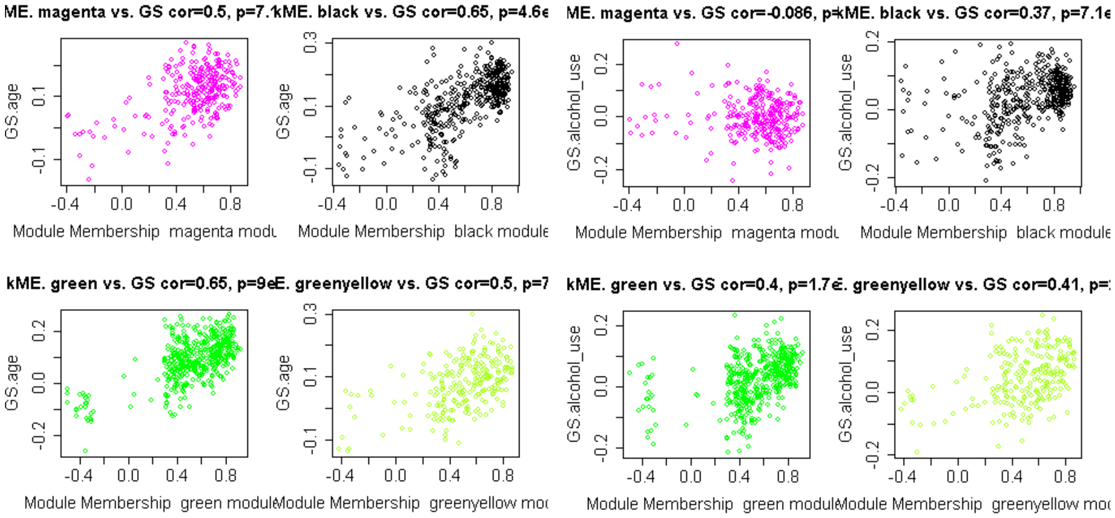


Figure 4.7: Gene significance versus module membership for the risk factor age and alcohol use.

A preliminary gene ontology analysis listed functions for the genes of identified modules.

As an example, the functions associated with genes with the three higher values of correlation in two of the modules more correlated with the risk factor age were listed.

These studies corroborate the statement that the identification of signatures through the study of the biological network topological properties can help the clinical identification of diseases. It is possible to relate risk factors with topological patterns found in the co-expression networks, which allows to study the evolution of diseases in different groups of the population and determine how these risk factors can alter the network topology.

4.4 Summary

Two investigations were described in this chapter. The first consisted of the characterization of PPI network models of the oral proteome. These models were constructed with different confidence scores and different prediction methods. These networks were characterized using several topological properties, previously defined, and the parameters of their degree distribution were also analyzed. It was verified that most of these networks are scale-free networks, with a high degree of modularity and some hierarchical organization. The second investigation consisted in the study of molecular mechanisms associated to the HNSCC disease and the contribution of existing risk factors in patients with this disease. For this investigation, a co-expression network was constructed and the WGCNA methodology was used, which uses as a biological criterion to consider that the networks are approximately scale-free. Relationships between gene modules and some risk factors of patients with this disease were identified and a gene ontology analysis was made for two of the modules that were most correlated with risk factor age.

Chapter 5

Denoising Protein Interactions Networks Using the Topology

Processes associated with mechanisms of life can be understood studying the various relationships existent between the entities of biological organisms. These relationships can be modeled using networks and therefore studied using graph theory. These networks can be characterized through the quantification of their topological properties, which allows to unveil snapshots of their structure and dynamics.

The process of modeling biological organisms through biological networks is noisy due to several motives, like the precision of the equipment used to obtain data, the limitations of the used methodologies and also the yet unknown knowledge about all of the processes of the living organisms. Consequently, in these biological networks models there are missing interactions and also there are represented interactions that do not exist.

To denoise biological networks, meaning that, to find interactions to withdraw because they are not real, and to find missing interactions, is relevant to obtain more precise models of the processes presented in the real networks. For that, a set of topological properties can be selected and quantified in the network models to identify patterns or distribution trends present in real networks that could be used to denoise the networks.

This chapter includes a review of current denoising methods, followed by the description of a new denoising methodology, named organization measurement (OM) method and based uniquely on the topology of the biological networks. Here are also described the topological properties used when the methodology was tested, including the proposed new topological measure, called neighborhood clustering (NC). Next, the experiments done, using this methodology, are presented. Those experiments include, the validation of OM method and the comparison with some other known methodologies. This chapter finishes with a summary of the results obtained and conclusions.

5.1 Introduction

Biological processes of all living organisms are not yet fully known. Some of those processes can be studied through PPI, because PPI are involved in many biological processes. Most of the proteins work in complexes and PPI analysis contributes to the understanding of the cellular organization, the cellular processes, and the cellular functions. Disease states may appear, if the physiological interaction between two proteins is disrupted [108]. PPI analysis can contribute to the identification of drug targets that will be used in the discovery and development of drugs to fight specific diseases. PPI can be modeled by networks, but these models are only approximations of the real PPI networks and one of the reasons of that is the lack of reliability and accuracy of the high-throughput experimental methods used for PPI identification. Besides that, these experimental methods take time, so the use of computational methods is a way of overcoming these constraints. Network-based methods are used to build models of these interactions.

PPI biological networks are a subset of complex biological networks that have specific topological properties, such as a high clustering coefficient, the presence of hierarchy, heterogeneity and a power-law-like degree distribution [104]. The guilt-by-association hypothesis states that two proteins sharing many interactive neighbors are likely to hold functional homogeneity and localization coherence [109]. These characteristics suggest that network topology alone may be a viable option for PPI network denoising.

Biological networks inference is the reconstruction of biological networks from high-throughput data, which can provide valuable information about biological mechanisms that contributes to biological and medical knowledge. A comparison between inference methods applied to gene regulatory networks can be found in [110].

Biological networks inference methods predict nodes and links, to find new biological entities and new relationships between them and use denoising methods that applied to real network models, allow to improve these models and approximate them to the real networks they represent.

Networks comparison can be used to predict nodes and edges of biological complex networks. Networks comparison may allow to discover new protein functions and disease-specific changes. Several network similarity measures can be applied for network comparison [46] and examples are: the edit distance, defined as the number of the edges that need to change to convert one network to another, the sampling distance, defined as the similarity to an ensemble of random networks, the cut distance and the similarity distance [111, 112], the spectral distances [113], and the comparison of the top-k nodes [114, 115]. Reviews about methodologies used to compare biological complex networks can be found in [7, 116]. Networks comparison can also be related with the network dynamics, by the analysis of sequential instances of network topology. Holme [117] discussed several network parameters that should be studied along temporal changes, like the connectivity, the diameter, the centrality, the motifs and the modules.

Computationally inferred interactions offer a useful resource in higher-level biological comprehension or in testing new interactions predicted from assumptions made by researchers. Link prediction attempts to estimate the likelihood of the existence of a link between two nodes, based on observed links and the attributes of nodes. In many biological networks, whether a link between two nodes exists must be demonstrated by field or laboratory experiments, which are usually very costly. Link prediction algorithms can be used, for example, to extract missing information, identify spurious interactions, and evaluate network evolving mechanisms, which can reduce the experimental costs if the predictions are sufficiently accurate [7].

Link prediction can be performed by comparing an appropriately selected network model, with a similar real world network, or with an ensemble of networks of the same type, or of multiple types, together with functional information [118–121]. Evolution models can also be built and used as link-predictors by analyzing sequential instances of these networks topology [46].

Link prediction algorithms are classified in three categories in [7]: similarity-based algorithms, maximum likelihood algorithms, and probabilistic models.

In the similarity-based algorithms category, for each pair of nodes x and y, a score S_{xy} is assigned, denoting a similarity measure (or index) between x and y. All non-observed links are ranked according to their scores, and the links connecting the nodes more similar are supposed to be of higher likelihoods. Similarity scores are classified into three categories: local indices, global indices and quasi-local indices. Local indices use local information and often use the edge neighborhood of the connected nodes, which may include all first neighbors, or all first and second neighbors. Global indices use the whole topological information and can provide much more accurate prediction than the local indices, but their calculation is very time consuming, usually infeasible for large-scale networks. Quasi-local indices do not require global topological information but make use of more information than local indices. Examples of indices belonging to these categories can also be found in [7]. Edge neighborhood may be compared by using the network degree, preferential attachment methods, fitness values, the community structure, a hierarchical structure model, a stochastic bloc model, probabilistic models, or by using hyper-graphs [46, 122].

Algorithms based on the maximum likelihood estimation presuppose some organizing principles of the network structure, with the detailed rules and specific parameters obtained by maximizing the likelihood of the observed structure. These algorithms are very time consuming and fail to deal with huge networks (millions of nodes). They are not probably among the most accurate ones, but provide very valuable insights into the network organization, which cannot be gained from the similarity-based algorithms or the probabilistic models.

Probabilistic models aim abstracting the underlying structure from the observed network and predict the missing links by using the learned model. Given a target network $G = (V, E)$, the probabilistic model will optimize a built target function to establish a model composed of a group of parameters Θ, which can best fit the observed data of the target network. Then the probability of the existence of a link (i, j) is estimated by the conditional probability $P(A_{ij} = 1 \mid \Theta)$ [7].

Different network-based methods were developed (reviewed in [123]), to identify false-positive interactions and missing links in biological networks. These methods use different strategies, such as using repeating experiments [124, 125], using prior knowledge about proteins [126, 127], using functional or structural annotations [128–132] and using comparisons with theoretical distributions constructed from known data and network topology-based approaches [133–137]. The herein proposed approach falls under the last category, using network topology-based approaches.

In a recent study, Lü *et al.* [138] proposed a "structural consistency" index and the structural perturbation method (SPM). On the one hand, the structural consistency index can reflect the inherent link predictability of a network without knowing its organization *a priori*, allowing to estimate the explicability of the organization of a network, and to supervise mechanistic changes during the evolution of the network. On the other hand, the SPM performs link prediction by removing a percentage of the edges in a network, thus perturbing the remaining network by that percentage. This is based on the strong correlation between independent network perturbations, which suggests that the missing links (*i.e.*, false negative (FN) interactions) can be identified by perturbing the networks with an additional set of known interactions (*i.e.*, true positive (TP) interactions).

Luo *et al.* [123] proposed the collaborative filtering method (CFM) to perform protein interactome mapping on sparse high-throughput screening (HTS)-PPI data, since the performance of network topology-based approach usually deteriorates when using sparse network data. This approach is based on the notion that interactome mapping and personalized recommendation have similar solution spaces. Each protein is represented as a feature vector that describes their interactions in the network. In addition, the feature vector is used to calculate the corresponding vector similarity that represents the interactions through functional similarity weight, creating an inter-neighbourhood similarity (I-Sim) for modeling PPI. Functional parameters for each protein in the dataset are obtained from GO, allowing the use of functional similarity metrics. Denoising of the input HTS-PPI data is performed via the integration of saturation-based strategies into the I-Sim, achieving a precise relationship model. Their method was applied to three different datasets and compared with three other algorithms (interaction generality [133], Czekanowski-Dice distance [134], and functional similarity weight [135]), showing better performance on large, sparse HTS-PPI datasets. Since they use GO annotations to characterize their proteins, this approach is likely to under-perform when considering less studied organisms.

A different strategy termed intrinsic geometry structure (IGS) was proposed by Yi Fang *et al* [139]. IGS is a geometry-based approach which uses heat diffusion in the PPI network to collect structural information about all paths connecting two given nodes, thus defining intrinsic relationships among them. They use a maximum likelihood-based algorithm to determine the optimal dissipation time, predicting the global structure of the PPI network from the local structure. After performing heat diffusion for the optimal dissipation time, the intrinsic geometric structure of the PPI network is revealed. One of the main advantages of the IGS method is its robustness against missing protein associations and sparse PPI data. Their method was tested with the *S. cerevisiae* network [140], a network of the bottle-nose dolphin community [141], and a network of known terrorist cells [142]. In addition, they compared the performance of IGS with two other methods, a multi-dimensional scaling-based (MDS) method [143] and the hierarchical random graph (HRG) method [144], showing that IGS performed slightly better than MDS and HRG for all datasets tested. In their analysis, they did not make a biological significance analysis in their work, establishing their work, only, using the area under the ROC curve (AUC) values.

In the following, the proposed new methodology, applied to denoising PPI networks, will be described. This methodology, the OM methodology, uses uniquely the network topology to find false interactions and predict absent interactions.

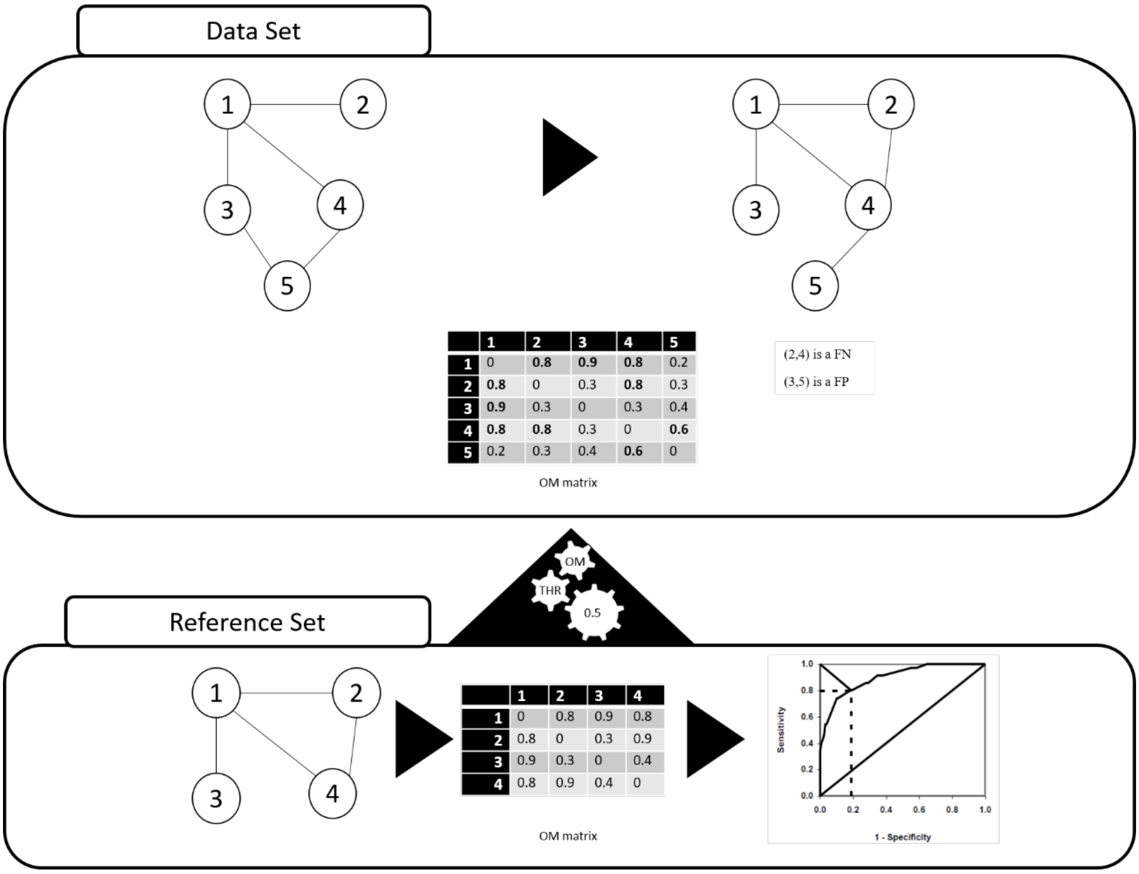
The OM methodology takes into considerations the limitations of the described methods and pretends to be a simple method, contributing to the denoise of PPI networks by using their inherent structure, to obtain a better model of the real network.

In this methodology, topological measures are used to find trends that characterize interacting and non-interacting proteins distributions. A high confidence set of protein interactions is used to construct a network, followed by the calculation of the weights of interactions and non-interactions in the network. The OM weighted matrix is obtained and used to find distribution trends that allow to distinguish interaction distributions from non-interaction distributions. The OM threshold value that better distinguishes these types of distributions is then used to identify false positive (FP) interactions and FN (novel) interactions. This way, an OM topological model is built to be used in the denoising of a network, resulting in a better approximation of the expected network.

5.2 Organization Measurement Methodology

OM methodology aims to solve two main problems: 1) The identification of FP PPI in a network; and 2) The prediction of new (FN) PPI, using exclusively the topology of PPI networks as input data.

This section will describe the OM methodology pipeline to denoise of networks (see Figure 5.1), and the topological measures used with the OM methodology, including the new NC proposed measure. When describing OM methodology, it will be described how to obtain the OM matrix of weights and how to determine the threshold value.



	1	2	3	4	5
1	0	**0.8**	**0.9**	**0.8**	0.2
2	**0.8**	0	0.3	**0.8**	0.3
3	**0.9**	0.3	0	0.3	0.4
4	**0.8**	**0.8**	0.3	0	**0.6**
5	0.2	0.3	0.4	**0.6**	0

	1	2	3	4
1	0	0.8	0.9	0.8
2	0.8	0	0.3	0.9
3	0.9	0.3	0	0.4
4	0.8	0.9	0.4	0

Figure 5.1: Diagram of the OM methodology pipeline. A Reference Set of an organism is used to create the model. The OM THR calculated using the Reference Set is applied to denoise a lower-confidence Data Set of the same organism.

For each organism, a set of high-confidence PPI interactions was collected. Although these PPI do not reflect the entirety of the protein interaction networks of the selected organisms, they are used to construct the known PPI network of each organism, named reference set.

In the application of the OM methodology, various topological measures were calculated to characterize those networks, based on the assumption that these measures will allow the identification of topological patterns to support network denoising. The term “denoising” is used to define the identification of FP and FN interactions, removing the former ones from the network while adding the latter. This methodology can also be used to rank the level of confidence of the interactions already presented in the network and also those missing. Different topological measures can identify different patterns and thus, here we consider that all topological measures can contribute to the denoising process.

5.2.1 Network Topological Measures

Protein interactions can be conveniently modeled as a network, where each node represents a protein and each edge represents an association between two proteins. The most commonly used technique to quantify the interaction profile similarity of a PPI network (or any other type of biological network) relies on association indices. Bass *et al.* [145] performed a comprehensive review on the selection of association indices for the analysis of gene similarity. In their work, the Jaccard (JC), Geometric and Cosine indices were shown to be the most versatile, as though not excelling in any particular task, their strengths were the most balanced out of all evaluated measures. A review of similarity indices can also be found in [146]. Daminelli *et al.* test the application of different association indexes to bipartite networks [147].

A more recent study reports that the JC measure performs better than three other measures in a specific model [148].

The JC measure is defined as the ratio of the intersection of the number of neighbors of nodes i and j divided by their union (i.e., the ratio of nodes shared between i and j divided by the total number of nodes connected to both):

$$\mathrm{JC_{ij}}{=}\frac{|\Gamma\,(\mathrm{i})\cap\Gamma\,(\mathrm{j})\,|}{|\Gamma\,(\mathrm{i})\cup\Gamma\,(\mathrm{j})\,|} \tag{5.1}$$

where Γ (i) is the set of neighbors of i. We also explored and tested additional measures, such as the betweenness (BETW) and Katz indexes.

The implementation of BETW used was:

$$\mathrm{BETW_{ij}}{=}(\mathrm{BETW_i}{+}\mathrm{BETW_j})/2 \tag{5.2}$$

where

$$\mathrm{BETW_i}{=}\sum_{\mathrm{l,m}\in\mathrm{V}}\frac{\mathrm{nsp(l,m|i)}}{\mathrm{nsp(l,m)}} \tag{5.3}$$

with V the set of nodes, nsp(l, m) is the number of (l, m) shortest paths and nsp(l, m|i) is the number of those paths passing through the node i.

The implementation of Katz used was:

$$\mathrm{KATZ_{ij}}{=}\frac{(\mathrm{KATZ_i}{+}\mathrm{KATZ_j})}{2} \tag{5.4}$$

$$\mathrm{KATZ_i}{=}\alpha\sum_{\mathrm{l}}\mathrm{Adj_{il}x_l}{+}\beta \tag{5.5}$$

with Adj the adjacency matrix of the network with eigenvalues λ. It was used $\alpha{=}1/\lambda_{\mathrm{max}}$ and $\beta{=}\ 0$, when Katz centrality is the same as the eigenvector centrality.

Based on the idea that closely associated proteins are more likely to interact, that the network modularity is associated with the clustering coefficient (CC) [149] and a high mean CC of a community can be used to identify those that are functionally homogeneous [150], we implemented a novel measure to emphasize the relevance of the CC concept associated to the neighborhood concept in a network. This measure was called NC measure and is defined as the ratio of the sum of the CC of the nodes shared between i and j divided by the sum of the CC of the total number of nodes connected to both i and j:

$$\mathrm{NC_{ij}}=\frac{\sum \mathrm{CC}\ (\Gamma\,(\mathrm{i})\cap\Gamma\,(\mathrm{j})\,)}{\sum \mathrm{CC}\ (\Gamma\,(\mathrm{i})\cup\Gamma\,(\mathrm{j})\,)} \tag{5.6}$$

where Γ (i) is the set of neighbors of *i*.

5.2.2 Organization Measurement Threshold Value Determination

One of the assumptions made in this work is that the PPI in the reference datasets are true. This assumption can be made due to the sparsity of protein interaction networks and the rigorous criteria chosen to filter TP interactions. However, the same cannot be said for the non-interactions, as the presence of FN PPI is highly likely.

The value that best distinguishes both interactions and non-interactions distributions was called the OM threshold value. First, we collect protein interactions data of a specific organism and then a network is built (Figure 5.1). Then, the respective adjacency matrix is constructed, followed by its transformation into a weighted matrix, the OM matrix, using the topological measures of interest. Finally, the receiver operating characteristic (ROC) curve is calculated and used to determine the optimal cut-off, corresponding to the threshold value that separates the interaction distributions from non-interaction distributions. We considered as the optimal cut the point closest to (0,1) in the ROC curve, where sensitivity equals specificity. Different topological measures were tested and the respective cut-off values were determined. The outcomes of these experiments will be described in Section 5.3.4.

5.2.3 Organization Measurement Methodology Application

An accepted assumption in network topology-based approaches is that interacting proteins in a local community and closer to one another in the network are most likely involved in similar functions, or part of the same pathways [151–153]. The use of topological measures that capture this information should be prioritized, as they are expected to better grasp patterns in incomplete networks, thus allowing the approximation of incomplete input networks to the real networks.

In brief, we used the PPI data of high confidence. From each of the resulting PPI networks we calculated their adjacency matrix. Then, after calculating the respective weights, the adjacency matrix is transformed into a weight matrix. Finally, the threshold that best separates PPI and non-PPI was determined through the finding of the optimal cut-off of the ROC curve. This threshold was applied to detect spurious and missing PPI in the network, to obtain a better approximation of the true network. In the example network shown in Figure 5.1, there are five nodes representing five different proteins, in addition to six edges that could represent the interactions between them (Data Set). Assuming the example network approximates the current knowledge of a given biological network, not all true interactions are represented and the existence of FP is expected. Once the threshold value is calculated, using the reference set (Reference Set), it is applied to the OM matrix, to identify FP and FN interactions. FP interactions are then removed from the network, whereas FN interactions are added.

5.2.4 Organization Measurement Matrix

In Figure 5.1 it is summarized the pipeline of the proposed OM methodology. Once the networks for the organisms are constructed, their respective adjacency matrices are built, followed by their transformation into a weighed matrix, the OM matrix. The OM matrix is used to find distribution trends that allow to distinguish interactions and non-interactions. The weights for interactions and non-interactions are calculated using topological measures and using the information about the interactions of the networks.

The adjacency matrix of the PPI network A, with N proteins and M interactions is defined as $\mathrm{adj_A}= [\mathrm{a(i,j)}]$, where $\mathrm{a(i,j)} = 1$, if there is an interaction in A between nodes i and j. Otherwise, $\mathrm{a(i,j)} = 0$. A topological measure is applied to A to determine a weight for each $\mathrm{(i,j)}$ to transform the adjacency matrix A into a transformed matrix $\mathrm{A_w}= [\mathrm{a_w(i,j)}]$, where $\mathrm{a_w(i,j)}$ is the weight of $\mathrm{(i,j)}$ in A, calculated using the topological properties of the network.

The weight $\mathrm{a_w(i,j)}$ represents the strength value of the edge $\mathrm{(i,j)}$ per the topological measure used and aims to capture patterns associated to the network that can originate signatures that identify the PPI network of each organism. This weight was used to characterize interaction and non-interaction distributions of the PPI network to determine the separation border between them.

5.3 Denoising Yeast and Human Protein-protein Interactions Using the Topology

The OM methodology was tested with different topological measures and was evaluated using different scenarios. Several experiments were done, namely:

- The analysis of different topological measures to identify the optimal threshold value, since the threshold value will be used to discriminate between protein interactions and non-protein interactions distributions;

- The evaluation of the OM methodology in different scenarios to assess whether the OM methodology is sensitive to the network topology, to assess its performance and to compare it with other network-based methodologies proposed by other researchers, that is:

  - Random network vs. reference set network comparison;
  - Random insertion of edges;
  - Random deletion of edges.

5.3.1 Data Sets

The STRING database [154] contains known and predicted protein interactions of various organisms. PPI in STRING derive from five main sources: 1) Genomic context predictions; 2) High-throughput experimental methods; 3) Conserved co-expression experiments; 4) Automated text mining, and; 5) Previous knowledge from third-party databases. Each interaction in STRING is associated with a combined score (CS) that ranges from 0 to 1000, indicating the degree of confidence of specific interactions. Calculation of the CS considers several parameters, such as the number and quality of different sources indicating that a PPI occurs.

The interactions derived by experimental methods with a score greater than 900 have been considered of high-confidence in multiple works [155, 156]. Therefore, the reference sets used in this work comprise experimentally determined PPI data obtained from STRING with a score greater than or equal to 900.

These data were collected from two different organisms, namely the *Yeast Saccharomyces cerevisiae* (Yeast) and *Homo sapiens* (Human). Using these data, an undirected network is constructed for each organism and the main component is extracted. Table 5.1 summarizes the characteristics of the reference set networks obtained for Yeast and Human, including the number of nodes, the number of edges, the average degree and network density. The observed average degree and density values are highly suggestive that these biological networks are sparse, *i.e.*, they have much less edges than the full network with the same set of nodes. Our high-confidence networks (*i.e.*, PPI obtained from the STRING database with experimental source score greater than 900) comprised 29,319 interactions between 3,937 proteins for the Yeast dataset and 16,931 interactions between 4,943 proteins for the Human dataset.

Additionally, it was used a high confidence external dataset compiled by Collins *et al.* [140] and referred to as CS2007 hereafter, to compare the proposed methodology with other topology-based denoising methods [139, 143, 144]. This dataset comprises 9,074 PPI between 1,622 unique proteins from *S. cerevisiae.* To ensure a direct comparison between the OM methodology and the existing methods we followed their approaches and only used the largest connected component. The largest connected component of the dataset compiled by Collins *et al.* [140] includes 8,323 interactions between 1,004 proteins (see Table 5.1).

Table 5.1: Topological characteristics of the Yeast, Human and CS2007 networks used as reference sets.

Organism	N^o of nodes	N^o of edges	Average degree	Density
Yeast	3,937	29,319	14.8941	0.0038
Human	4,943	16,931	6.8505	0.0014
CS2007	1,622	8,323	10.2626	0.0063

5.3.2 Identify the Optimal Threshold Value

A key component of the proposed method is the determination of the threshold value to discriminate between protein interactions and non-protein interactions distributions. As such, it was decided to test the OM methodology with different topological measures to determine which better discriminates PPI from non-PPI. The four topological measures used were the JC, the BETW, the KATZ and the proposed new measure, the NC (see section 5.2.1). They were calculated also by normalizing them between 0 and 1.

In addition to testing the OM methodology with these measures, the cut-off values for both the optimal cut and the accuracy cut were calculated, using the JC and NC measures, those that gave better results. The optimal cut calculates the point closest to (0,1) in the ROC curve, where sensitivity equals specificity, whereas the accuracy cut calculates the maximum accuracy and the respective cut-off value. Results will be shown in section 5.3.4.

5.3.3 Evaluation of the Organization Measurement Methodology in Different Scenarios

To assess whether the OM methodology is sensitive to the network topology, OM methodology was applied to a randomly generated protein network, with the same number of nodes and edges as their respective reference sets (for Yeast and Human). If the OM methodology can distinguish between interactions and non-interactions in the reference data sets, but fails to do so in the random networks, one can assume that it captures the inherent topological structure of a real network.

To further evaluate the performance of OM methodology, two other experiments were performed. First, while maintaining the same number of nodes (proteins), we randomly added incrementing percentages of edges (proteins interactions), 20%, 40%, 60% and 80%, not belonging to the reference set network, building four networks and removed the same percentages of edges from the reference set network, building more four networks. This was performed for the Yeast, Human and CS2007 reference set networks. After each addition or removal, we used the OM methodology to denoise the networks. More, to assess the ability of the proposed methodology for network denoising, it was determined the percentage of inserted true negative (TN) removed from the respective CS2007 perturbed networks and the percentage of TP retrieved from the respective CS2007 perturbed networks. OM was compared to MDS and IGS methodologies [139]. Next, there is a thorough description of these experiments.

OM Methodology Performance Comparison: Random Network vs. Reference Set Network

The only criteria selected to generate the random networks was that the resulting randomize networks were required to comprise the same number of nodes and edges. Thus, 10 networks were generated for each organism to be tested using the NC measure.

Random Insertion of Edges

To evaluate the performance of the OM methodology for denoising PPI networks, the networks of the reference sets were perturbed by randomly adding incrementing percentages of edges to the networks of the Yeast and Human reference sets and the CS2007 reference set.

Four noisy networks were created for each data set, adding 20% more edges to the original network, followed by 40%, 60% and 80%. These intervals were selected following the research conducted by Yi Fang *et al.* [139].

To be able to compare the performance of OM methodology with other network-based methodologies proposed by other researchers, the CS2007 network was also perturbed by randomly inserting edges in the same proportions previously described.

After denoising the networks with the OM methodology, the percentage of FP interactions that were removed was also calculated.

Random Deletion of Edges

To evaluate the performance of the OM methodology for the identification of missing interactions, four new networks were created for each dataset (*i.e.*, Yeast, Human, and CS2007 reference sets) by removing increasing percentages of edges from the respective reference set networks. Edge removal was performed in the same proportion as edge addition: 20%, 40%, 60% and 80%.

5.3.4 Results and Conclusion

The results will be presented following the same organization for a better understanding.

Identify the Optimal Threshold Value

The ROC curves obtained after using this methodology with four different topological measures for the Yeast and Human organisms, are shown in Figure 5.2.

Figure 5.3 shows the same information, but after data normalization between 0 and 1. Table 5.2 shows the respective AUC values obtained. It can be observed that the best results were achieved when using OM methodology with the JC and NC measures.

The cut-off values for both the optimal cut and the accuracy cut were calculated, using the JC and NC measures, those that gave better results. The optimal cut calculates the point closest to (0,1) in the ROC curve, where sensitivity equals specificity, whereas the accuracy cut calculates the maximum accuracy and the respective cut-off value. Table 5.3 and Table 5.4 show the AUC, optimal cut and accuracy cut values in Yeast and Human datasets respectively, using the JC and NC measures.

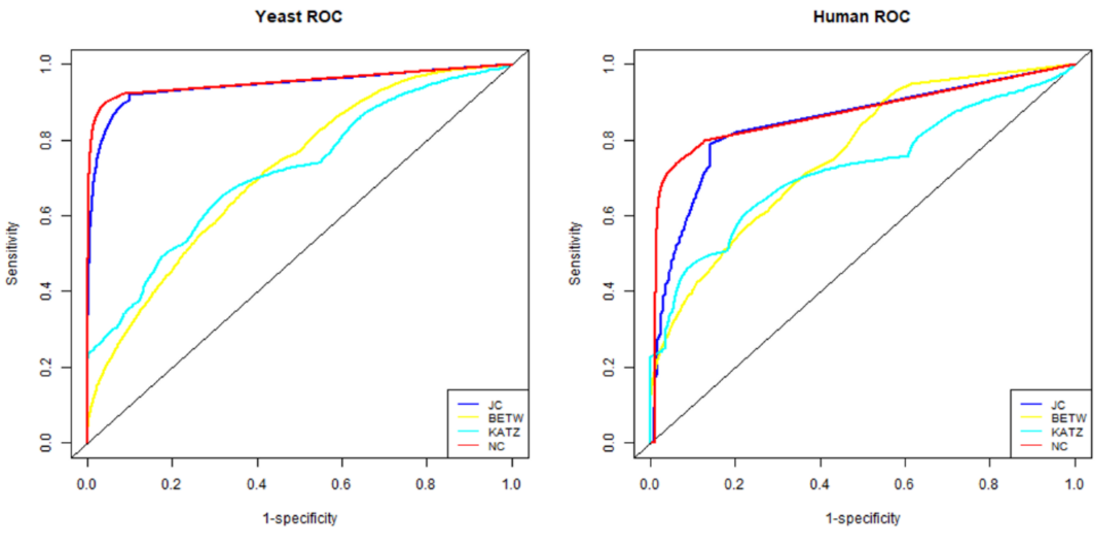


Figure 5.2: OM methodology ROC curves. ROC curves obtained by OM application with JC, BETW, KATZ and NC measures in Yeast and Human datasets.

Table 5.2: AUC in Yeast and Human datasets.

Topological Measures	Yeast AUC		Human AUC	
	Not normalized	Normalized	Not normalized	Normalized
JC	0.9462	0.9460	0.8438	0.8458
BETW	0.7142	0.7676	0.7659	0.8004
KATZ	0.7151	0.7714	0.7258	0.7944
NC	0.9534	0.9526	0.8708	0.8700

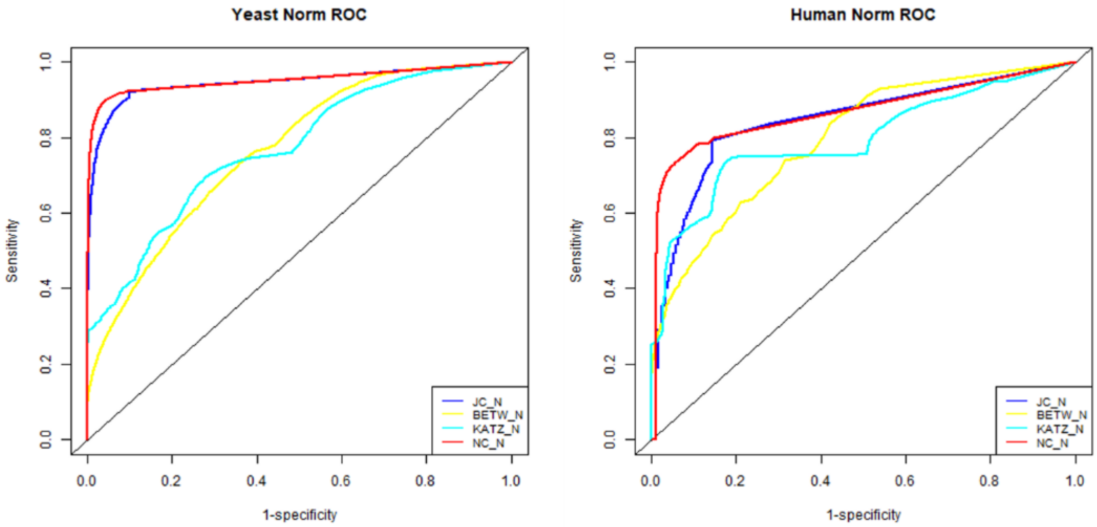


Figure 5.3: OM methodology ROC curves with normalized data. ROC curves obtained by OM application with JC, BETW, KATZ and NC measures in Yeast and Human datasets after data normalization between 0 and 1.

Table 5.3: AUC, optimal cut and accuracy cut values in Yeast.

Yeast	AUC	Optimal cut		Accuracy cut	
JC	0.9462	sensitivity specificity cut-off	0.9246 0.9000 0.0008	accuracy cut-off	0.9123 0.0008
NC	0.9534	sensitivity specificity cut-off	0.9057 0.9471 0.0021	accuracy cut-off	0.9282 0.0044

Table 5.4: AUC, optimal cut and accuracy cut values in Human.

Human	AUC	Optimal cut		Accuracy cut	
JC	0.8438	sensitivity specificity cut-off	0.8069 0.8531 0.0005	accuracy cut-off	0.8300 0.0005
NC	0.8708	sensitivity specificity cut-off	0.7989 0.8735 0.0001	accuracy cut-off	0.8400 0.0014

Evaluation of the OM Methodology in Different Scenarios

- Random network vs. reference set network comparison;
- Random insertion of edges;
- Random deletion of edges.

OM Methodology Performance Comparison: Random Network vs. Reference Set Network. Ten networks were generated for each organism to be tested using the NC measure. Figure 5.4 shows the ROC curves, the separation of classes (PPI e non-PPI) curves and the accuracy curve, when applying the OM methodology with the NC topological measure to one of the Yeast (left column) and Human (right column) random networks, generated with the same number of nodes and edges of the respective reference sets. Analyzing their ROC curves, we can see a clear distinction in performance between the application of OM methodology to the random network (see Figure 5.4) and the subsets of the real networks (see Figure 5.2).

The AUC obtained after using the proposed method in all 10 random networks generated was close to 0.5 for both organisms (Yeast and Human), while for the subset of Yeast network and Human networks the AUC was 0.9534 and 0.8708, respectively.

Random Insertion of Edges. Four noisy networks were created for each data set, adding 20% more edges to the original network, followed by 40%, 60% and 80%. Figure 5.5 shows the ROC curves when the OM methodology is applied to the networks of the Yeast and Human reference sets and to the four noisy networks generated from each of them. It can be observed a decreasing of performance when we increase the percentage of the random edges added.

The OM methodology was applied to the four noisy networks obtained from the CS2007 network and compared to the results obtained when MDS and IGS methodologies were applied [139]. The graphical representation of the resulting AUC values are shown in the graphical representation of the Figure 5.6. It can be observed that the proposed OM methodology outperforms MDS and IGS methodologies.

After calculated the percentage of FP interactions that were removed, it could be observed that the OM methodology could remove 97% of the FP of the 20% added and 89% of the 80% added (see Table 5.5).

Random Deletion of Edges. Edge removal was performed in the same proportion as edge addition: 20%, 40%, 60% and 80%. The results are shown in Figure 5.7, representing the ROC curves, when the OM methodology is applied to the eight noisy networks referred previously, of the Yeast and Human reference set networks.

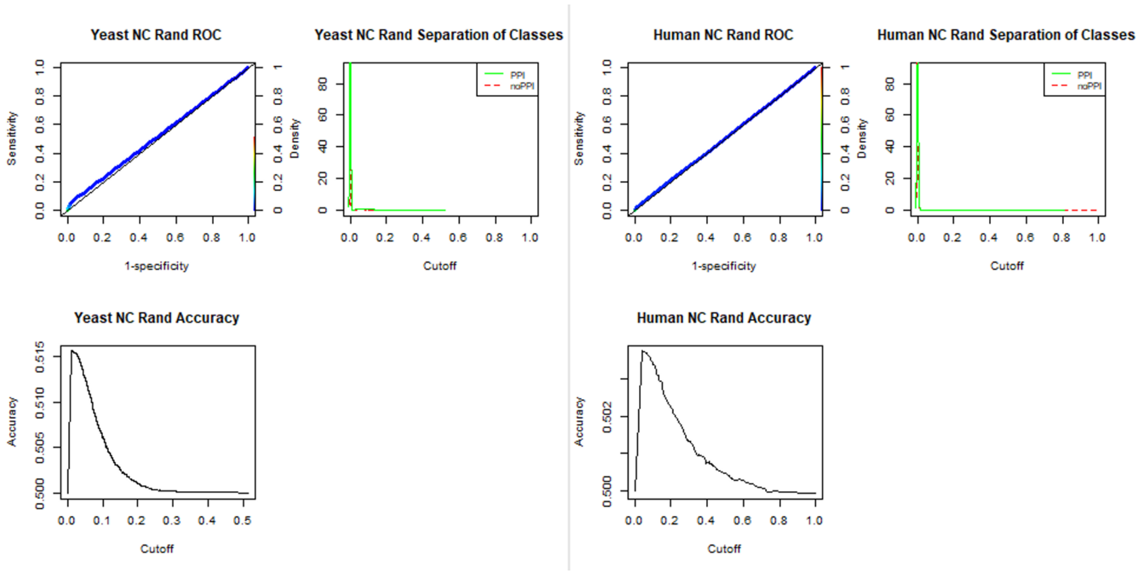


Figure 5.4: OM methodology application with the NC topological measure, to Yeast and Human random networks. OM methodology application ROC curves, separation of classes (PPI and non-PPI) curves and accuracy curve with the NC topological measure in one of the random networks generated with the same number of nodes and edges as the Yeast (left column) and as the Human (right column) reference sets.

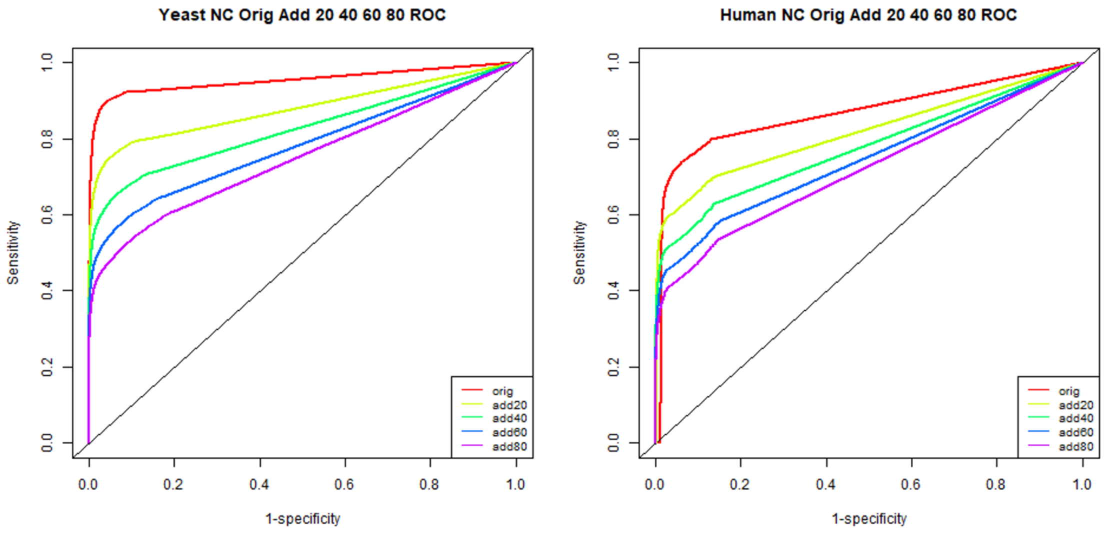


Figure 5.5: Application of the OM methodology with the NC topological measure, when an increasing percentage of edges was added randomly to the Yeast and Human reference set networks. ROC curves of the reference sets, and the other 4 networks, when 20%, 40%, 60% and 80% of edges were added to the reference set network for Yeast (left column) and Human (right column).

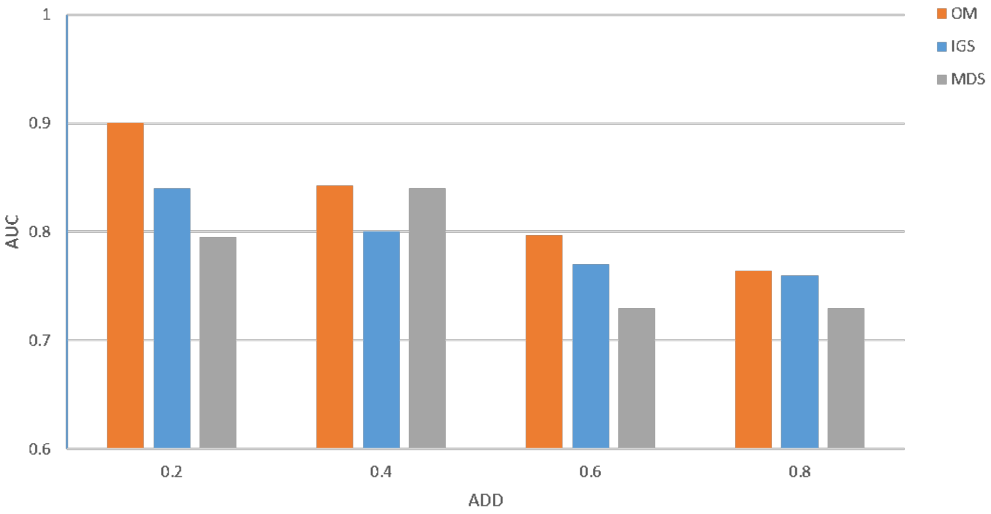


Figure 5.6: Application of OM methodology with the NC topological measure, when an increasing percentage of edges was added randomly to the CS2007 reference set network compared to the MDS and IGS methods. AUC values of the CS2007 perturbed 4 networks when a percentage of random 20%, 40%, 60% and 80% of edges were added to the reference set networks, using OM, IGS and MDS methods.

Table 5.5: Percentage of FP removed after applying OM methodology to the noisy networks of the CS2007 dataset.

% added	# FP added	# FP removed	% FP removed
20	1,665	1,607	97
40	3,329	3,114	94
60	4,994	4,560	91
80	6,658	5,926	89

These results show a scenario alike the one observed after randomly adding edges, as greater reductions in the number of edges result in greater performance drops, but the performance drops are steeper in the Human organism.

Figure 5.8 shows a graphic of the AUCs values, when the OM methodology is applied to the four noisy networks obtained from the CS2007 network, when edge removal was performed in the proportion, compared to the MDS and IGS methodologies [139].

OM methodology has a better performance compared to the IGS and MDS methodologies, except when 80% of the interactions are removed from the CS2007 reference set, where the application of IGS gives better results (see Figure 5.8).

Further details are in Table 5.6, where we can observe that 95% of the TP removed could be detected when the OM methodology is applied to the perturbed network, when 20% of the interactions reference set were removed and 40% could be detected when 80% were removed.

Analysis

Different topological measures were used to identify the optimal threshold, with the Yeast and Human reference sets and comparative testing showed (see Figure 5.2, Figure 5.3 and Table 5.2) that the best results were obtained using the JC and the NC measures, and thus it was decided to use both in some experiments of this work. JC is a widely known measure frequently used in network denoising and missing link prediction. It also considers the neighborhood information, which is aligned with the “guilt-by-association” principle. Same applies to the NC index, proposed herein, where the concept of CC is also taken into account.

The OM methodology was then applied to the Yeast and Human datasets, using the JC and the NC measures and after analyzing Table 5.3 and Table 5.4, where the AUC values and the cut-off values, for both the optimal cut and the accuracy cut for the Yeast and Human reference sets, obtained are shown, it can be seen that the NC measure performed better than the JC measure at discriminating between protein interactions and non-interactions and for this reason the NC measure was used in the evaluation of the OM methodology.

Three different scenarios were considered to evaluate the OM methodology. The first one uses randomly generated protein networks, with the same number of nodes and edges as their respective reference sets (Yeast and Human). Observing the Figure 5-4, it can be seen that the AUC, obtained when applying the OM methodology to one of the random networks, was close to 0.5 for both organisms (Yeast and Human), while for the Yeast and Human reference sets, the AUC was 0.9534 and 0.8708, respectively, which shows that OM is sensitive to the inherent topological structure of a real network. These results show that the OM methodology cannot distinguish between interactions and non-interactions in random networks, but can capture the inherent rules of biological networks, not present in random networks.

The second scenario used to evaluate the performance of OM methodology, consisted in applying OM to networks obtained from the two Yeast and Human reference sets, where

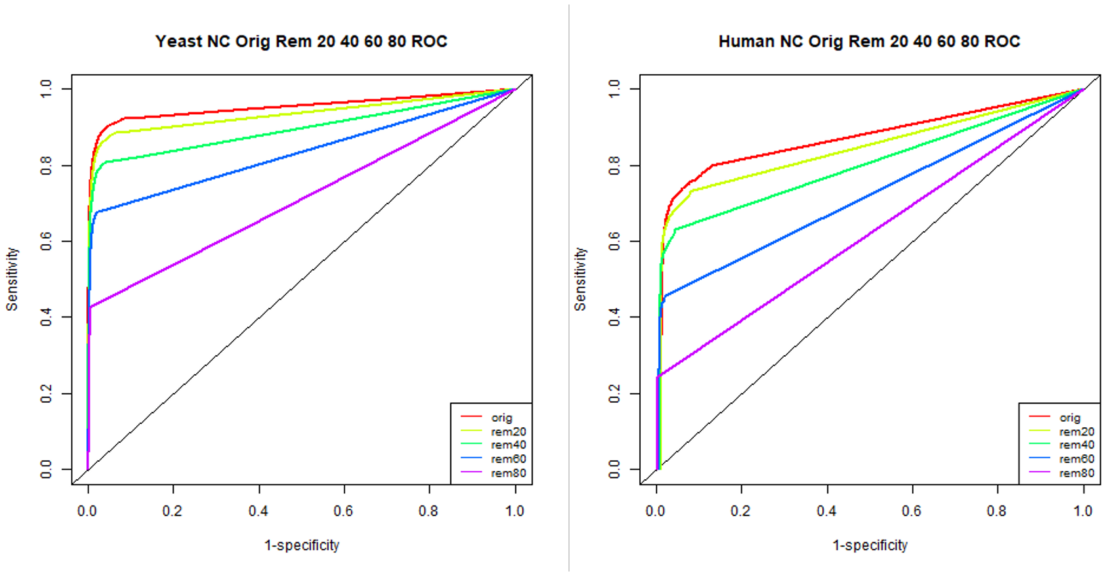


Figure 5.7: Application of the OM methodology with the NC topological measure, when an increasing percentage of edges was removed randomly to the CS2007 reference set network. AUC values of the CS2007 perturbed 4 networks when 20%, 40%, 60% and 80% of edges were removed to the reference set networks, using OM, IGS and MDS methods.

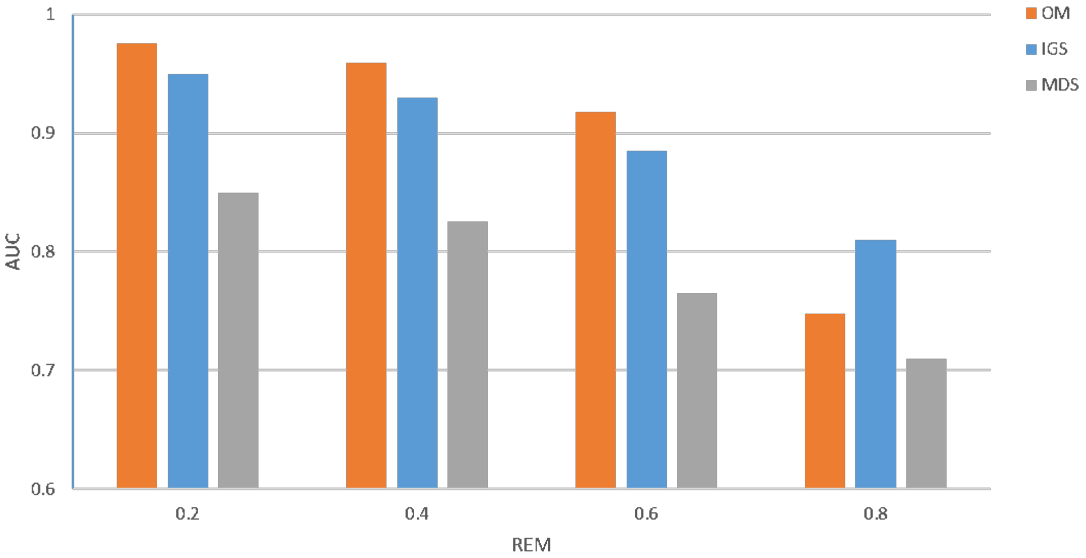


Figure 5.8: Application of OM methodology with the NC topological measure, when an increasing percentage of edges was removed randomly to the CS2007 reference set network compared to the MDS and IGS methods. AUC values of the CS2007 perturbed 4 networks when a percentage of random 20%, 40%, 60% and 80% of edges were removed to the reference set networks, using OM, IGS and MDS methods.

Table 5.6: Percentage of TP inserted after applying the OM methodology to the incomplete networks of the CS2007 dataset.

% removed	# TP removed	# TP inserted	% TP inserted
20	1,665	1578	95
40	3,329	2993	90
60	4,994	4011	80
80	6,658	2633	40

the number of nodes (proteins) was maintained, but where a random percentage of edges (proteins interactions), 20%, 40%, 60% and 80%, not belonging to the reference set network, were added, and the third scenario is similar to the second but instead of adding, the same random percentages of edges were removed from the reference set network.

In the second scenario (random insertion of edges), as expected, greater increments of random edges resulted in greater performance reductions (Figure 5.5). The performance reductions were steeper in Human, which could be attributed to one major reason: the percentage of FN is most likely greater in the Human interactome than in the Yeast interactome. Thus, it could be argued that the Yeast reference set is a more reliable, better representation of the actual Yeast interactome, than the Human reference set is of the real Human interactome. When these percentages of random edges were added, the inherent structure of these biological networks becomes deteriorated, because TP were probably added.

In the third scenario (random deletion of edges), greater reductions in the number of edges result in greater performance drops compared to the second scenario, but the performance drops are steeper in the Human organism (Figure 5.7). This could be explained by the fact that we are removing TP from both networks. However, since the Yeast network seems to be a closer representation of its true network than the Human network, the accentuated deterioration in the structure of the Human network could explain this behavior.

So, when comparing the results between edge addition and edge removal in Yeast and Human reference sets (Figure 5.5 and Figure 5.7), it can be witnessed that the overall performance reductions were quite dissimilar. Adding just 20% more edges contributed to a reduction of approximately 0.08 in AUC for Yeast, and 0.06 AUC for Human. Further addition of edges beyond this point did not decrease the AUC as sharply. Contrariwise, after removing 20% of the existing edges, the AUC decreased by roughly 0.02 for both Yeast and Human, with greater performance drops after each percentage of edges removal.

The better performance observed for the Yeast interactome could be explained by its smaller size compared to the Human interactome, in addition to being relatively well-studied, meaning that input data quality plays an important role in the performance of computational methods. Additionally, the negative impact in performance observed after randomly adding edges suggests that the OM methodology is very sensitive to high percentages of FP and FN.

To compare the performance of OM methodology with other network-based methodologies proposed by other researchers, CS2007 network reference set was perturbed by randomly inserting edges in the same proportions previously described in scenario two and by randomly deleting edges in the same proportions previously described in scenario three. OM methodology was compared to the MDS and IGS methodologies [139]. Figure 5.6 and Figure 5.8 show the AUC values when these methodologies were applied to this dataset and it can be seen a general improvement in the performance when the OM methodology is applied compared to the MDS and IGS methodologies.

Further analysis was conducted for these last networks with added random percentages of FP interactions. OM methodology was applied and the percentage of FP interactions removed was calculated (Table 5.5). Interestingly, most of the randomly inserted FP interactions were promptly identified, even when the network was heavily perturbed, with 89% of the FP removed after contaminating the network with 6,658 random interactions. These results suggest that the OM methodology can indeed capture the inherent topology of biological networks. Interestingly, it was observed that the number of TP identified after randomly removing edges from the CS2007 dataset plummets after removing 60% of TP (Table 5.6). Still, the OM methodology seems to identify most missing links up to that point.

These findings suggest that the OM methodology can assess whether the topological structure of a network is according to the characteristic topology of biological networks. OM methodology could still work well in less-studied interactomes, when the subset of the interactome of interest is a representative sample of the structure of the entire interactome, meaning that the percentage of FP and FN cannot hide the inherent structure behind the biological networks of the organisms.

Currently, low-throughput experimental methods are the only effective way to validate protein interactions. While high-throughput experimental methods to obtain PPI exist, the obtained results have very high noise. As such, computational methods are required to speed data acquisition and to reduce the data contamination. Methods relying exclusively in the topology of biological networks are simpler and faster, as it appears that networks topology may reveal patterns or signatures associated with the kind of organism and the type of interactions. If we can use, effectively, only the topology to denoise biological networks, we have a simple computational method suitable for incomplete interactomes, without the need of extra biological knowledge.

This research introduced the OM methodology for denoising biological networks, a methodology that: a) uses exclusively the topology of the network; b) enables, easily, to distinguish the distributions of interaction and non-interaction proteins in PPI networks; c) does not use known distributions as approximations; and d) provides a topological way of detecting FP interactions and find new interactions. The main innovation of the OM methodology is related with its ability to combine the advantages of using exclusively the topology without taking approximations to known distributions and without using external knowledge to detect interactions that do not exist or to find new interactions with a better performance than some documented used methodologies. This research also introduced a new network topological measure, the NC, which was used with the OM methodology and yielded better results, compared to other known and current topological measures.

So, the OM methodology is sensitive to the topological structure of the biological networks and can be used for network denoising. The obtained results suggest that the present approach can efficiently be used to denoise PPI networks. The OM methodology can be explored in the future by applying it in networks belonging to other domains, where there is an inherent structure, to predict new interactions and eliminate spurious interactions.

5.4 Summary

As the biological processes of organisms are not yet fully understood, and as the methods used to determine the interactions between their bio-entities are not accurate, in constructing network models of these interactions, these models will contain non-existent interactions and will not have interactions that should exist.

In this chapter a new method for the denoise of PPI networks has been described, the OM method, which is a method based exclusively on the topology of the networks. It was also proposed a new topological measure NC. This method was tested in two organisms and it was found that when compared to other network-based methods, in the great majority of the cases tested, it presents superior results.

Chapter 6

Prediction of Diseases on Dynamic Biological Networks using Supervised Methods

Diagnostic, prevention and cure of diseases has a growing social and economic impact. Diseases have been studied for some time, but there is still a lot of unknown knowledge behind their mechanisms and biological processes. Cancer is one of those diseases, which is associated to genetics and environment factors. Cancer diseases are very heterogeneous and have several stadiums that can be characterized according to various perspectives.

Diseases, like cancer, must be studied in a systemic way, since several bio-entities contribute to its appearance and evolution. To model the interaction between the bio-entities, network models can be used. Diseases alter the normal functioning of the organism, changing the interactions between their bio-entities. The interactions between bio-entities, for example, in PPI, interactions do not occur all at the same time, because some proteins may be inactive and their coded genes being expressed at different instants, or according to different conditions or individuals.

To characterize and differentiate those networks, global or local topological properties measures can be used. It is necessary to find those properties that better describe the networks according to the biological question posed and that help to find patterns associated to relevant biological processes and consequently could help in the identification of signatures associated with embedded disease processes. Those processes are dynamic, so is necessary to capture their dynamics that can vary according to various conditions, evolution through time, different states, specific traits, and from individual to individual. Because not all genes are expressed at the same time across different conditions, using information from genes expression to create PPI subnetworks allows to have several snapshots, each a PPI subnetwork, that, in a whole, are representative of the dynamics of the process in study.

The fact that there exists already a subset of biological data with known and curated disease classification allows the use of supervised learning techniques from machine learning, where samples of data are used to train a model that afterwards can be applied to other datasets with unknown classification to preview the existence of the studied disease.

6.1 Introduction

Anomalies in a gene, protein or other bio-entities can cause diseases, and since the arrival of the NGS, more evidences of human genes being correlated to diseases were found. Data from November 22^{th}, 2016 obtained in the on-line mendelian inheritance in man database (OMIM) [157] shows that there are 5861 phenotypes for which the molecular basis is known and 3642 genes with phenotype-causing mutation and data from October 15^{th} 2018 increased to 6275 and 3973 respectively .

Genomic changes that are translated to proteins can alter biological functions and a system-based approach modeled through complex networks can assist the discovery of signatures related to disease mechanisms, through the analyses of their topology [52, 82, 158, 159].

Most disease genes are not essential, and essential genes are associated to the hubs of the biological networks [158, 160], and since these networks are approximately scale-free, they are robust to random failures [16] and have compensatory mechanisms, when there are function failures. These compensatory mechanisms are one of the reasons why diseases resist to some drugs. Also, it is known that genes and proteins that are involved in the same phenotype are network neighbors [161], and that a disease phenotype can be associated to interactions in a biological complex network, that models these biological processes [162].

Several data mining methods have been applied to explore biological data and understand the mechanisms that regulate genetic and metabolic diseases, like data mining classification techniques, which are supervised learning methods that have been used to look for signatures in cancer diseases. The underlying hypothesis is that the identification of signatures can help the clinical identification of diseased tissues. To use supervised methods, it is necessary to choose the best features to use, to build the classification model, and also test different supervised learning algorithms to find those that best generalizes the classification.

A common trend consists in combining the information obtained from gene expression and protein-protein interactions to build series of complex networks to model system dynamics. Many different methodologies have been tested using unsupervised methods and supervised methods.

If topological properties measures could be used as features to predict diseases, without any prior knowledge, those found more relevant could be associated to topological signatures that could be used in the prevision and posteriorly in the identification of genetic targets that can be tested in the prevention or cure of those diseases.

Other researchers used topological properties of biological networks to study existent disease mechanisms. Global and local topological measures, were used in [23] to show that the structure of yeast PPI networks is closer to the geometric random graph model relatively to graphlet frequency and in [24], a new network similarity measure is defined based on the graphlet degree distribution as a generalization of the degree distribution (see Section 2.3 for definition). Cliques also helped to understand the mechanisms involved in cancer, since they are fully connected subnetworks more conserved in biological networks. In cliques, genes are functionally related and highly expressed. In [163] it was proposed a topological and biological feature-based network approach, integrating the expression data, along with network topological information and biological information. Cliques were scored-based on this information and were considered as gene signatures for the colorectal cancer (CRC). DNA, RNA and protein changes data were integrated to understand breast cancer metastasis process in [164]. In [165] was described how to build a PPI network representative of the CRC, where nodes are genes/proteins obtained from gene set enrichment analysis (GSEA).

This chapter is based on the paper [166] and describes a methodology that uses exclusively the topological properties of PPI networks induced by expressed genes. Those properties were used as features, to capture the dynamics embedded in different samples, to preview the existence of a certain disease and using supervised learning methods. The emphasis is on cancer diseases. This chapter begins with an introduction to feature selection algorithms and their use with cancer diseases data, modeled by networks. Then, is given a summary of some supervised learning algorithms and, after a contextualization, is described the proposed new methodology called sample series networks (SSN), which was applied in the prediction of cancer. This methodology describes how to capture the dynamics in the subset of the samples for each group of classification, usually different individuals or different stages of the same individual, which is an innovative aspect of capturing dynamics, assuming that the dynamics is not only present through different states or through time, but also through different individuals or different stages of the same individual. This methodology pretends to capture new patterns that could complement patterns found with other dynamic conditions. Different lays of dynamics can contribute to the improvement of the knowledge of the system in their steady and dynamic states.

6.2 Features Selection

Features selection (FS), to be used in supervised classification, consists of identifying the minimum subset of variables (features) that allows the prediction, while assuring high performance. Feature selection contributes to faster and more cost-effective models of the biological processes that generated the studied data.

Typically, gene expression studies (DNA micro-arrays or RNA-Seq) have few samples compared to the high number of genes tested for each sample. Even big studies have a couple hundred samples for about 20, 000 genes. Despite that, it is assumed that only a small portion of those genes have a direct impact to the experimental condition under study. FS is used to reduce the dimensionality of a dataset by selecting an optimum subset of features and using only this subset of features in the analysis. The optimum subset is much smaller than the entire feature set, and consequently, the computation time of the analysis is greatly reduced, and the learners being built using the reduced subset perform on a similar level to those built using the entire feature set, and in some cases perform better, as it is shown in [167], when filter and wrapper gene selection procedures are applied.

Several FS approaches have been used when classifying cancer data sets. A classification of lung cancer tumors as small cell lung cancer (SCLC), non-small cell lung cancer (NSCLC) and COMMON classes, using the structural and physiochemical properties of protein sequences obtained from genes using micro-array analysis can be seen in [168]. Several FS methods and prediction techniques were used. Best results were obtained using Bayesian network learning (BNL) with gain ratio.

A model for predicting the survival rate of patients affected by lung cancer, can be found in [169], where different FS algorithms were applied.

Five FS methods using only gene expression levels were used in [170] with a new network-based supervised classification method to predict cancer. Here, it was applied to different datasets, (lung, breast, leukemia, and colon cancers).

FS techniques can be organized into three categories: filter methods, wrapper methods and embedded methods. A review about the main contributions of feature selection research in bioinformatics applications can be read in [171]. A summary of the reasoning of this classification and some advantages and disadvantages is presented in the following tables, where for each category are described the procedures used, the advantages and disadvantages and their classes, when applicable.

Filter FS techniques (Table 6.1), assess the relevance of features only with the intrinsic properties of the data, using statistical metrics.

Wrapper FS techniques (Table 6.2), builds classification models and uses its performance as evaluation criterion to determine the importance of features.

Embedded FS techniques, described in Table 6.3, are built into the classifier construction. Different FS methods can be combined using ensemble FS approaches.

A list with a description of classical FS methods applied to DNA micro-array datasets analysis can be found in [172]. Recently, different FS methods have been combined and used in hybrid or embedded FS methods.

The stability of features selection techniques is important to obtain reliable results. The

Table 6.1: Filter features selection techniques.

Procedure	Advantages	Disadvantages	Classes
A score is calculated, and low-scoring features are removed. This subset of features is presented as input to the classification algorithm.	Scalable to very high-dimensional datasets; Computationally simple and fast; Independent of the classifier; Performed only once.	Ignores the interaction with the classifier; Univariate filters ignore feature dependencies.	Feature Ranking or univariate; Subset evaluation or multivariate.

Table 6.2: Wrapper features selection techniques.

Procedure	Advantages	Disadvantages	Classes
Evaluation of a specific subset of features is obtained by training and testing a specific classification model .	Interaction between feature subset search and model selection; Takes into account feature dependencies.	Feature subsets grows exponentially with the number of features; Heuristic search methods are used to guide the search for an optimal subset; Higher risk of over-fitting than filter techniques; Very computationally intensive, because they build a learner for every test.	Deterministic search algorithms; Randomized search algorithms.

Table 6.3: Embedded features selection techniques.

Procedure	Advantages	Disadvantages	Classes
Runs a learner with embedded feature selection and learner performs feature selection prior to analysis.	Includes interaction with the classification model; Less computationally intensive than wrapper.	Still computational intensive.	—

role of stability in feature selection with DNA micro-arrays data is addressed in [173].

Measuring the stability of FS methods requires a testing procedure and a stability measurement. There are several techniques to test and measure the stability of feature selection methods, like using dataset perturbation, the cross validation method and the fixed overlap partitioning [171].

Methods to improve FS stability can be classified into two categories, group features selection category and ensemble features selection category. The following tables (Table 6.4 and Table 6.5) describe the steps and methods used and respective descriptions for each of the categories.

In group features selection, the selection of one feature from each group handles with the problem of redundancy and the stability of the feature subset is increased because all of the features in the dataset are grouped by correlation rather than removed due to redundancy.

Table 6.4: Group features selection category.

Steps	Steps description	Methods	Methods description
Group formation	Different groups of correlated features are identified.	Knowledge driven. Data driven.	Depend on domain knowledge, which are used in dividing the correlated features into groups. Takes only the original dataset into consideration.
Feature selection	From the resulting groups formed selects one feature from each group.	—	—

An application of the ensemble features selection is used in [174], where each feature's final rank is the sum of its rank in the different lists.

In ensemble features selection, several aggregation function can be used, like exponential aggregation, mean and median aggregation, and threshold based aggregation. After multiple ranked lists are created using one of the above methods, the second step of creating an ensemble feature ranker is to use one of many available aggregation functions to aggregate the results that are generated in the first step.

When using supervised methods, a large number of variables can be used to characterize cancer and non-cancer biological datasets, so it is necessary to choose the most relevant ones.

There are several feature selection algorithms, but due to their simplicity and low computational cost, filter algorithms are the most used. One of those algorithms is the ReliefF feature selection method [175], an extension of the original Relief algorithm, here described because it was used in this research.

Given a randomly selected sample $\mathrm{S_i} \in \mathrm{S}$, for $\mathrm{i} = 1, \cdots, \mathrm{s}$, described by a vector of features

Table 6.5: Ensemble features selection category.

Steps	Steps description	Methods	Methods description
Multiple ranked lists are created using one of the above methods; Then uses an aggregation method to aggregate ranks results.	Applies feature selection algorithms multiple times and combines the results into one decision.	Through data diversity. Through functional diversity. Through hybrid ensemble techniques.	Applies a single feature selection method to a number of differently sampled versions of the same dataset and results are aggregated. Applies a set of different feature selection techniques on the same dataset. Applies different feature selection techniques to different sampled versions.

$f_i \in F$, for i = 1, $\cdots$, f, the Relief algorithm is described in Figure 6.1

```
for $i$ := 1 to f
        W[$f_i$]:= 0.0;
for $iter$ := 1 to m {
        randomly select a sample $S_{iter}$;
        find H and M;
        for $i$ := 1 to f
                W[$f_i$] := W[$f_i$] - diff($f_i$,$S_{iter}$,H)/m + sdiff($f_i$,$S_{iter}$,M)/m;
}
```

Figure 6.1: Relif algorithm.

Relief finds for the two nearest neighbors of S_{iter}. One from the same class, the nearest hit H and the other from the other class, the nearest miss M. The number of iterations m is a user defined parameter.

ReliefF is an extension of Relief. The Relief algorithm deals with discrete and continuous values features, but is limited to two class problems and cannot deal with incomplete and noisy data. The ReliefF algorithm can deal with multi-classes and with incomplete and noisy data. ReliefF can be applied in all situations, has a low bias and captures well local dependencies. ReliefF algorithm is efficient and is multivariate, being suitable when there is much feature interaction and ranking well the quality of features, when there is a strong dependency between them [176].

Considering that the diff (f,S_1, S_2) function calculates the difference between the values of f for two samples S_1 and S_2. For discrete values of f we have

$$\text{diff}\,(\text{f,S}_1,\text{S}_2) = \begin{cases} 0, \quad if \quad value(f,S_1) = value(f,S_2) \\ \qquad\quad 1, \quad otherwise \end{cases} \tag{6.1}$$

and for continuous values of f

$$\text{diff}\,(\text{f,S}_1,\text{S}_2) = \frac{|\text{value}\,(\text{f,S}_1) - \text{value}\,(\text{f,S}_2)|}{\max\,(\text{f}) - \min\,(\text{f})} \tag{6.2}$$

The value of feature f on sample S_1 is the value (f,S_1), and max (f) is the maximum value f gets. The function diff() is also used to calculate the distance between samples to find the nearest neighbors and the total distance is the sum of instances for all attributes (Manhattan distance).

The ReliefF algorithm is then described in Figure 6.2.

```
for $i$ := 1 to f
        W[$f_i$]:= 0.0;
for $iter$ := 1 to m {
        randomly select a sample $S_{iter}$;
        for $j$ := 1 to k
                find $H_j$
        for each class $C \neq class(S_{iter})$
                for $j$ := 1 to k
                        find $M_j$(C)
        for $i$ := 1 to f
                W[$f_i$]:=W[$f_i$]-$\sum_{j=1}^{k}$ diff($f_i$,$S_{iter}$, $H$)$(m,k)$+
                $\sum_{C \neq class(S_{iter})} \left[ \frac{P(C)}{1-P(class(S_{iter}))} \sum_{j=1}^{k} \text{diff}(f_i, S_{iter}, M_j(C))/(m,k) \right]$
}
```

Figure 6.2: RelifF algorithm.

P (C) is the prior probability of class C, estimated from the training set, and $1 - \text{P}(\text{class}\,(\text{S}_{\text{iter}})\,)$ is the sum of probabilities of classes misses.

Choosing k hits and misses makes the algorithm robust to noise. Missing values are treated probabilistically. If one sample S_1 has an unknown value

$$\text{diff}\,(\text{f,S}_1,\text{S}_2) = 1 - \text{P}(\text{value}\ \ (\text{f,S}_1)\,|\text{class}\,(\text{S}_1)\,) \tag{6.3}$$

If both samples S_1 and S_2 have unknown values

$$\text{diff}\,(\text{f,S}_1,\text{S}_2) = 1 - \sum_{\text{V}}^{\#\text{values(f)}} (\text{P}(\text{V}|\text{class}\,(\text{S}_1)\,) \times (\text{P}(\text{V}|\text{class}\,(\text{S}_2)\,) \tag{6.4}$$

The relative frequencies from the training set can approximate conditional probabilities.

Another extension for ReliefF is the ReliefF for regression.

In [177], the ReliefF FS method is claimed to be the best method among several tested for cancer classification using gene expression data.

6.3 Supervised Learning Algorithms

The field of machine learning can be divided in the two main categories: unsupervised and supervised learning.

In unsupervised learning, the goal is to explore the data and discover similarities between objects, where classes of the objects are unknown and using a similarity measure to define groups of objects, referred to as clusters, such that objects in one cluster are more similar and in separate clusters are less similar.

In supervised learning, objects in a given collection are classified using a set of attributes, variables, or features. The goal in supervised learning is to design a system able to accurately predict the class membership of new samples based on the available features and a subset of samples with known classes. The classification model is built partitioning the data into a training set that is used to build the model, by applying a classification algorithm and a test set that is used to validate the model and determine its accuracy. The model must have a good generalization capability, in order to predict class labels of new samples.

When only a few labeled objects (samples) are available, and, if there are available many other objects (samples) with unknown classes, a better classifier can be obtained using a semi-supervised learning technique. One way of doing this is assuming that objects with unknown classes from a cluster feature space belong to the same class of known class objects of the cluster [178].

A general review about machine learning applications in genetics and genomics can be found in [179] and a review about their applications to cancer prognosis and prediction can be found in [180].

6.3.1 Pre-Processing and Data quality

To make the data more suitable for unsupervised and supervised learning, some pre-processing tasks can be done, like:

– Aggregate two or more attributes (or samples) into a single attribute (or sample), which allows data reduction;

– Sampling, to find the adequate samples for each study, because when an entire data set cannot be used or may not be necessary, a sample which is a representative subset of the original data can be used, if results are approximately the same, as if the entire data set was used. There are several sampling approaches, like random sampling and stratified sampling. The last is used when subsets vary considerably, and in this case is applied random sampling within each of the subsets.;

- Discretization that converts a continuous attribute to a discrete attribute, and binarization that converts a continuous attribute to a discrete attribute with two possible values, which sometimes is required by some data mining algorithms;
- Attributes transformation, changing all values of an attribute by function transformations or by normalization.

Dimensionality reduction is important for modeling, because many data mining algorithms work better if the number of attributes is lower and if irrelevant features are eliminated or if noise is reduced. With dimensionality reduction, the quality of results can be better, the model can be more understandable and data can be more easily visualized. There are several approaches to dimensionality reduction, and one of the strategies that can be applied to continuous attributes is the principal component analysis (PCA), that uses a linear or non-linear projection of data from a high dimensional space to a lower dimensional space. Another strategy is feature selection already referred in the previous section.

Also data quality must be assured and it is necessary to take care of:

- The noise, because most methods and technologies used are not precise;
- The outliers, because some data can be inconsistent regarding the model used to analyse them, having deviations from normal behavior;
- Missing values, and one way to deal with them is to eliminate attributes that have missing values or estimate them. If the attributes are eliminated, this can cause removing a large number of objects. Missing values can also be estimated if the data set that has many similar data points. To estimate them, the nearest neighbors can be used. For example, if the attribute is continuous, it can be used the average value of the nearest neighbors, and if it is categorical, the most commonly occurring attribute value. There are many data mining approaches that can be modified in order to ignore missing values;
- Duplicate data and inconsistent values should be removed too.

6.3.2 Classification Algorithms

Various classification algorithms have been applied by several researchers to predict cancer diseases. In the following, some of those applications are mentioned.

A comparison between single-gene, gene-set and two PPI network-based methods, using gene expression micro-arrays data, applied to melanoma and ovarian cancer can be found in [181]. In single-gene, features are the expression values of informative genes identified via differential expression analysis. In the gene-set method, genes are grouped into sets using biological knowledge, which are used as features for classification. Three classifiers were used, namely random forest (RF), diagonal linear discriminant analysis and SVM, with 5-fold cross-validation. It concludes that including network information may lead to the identification of more stable gene expression signatures.

In [164], PPI subnetwork markers are found to distinguish between metastatic and non-metastatic tumors, using a score function. Candidate subnetworks are built starting with a single protein and are expanded using the PPI network, until the score stops to increase. The activity scores calculated from the average of the expression levels of each subnetwork were used as feature values. The classifiers used were based on logistic regression and SVM using 5-fold cross validation.

In [182], a score is calculated for the expression values of genes to select those with highest scores as features in the classification withhold-one-out cross validation. Tests were made, and the best results were obtained with 50 genes. [168] uses a BNL prediction to classify lung cancer tumors as SCLC , NSCLC and COMMON classes, using the structural and physiochemical properties of protein sequences obtained from genes using micro-array analysis.

A model for predicting the survival rate of patients affected by lung cancer can be found in [169]. The classification algorithms used were, the decision tree (DT), BNL and neural network (NN).

A new network-based supervised classification method to predict cancer, named NBC and using only gene expression levels is presented in [170]. It was applied to different datasets, (lung, breast, leukemia and colon cancers) using five classification algorithms, namely SVM, k-nearest neighbours (KNN), naive Bayes (NB), C4.5 and RF with 10-fold cross validation. High accuracy classification was obtained with less than 100 genes.

In this research were used three classification algorithms that will be now briefly described: SVM, the KNN and the RF.

Support Vector Machines

SVM [183] is a learning algorithm widely used in computation biology for classification [184, 185]. SVM algorithm was originally proposed to construct a linear classifier [186] and aims to create a decision boundary, called a hyperplane, between two classes, as far as possible from the closest data points from each of the classes, which enables the prediction of labels from one or more feature vectors. These closest points are called support vectors [187].

Considering $\{(\mathbf{x}_\mathrm{i}, \mathrm{y}_\mathrm{i})\}$, for $\mathrm{i} = 1, \cdots, \mathrm{n}$, and $y_i \in \{-1, 1\}$, where $\mathbf{x}_\mathrm{i}$ is a feature vector representation and y_i the class label (negative or positive) of a training component i, the optimal hyperplane can then be defined as:

$$\mathrm{wx}^{\mathrm{T}} + \mathrm{b} = 0 \tag{6.5}$$

where w is the weight vector, x is the input feature vector, and b is the bias.

For all elements of the training set, w and b would satisfy the following inequalities:

$$\begin{cases} \mathrm{wx_i}^{\mathrm{T}} + \mathrm{b} \geq 1 \text{ if } \mathrm{y_i} = 1 \\ \mathrm{wx_i}^{\mathrm{T}} + \mathrm{b} \leq -1 \text{ if } \mathrm{y_i} = -1 \end{cases} \tag{6.6}$$

When training an SVM model, we are looking for w and b, so that the hyperplane separates the data and maximizes the margin $1\ /\|\mathrm{w}\|^2$.

The support vectors are the vectors $\mathrm{x_i}$ for which $\mathrm{wx_i}^{\mathrm{T}} + \mathrm{b} = 1$ if $\mathrm{y_i} = 1$ and the vector $\mathrm{x_i}$ for which $\mathrm{wx_i}^{\mathrm{T}} + \mathrm{b} = -1$ if $\mathrm{y_i} = -1$.

Real world data analysis requires often nonlinear methods. To model higher dimensional, non-linear models, it can be used the kernel method [188], which allows to add additional dimensions to the data and make it a linear problem in the resulting higher dimensional space. The kernel corresponds to a dot product in a (usually high-dimensional) feature space and, in this space, estimation methods are linear.

The Kernel function can be defined as

$$\mathrm{K}\left(\mathrm{x},\mathrm{x}'\right) = \langle \mathrm{f}\left(\mathrm{x}\right),\ \mathrm{f}(\mathrm{x}') \rangle \tag{6.7}$$

where x, x' are n dimensional inputs. f is used to map the input from a n dimensional to a m dimensional space, and <.,.> denotes the dot product. Using kernel functions, the scalar product between two data points can be calculated, in a higher dimensional space, without explicitly calculating the mapping from the input space to the higher dimensional space. A kernel function projects data from a low-dimensional space, where usually the data cannot be separable to a space of higher dimension, where the data will become separable in the resulting higher dimensional space. That depends on the kernel function chosen.

Several researchers have been using SVM classifiers to investigate cancer diseases. A linear SVM was used in [189] to classify two different types of leukemia using gene expression micro-array data. SVM was also applied in a colon cancer tissue classification using selected features in [190]. In [191] SVM was used to detect persons with diabetes and pre-diabetes in a cross-sectional representative sample of the U.S. population. Breast cancer subtypes were classified using an SVM model and using proteomics data in [192] and to unveil cancer and breast cancer were also used with SVM classifiers using single nucleotide polymorphisms (SNPs) data.

K-nearest Neighbors

The KNN is non parametric algorithm, meaning that it does not make any assumptions on the underlying data distribution, which is useful, since in real world, most of the data does not fit exactly to the theoretical hypothesis made. KNN can be used for both classification and regression predictive problems.

KNN assumes that the data is in a feature space, so data points, which can be scalar or multidimensional vectors, are in a metric space and distances (Euclidean or other) can be calculated.

A sample can be classified according to the class label of its neighbors. In classification it determines the class label of a sample with unknown class label, using the class labels of the k number of nearest neighbors of this sample and a distance metric to compute distance between samples to choose the smallest distances. A nearest neighbor classifier represents each instance as a data point embedded in a d dimensional space, where the number of continuous attributes is d.

A comparison between several nearest neighbor techniques can be found in [193]. Some of them are improvements of the basic KNN to gain speed and space efficiency. In [194], KNN is used combined with genetic algorithm to diagnosis the heart disease. The KNN algorithm was also used for the classification of lymph node metastasis in gastric cancer.

Random Forest

A RF is an ensemble classifier that uses many decision trees models. A different subset of the training data is selected, about 2/3, with replacement, to train each tree. Then, remaining training data is used to estimate error and variable significance information, and the class assignment is made by the number of votes from all of the trees [195].

Some applications of this algorithm are in ecology [196], in multi-class object detection [197].

A new method of gene selection in classification problems based on RF of micro-array data is presented in [198]. The authors concluded that RF has comparable performance to other classification methods, including KNN, and SVM. In [199], a comparison of RF and SVM for micro-array-based cancer classification was made. Here authors claim that on average and in the majority of micro-array datasets, SVM outperform RF.

A domain-based RF of DT decision trees to infer PPI protein interactions is applied in [200], comparing the results with the maximum likelihood approach.

A survey of RF developments is presented in [201], with emphasis in bioinformatics and computational biology application. They mention several related aspects, like the available implementations of the selection of parameters.

A review of RF applications to genomic data, including prediction, variable selection, pathway analysis, genetic association, and epistasis detection can be found in [202].

6.3.3 Performance Evaluation Metrics

To assess the performance of inference methods several measures have been proposed [110]. They were divided in three categories: general statistical-based measures, ontology-based measures and network-based measures.

Network-based measures consider the network structure, and they can be categorized as network-based measures that use topological descriptors, graphlets or motifs and global network-based measures.

Ontology-based measures use biological information when trying to quantify the biological relevance of the inferred network [110].

General statistical-based measures evaluate networks performance by scalar values and assume that the inference process is homogeneous.

Some of the general statistical-based measures used are defined in Table 6.6. The sensitivity, also called recall and true positive rate (TPR), that, in binary classification, is the number of positives correctly identified. The specificity is equal to 1- false positive rate (FPR) and measures the number of negatives correctly identified. The precision that indicates the reproducibility of the measurement. The F1-score (F1), which is the harmonic mean of precision and sensitivity. The classification accuracy (CA), that is the proximity of measurement results to the true value. At last, the AUC, which is the area under the ROC curve. The ROC curve is obtained by plotting the TPR against the FPR at various threshold values.

6.4 Dynamic Network Models

Applying networks to model biological processes allows the representation of the interactions between the bio-entities involved. The calculus of the topological properties of these networks seeks to quantitatively characterize these networks for comparison, to discover their inherent topological patterns.

These networks vary, depending on the conditions used to build them. Evaluating a biological system implies to evaluate their associated dynamics through a condition or several conditions, by building a set of networks representative of the interactions between bio-entities of the system to create a topological dynamic model. Analyze the dynamics implies to discover patterns that characterize it, and each set of networks, built for each condition, represents the system dynamics through the respective condition used to build the set. This a view of the overall dynamics of the biological system, being each network a snapshot for the specific condition of the dynamic biological system modeled.

Combining the information obtained from gene expression and protein-protein interaction networks analyses, and building series of complex networks to model system dynamics is a recent common trend, used that have been contributing to the identification of diseases. By

Table 6.6: General statistical-based measures.

Measure	Equation
Sensitivity	$$sensitivity = \frac{TP}{TP + FN} \quad (6.8)$$ TP are the true positives FN are the false negatives
Specificity	$$specificity = \frac{TN}{TN + FP} \quad (6.9)$$ TN are the true negatives FP are the false positives
Precision	$$precision = \frac{TP}{TP + FP} \quad (6.10)$$
F1-score	$$F1 = \frac{2 \times sensitivity \times precision}{sensitivity + precision} = \frac{2TP}{2TP + FP + FN} \quad (6.11)$$
Classification Accuracy	$$CA = \frac{TP + TN}{TP + TN + FP + FN} \quad (6.12)$$
AUC	$$AUC \ = \ \frac{1}{2} (\frac{TP}{TP + FN} + \frac{TN}{TN + FP}) \quad (6.13)$$

constructing series of different complex networks across multiple conditions, like species, time, evolutionary states, or specific traits, the dynamics can be captured and modeled. Those series of networks are usually constructed using different parameters and through the analysis of their topology and of the finding of specific topological structures or signatures, allow the understanding of their similarities and dissimilarities. Extracting the network topology dynamics embedded in a disease system can improve the capacity of diseases prediction and of the understanding of the mechanisms behind their evolution.

In [203], normal, benign and malignant states of breast cancer are differentiated, building a gene regulatory network representative of each state and comparing several network topological properties, like the in and out-degree, the betweenness, the cluster coefficient and the closeness. Gene ranking was made selecting 53 hub genes.

Another approach used differential co-expression analysis and PPI networks for studying human hepatocellular carcinoma progression at five different stages [204]. Dominietto et al. [205] show how to integrate imaging data into networks to define tumor fingerprints through both network topology and the detection of dynamic connectivity patterns.

6.5 Proposed Model Based on Sample-series Networks

SSN is the proposed novel approach, based on the hypothesis that different individuals or different conditions (healthy and non-healthy cells of the same individual) have inherent new patterns important to be captured, that are part of the dynamics of the system to be studied [166]. The new methodology used to obtain SSN network-based features is schematically described in Figure 6.3, where is shown how to build SSN from genes expression data and PPI data, and to determine the SSN network-based features. SSN are undirected PPI subnetworks of the whole PPI network.

Details about the methodology to build the SSN and calculate their topological properties is described in the algorithm of Figure 6.4. The SSN topological properties will be used as features to create the prediction models using the classification algorithms. Using micro-arrays technology, each sample represents one instance of those conditions where different gene expression values were measured. These different values, obtained from each sample, are used to build a set of PPI networks, each of them capturing the interactions representative of the sample from which were obtained.

Samples dynamic may reveal new signatures to help the identification of diseases, for example distinguishing cancer from non-cancer tissues, having in account samples dynamics between groups.

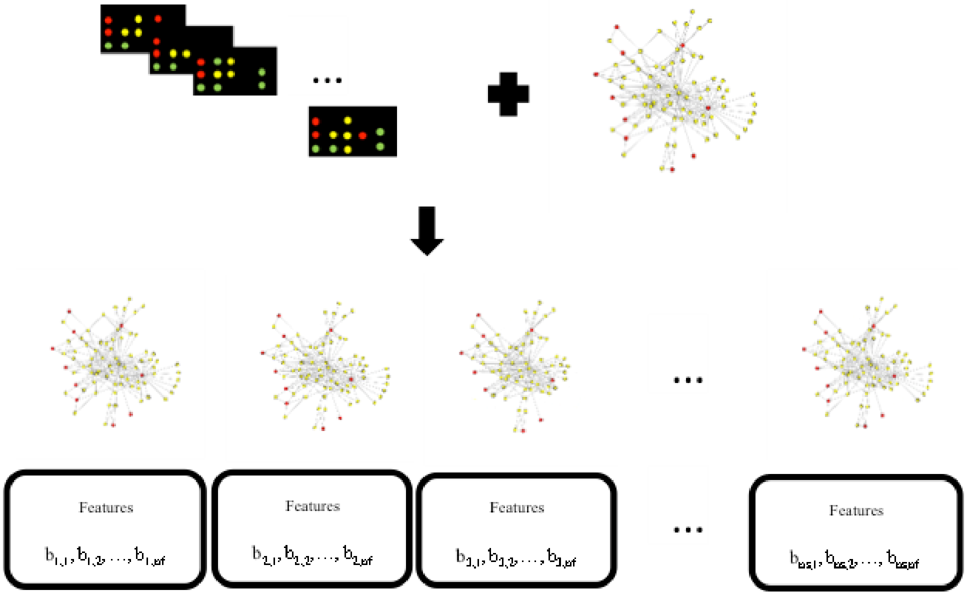

...
Features
$b_{1,1}, b_{1,2}, \ldots, b_{1,nf}$
Features
$b_{2,1}, b_{2,2}, \ldots, b_{2,nf}$
Features
$b_{3,1}, b_{3,2}, \ldots, b_{3,nf}$
...
Features
$b_{ns,1}, b_{ns,2}, \ldots, b_{ns,nf}$

Figure 6.3: SSN network-based features, where ns is the number of samples and nf is the number of features.

6.6 Prediction of Cancer using Networks Topological Features

Cancer is a complex genetic disease that affects an increasing number of citizens all over the world. In 2015, more than 1.6 million new cancer cases were expected in the United States, from which around 15% correspond to breast cancer [206], and in 2018 are expected more than 1.7 million new cancer cases diagnosed in the United States, from which around 15% correspond to breast cancer. This corresponds to an increase of approximately 1 million in 3 years [207]. Understanding the underlying biological mechanisms behind this disease has been the goal of many and continuous research initiatives.

Considering that there are many high-throughput experiments, regarding different diseases and with different samples, that can represent different individuals or different conditions of the same individual, a supervised approach can be used to find models, that will allow the static and dynamic study of these diseases.

So, one way to study cancer is using high-throughput technologies, which allows the parallel analysis of genes expression in several samples together with the PPI data induced by the expressed data to build models of this disease system. In a network-based approach, bio-entities such as genes and proteins can be represented as nodes and their relationships as edges. Using this approach, biological processes can be modeled to be analyzed using graph and network methods. More, the construction of network-based models to study complex phenomena, like cancer diseases, allows the capture of the embedded systems dynamics, when

Step 1: Obtain the e matrices, for $e = 1, \ldots, ne$, where the number of microarray experiments is ne, the number of samples of the experiment e is ns_e and the number of genes is ng.

$$EXP_e = [exp_{ij}], \quad i = 1, \ldots, n_e; \quad j = 1, \ldots, ng \tag{6.14}$$

Step 2: Obtain the lists of the top genes ranked by decreasing order of ReliefF

$$LTGR_e, \quad for \quad e = 1, \ldots, ne \tag{6.15}$$

Step 3: Obtain the union of the previous lists, for a threshold value, thr_rf that defines the number of top elements of the lists to be considered.

$$LUNION = UNION(LTGR_e), \quad for \quad e = 1, \ldots, ne \tag{6.16}$$

Step 4: Obtain the sub-matrices of $EXP_e, \quad for \quad e = 1, \ldots, ne$, obtained in Step 1, for the genes selected in Step 3.

$$EXP = [sexp_{ij}], \quad for \quad i = 1, \ldots, sum_e(ns_e); \quad j = 1, \ldots, thr\_rf \tag{6.17}$$

Step 5: Obtain the lists of the top thr_me most expressed genes in $SEXP$, for each sample from e experiments, for $e = 1, \ldots, ne$.

$$\begin{aligned} LGME = [lgme_{ij}], \quad for \quad i = 1, \ldots, sum_e(ns_e);& \\ j = 1, \ldots, thr\_me& \end{aligned} \tag{6.18}$$

Step 6: Obtain the lists of the proteins encoded by the genes of the $LGME$ matrix, $P(LGME)$ for each sample of the experiments e, for $e = 1, \ldots, ne$.

$$LPME = P(LGME)_i, \quad for \quad i = 1, \ldots, sum_e(ns_e) \tag{6.19}$$

Step 7: Obtain the SSN, the PPI human interaction sub-networks induced by $LPME$.

$$SSN = [ssn_i], \quad for \quad i = 1, ..., sum_e(ns_e); \quad e = 1, \ldots, ne \tag{6.20}$$

Step 8: Calculate the five groups of topological properties for each ssn belonging to SSN, where the number of features, nf, is the number of descriptors used.

$$\begin{aligned} FEAT = [feat_{ij}], \quad for \quad i = 1, \ldots, sum_e(ns_e); \quad j = 1, ..., nf;& \\ e = 1, ..., ne& \end{aligned} \tag{6.21}$$

Figure 6.4: Algorithm to build SSN and calculate their topological properties.

we obtain different snapshots of the PPI networks induced by the expressed genes.

A new methodology was developed, that uses a supervised classification approach and is described in [166] for extracting the network topology dynamics embedded in a disease system, to improve the capacity of cancer prediction, using exclusively the topological properties of biological networks as features. It is a system-based approach that classifies between cancer and non-cancer tissues and aims to contribute to find signatures that distinguish disease biological processes from healthy biological processes, using the topological properties of the built networks without prior biological knowledge and considering the network topological dynamics embedded in the disease and health systems.

6.6.1 Data Sets

In this work were used four datasets, three from breast cancer micro-array experiments and one from a liver cancer microarray experiment. The experiments were obtained from ArrayExpress [98]: E-GEOD-65194 (178 samples, where 167 are from breast cancer tissue cells), E-GEOD-54002 (433 samples, where 417 are from breast cancer tissue cells), E-GEOD-29044 (124 samples, where 75 are from breast cancer tissue cells) and E-MTAB-950 (276 samples, where 179 are from liver cancer tissue cells). To assure probes and samples uniformity all experiments share the same array design A-AFFY-44 and all samples were labeled as belonging to one of the two classes, Cancer or Healthy.

DAVID and universal protein resource (UniProt) were primarily used [137, 208] to obtain the mapping of identifiers from probe-set ids and gene names to proteins. The human PPI dataset was obtained from STRING, an on-line database resource with several distinct types and sources of PPI information.

The number of genes obtained from the experiments was 54673 genes, which were sorted by decreasing values of ReliefF. For each experiment, the top ReliefF 100 genes were selected and merged in one matrix of 735 samples and 276 genes for breast cancer and one matrix of 124 samples and 276 genes for the liver cancer.

6.6.2 Network-based Approach

The proposed methodology to construct protein-protein interaction networks to capture the existent system dynamics beneath their topology, using genes expression data, uses SSN. These networks were built from a group of cancer and healthy micro-array samples. Using a supervised binary classification methodology, to classify between cancer and non-cancer tissues samples, features information are exclusively obtained through the analysis of the topological properties of these networks. The methodology to build SSN and to determine the features from them was describe in the Section 6.5.

To capture the structural complexity and dynamics embedded in these biological networks, network topological properties values were obtained from the topological analysis of each network belonging to the SSN. Five groups of descriptors were used, a first group of 15 descriptors, named D0, a second group of 16 descriptors, named D1, a third group of 14 descriptors, named D2, a forth group of 6 descriptors, named D3 and a fifth group of 58 descriptors, named D4. All these descriptors were described in Section 2.2.

The first D0 group of 15 descriptors was calculated using the NetworkX from Python [209] (Table 6.7). The second D1 group of 16 descriptors that uses distances between nodes to capture the structural complexity of the network, the third D2 group, of 14 descriptors and the forth D3 group of 6 more recent descriptors, all of them were calculated using the QuACN R package [210] (see Table 6.8 and tables of the Section 2.2.2). The fifth D4 group of 58 descriptors were calculated using the gtriesScanner software [211] (Table 6.9).

In the 58 descriptors of D4 group, the first 29 correspond to the relative frequency values of 3 to 5 nodes subgraphs (graphlets) if they are a motif, zero if they are not a motif, and the last 29 were the correspondent z-score values, which were calculated using 1000 random networks (Table 6.9). These were described in Section 2.3 and in [23].

Table 6.7: Group D0 of topological network-based descriptors.

D0.1: Number of nodes	D0.9: Size of the largest clique
D0.2: Number of edges	D0.10: Number of maximal cliques
D0.3: Density	D0.11: Degree assortativity coefficient
D0.4: Number of connected components	D0.12: Estrada index
D0.5: Number of nodes of the largest component	D0.13: Graph transitivity
D0.6: Number of edges of the largest component	D0.14: Average clustering coefficient
D0.7: Diameter of the largest component	D0.15: Average shortest path length
D0.8: Global clustering coefficient	

A motif is a subgraph that is frequent compared to their frequency in a set of similar random networks and in this work a subgraph was considered a motif [25], if the frequency of the subgraph in the network is superior to 4, the difference between the frequency in the network (f) and the average frequency in 1000 similar random networks (avg_{fr}) is greater or equal to 0.10 of the average frequency in those random networks, and $|\mathrm{z_{score}}| > 2$, where $\mathrm{z_{score}} = (\mathrm{f} - \mathrm{avg_{fr}}) / \mathrm{sd}$, with sd as the standard deviation.

The network-based method used and described in Section 6.2, builds a set of PPI networks, one for each sample belonging to the SSN and uses the ReliefF algorithm [175] to rank a subset of genes. The ReliefF algorithm can be applied to both continuous and discrete values.

Using the human PPI dataset, several networks were constructed, one for each sample, representing the entire set of PPI for all different samples. In each SSN network, nodes are proteins coded by a subset of the most expressed genes of the top ReliefF genes, and edges indicate that the proteins coded by those genes interact physically. A score was used as a threshold for the PPI. Only PPI with score greater or equal to 300 were considered. These networks were constructed as undirected, unweighted and with no self-edges.

To build the binary classification models, three different supervised learning algorithms were used. The set of supervised learning algorithms used were the KNN classifier, the SVM classifier implemented using an radial basis function (RBF) kernel, and the RF, all with default parameters.

Network topological properties were used as features of the supervised binary classification methodology used, and their values were obtained from the topological analysis of each SSN network. These classification models were evaluated using several statistical measures.

Three strategies were used, the first one with classification results obtained by 5-fold cross-validation and the others two using a separate test data, one from the same type of cancer, using data obtained from two of the breast cancer datasets as train set, and data obtained from the other breast cancer dataset as test set, and the other from a different type of cancer, using data obtained from the three breast cancer datasets to train the dataset and, for testing, using data obtained from the liver cancer dataset.

The first strategy, named Case 1 (C1), used 5-fold cross-validation on the network-based features values calculated from the three breast cancer micro-array datasets.

The second strategy included two types of tests that were named Case 2 (C2) and Case 3 (C3) and here two of the breast cancer datasets were used to calculate network-based features values for the training dataset, and the remaining one was used to calculate the network-based features values for the test dataset. In C2, the training set used was the E_GEOD-65194 and the E-GEOD-54002 micro-array datasets and the test set used was the E_GEOD-29044 dataset and in C3, the training set used the E_GEOD-54002 and the E-GEOD-29044 datasets and the test set used was the E_GEOD-65194 micro-array dataset.

The third strategy, named Case 4 (C4), used data from the three breast cancer micro-array datasets for the training dataset and the liver micro-array dataset was used for the test dataset. This third strategy was used to check if the classification models with datasets of one type of cancer can be generalized for another cancer type.

To analyze which of the network-based features contributed more for the classification model a ranking list of the top 5 features was done.

The statistical measures used to evaluate the performance of the binary classification models were (see Table 6.6), the CA, the precision (Precision), the recall (Recall), the F1 and the AUC [212] and values were obtained using three strategies, with different sets of features

Table 6.8: Groups D1, D2 and D3 of topological network-based descriptors.

D1.1: Wiener	D2.1: Total adjacency	D3.1: Medium articulation
D1.2: Harary	D2.2: Zagreb 1	D3.2: Efficiency
D1.3: BalabanJ	D2.3: Zagreb 2	D3.3: Graph index complexity
D1.4: Mean distance deviation	D2.4: Modified Zagreb	D3.4: Off diagonal
D1.5: Compactness	D2.5: Augmented Zagreb	D3.5: Spanning tree sensitivity STS
D1.6: Product of row sums	D2.6: Variable Zagreb	D3.6: Spanning tree sensitivity differences STSD
D1.7: Hyper distance path index	D2.7: Randic	
D1.8: Dobrynin eccentricity graph	D2.8: Complexity index B	
D1.9: Dobrynin avgecc of G	D2.9: Normalized edge complexity	
D1.10: Dobrynin eccentric graph	D2.10: Atom bond connectivity	
D1.11: Dobrynin graph integration	D2.11: Geometric arithmetic 1	
D1.12: Dobrynin unipolarity	D2.12: Geometric arithmetic 2	
D1.13: Dobrynin variation	D2.13: Geometric arithmetic 3	
D1.14: Dobrynin centralization	D2.14: Narumi Katayama	
D1.15: Dobrynin average distance		
D1.16: Dobrynin mean distance vertex deviation		

Table 6.9: Group D4 of topological network-based descriptors.

D4.1_j: 3_j_fr, j=1,... 2	D4.4_j: 3_j_zsc, j=1,... 2
D4.2_j: 4_j_fr, j=1,... 6	D4.5_j: 4_j_zsc, j=1,... 6
D4.3_j: 5_j_fr, j=1,... 21	D4.6_j: 5_j_zsc, j=1,... 21

groups.

For these statistical measures, TP is the number of correctly predicted samples that belong to the class, TN is the number of correctly predicted samples that do not belong to the class, FP is the number of wrongly predicted samples that belong to the class and FN is the number of wrongly predicted samples that do not belong to the class. CA measure calculates the proximity of measurement results to the true value and gives the global efficacy of a classifier. Precision (Precision) measure specifies the positive labels given by the classifier that are correct. Recall or sensitivity measure shows the efficacy of a classifier to identify positive labels. F1 is the harmonic mean of precision and recall and is between 0 and 1, being 1 the best value. Finally, the AUC is the classifier's capacity to avoid false classification.

6.6.3 Results and Conclusion

The objectives of this research were, to find out if there were evidences of signatures beneath the SSN, that allowed the classification of samples as cancer or non-cancer samples, to select the topological measures that give better results as classification features among the several groups considered and to discover if a classification model can distinguish different types of cancer.

The two sets of features considered, were the set of all of the network-based features (groups D0 to D4) and the set of network-based features of the group D4. The results obtained are shown in Table 6.10. In the C1 case, 5-fold cross validation was used, with results, above 0.95, for all the statistical measures considered. To test if the classification obtained in C1 was suffering from over fitting, different datasets were used as a train set and as test set, the cases C2 and C3 and the results obtained were, for example for CA, above 0.80 for C2 and above 0.92 for C3, which evidence good performance of the classifier. The difference between the values of C2 and C3 may be explained by the imbalance between cancer and non-cancer samples.

To check if the classification models with datasets of one type of cancer can be generalized for another cancer type, the classification model was trained with data from three breast cancer datasets and tested with data obtained from a liver cancer dataset, in case C4. Values obtained and shown in Table 6.10, are still positive, probably due to the fact of all being cancer diseases, but worse than the previous ones.

To analyze which of the network-based features contributed more for the classification, model features were ranked and the list of the top 5 network-based features found is shown in Table 6.11.

From the analysis of which features are more informative, it can be stated that the most relevant features belong mainly to group D0 and group D4. When all groups of topological features are used as features variables, it can be seen that the size of the largest clique and the number of nodes are better ranked. Motifs of size 4 and 5 are the most informative motifs.

Table 6.10: Statistical evaluation (CA, Precision, Recall, F1 and AUC) of the binary classification C - Cancer and H - Healthy for the cases C1, C2, C3 and C4 using the three classifiers KNN, SVM and RF, for all of the network-based features and for the group of network-based features D4 for the class C.

Cancer		D0+D1+D2+D3+D4			D4		
		KNN	SVM	RF	KNN	SVM	RF
CA	C1	0.98	0.96	0.96	0.95	0.96	0.96
	C2	0.80	0.81	0.88	0.76	0.81	0.85
	C3	0.97	0.96	0.98	0.92	0.96	0.98
	C4	0.61	0.62	0.60	0.58	0.62	0.57
Precision	C1	0.98	0.99	0.98	0.98	0.98	0.97
	C2	0.76	0.78	0.88	0.73	0.79	0.84
	C3	0.98	0.99	0.98	0.95	0.97	0.98
	C4	0.64	0.70	0.69	0.67	0.76	0.64
Recall	C1	0.97	0.97	0.98	0.97	0.97	0.98
	C2	0.97	0.95	0.93	0.95	0.93	0.93
	C3	0.99	0.96	1.00	0.97	0.99	1.00
	C4	0.88	0.72	0.69	0.68	0.60	0.79
F1	C1	0.98	0.98	0.98	0.97	0.98	0.98
	C2	0.85	0.86	0.90	0.83	0.85	0.89
	C3	0.99	0.98	0.99	0.96	0.98	0.99
	C4	0.75	0.71	0.69	0.67	0.67	0.71
AUC	C1	0.96	0.98	0.98	0.96	0.97	0.97
	C2	0.86	0.94	0.94	0.82	0.87	0.93
	C3	0.94	0.98	0.99	0.93	0.99	0.99
	C4	0.55	0.64	0.60	0.56	0.60	0.51

The statistical evaluation results were obtained using only topological properties as features variables, measured in the SSN, which are PPI networks built from the expressed genes without any other biological information and the results seem to indicate that there are signatures embedded in the topology dynamics, modeled through the SSN, which can distinguish cancer from non-cancer cells for each type of cancer.

From the three different supervised learning algorithms used, the KNN, SVM with RBF kernel and RF classifiers, all with default parameters, all gave similar results, with a slight advantage in some statistical measures for RF, when using information from breast cancer datasets.

This new methodology of creating SSN allows the capture of the topology dynamics of the biological system through the set of samples and allows data to be reduced and be computationally manageable, keeping the more informative data. These statements are supported by the good results obtained. This novel approach is worth, and gives different contributions compared to previous works, namely: the number of considered topological properties is much higher; the exclusive use of topological properties (global and local) with good binary classification results obtained; the topological dynamics of the system captured through each sample, different from other works that use time or states for example, which can contribute to the capture of different signatures that could help in the differentiation between disease and healthy systems.

Table 6.11: Top five ranking of network-based features.

Tests	1st	2nd	3rd	4th	5th
D0+D1+D2+D3+D4 (case 1)	D0.9	D0.5	D0.1	D3.2	D4.3_1
D0+D1+D2+D3+D4 (case 2)	D4.2_6	D0.1	D0.5	D0.6	D0.2
D0+D1+D2+D3+D4 (case 3)	D0.9	D0.4	D4.2_2	D0.13	D0.1
D0+D1+D2+D3+D4 (case 4)	D0.9	D0.5	D0.1	D3.2	D4.3_1
D4 (case 1)	D4.3_1	D4.1_1	D4.1_6	D4.2_6	D4.5_3
D4 (case 2)	D4.2_6	D4.3_17	D4.3_4	D4.3_2	D4.3_1
D4 (case 3)	D4.3_18	D4.3_19	D4.2_5	D4.3_17	D4.1_2
D4 (case 4)	D4.3_1	D4.1_1	D4.2_1	D4.2_6	D4.5_3

The results obtained show that classification models should be different according to the cancer disease type considered. More, the knowledge of which features are more informative can be used, in the future, to look for signatures, based in these features that could help in the identification of certain cancer types.

Cliques are fully connected subnetworks where genes are functionally related and highly expressed and were considered by some researchers as gene signatures [163] and one of the most discriminative feature obtained was the size of the largest clique. Motifs of size 4 and 5 were also two of the most discriminative features obtained, so the relative frequency and z_{score} of some motifs as local topological properties measures, showed to be discriminatory features, indicating that there are clues that some small subnetworks could help to distinguish cancer samples. Adding more biological information to the more discriminative features found in the classification may reveal important signatures like subgraphs markers of cancer diseases. This approach also seems worth to be further explored.

The obtained results corroborate the potential of the proposed methodology to predict a certain type of cancer and the necessity of applying different classification models to different types of cancer.

Finally, the proposed methodology for creating SSN is a novel contribution that can be extended to other types of networks, besides PPI, adding information that can differentiate samples and capture their topological dynamics helping to uncover new signatures that can be biologically relevant for the identification of diseases.

6.7 Summary

This chapter began by describing the importance of the study of diseases such as cancer, due to the number of cases and increasing number of occurrences. It was proposed a supervised method that builds a set of PPI networks, the SSN, that captures the dynamics of the system across samples, using genes expression values of the genes that codify those proteins. It was also described the great variety of topological properties used as features for the construction of the classification models, the application of this methodology in the prediction of some types of cancer diseases, and the statistical analysis performed, that allowed to justify the good results obtained.

Chapter 7

Final Remarks

We started this work with the hypothesis that biological networks can be successfully used to get a better insight of the diseases mechanisms that could be potentially used to develop new diagnostic strategies. This research is relevant both from the economic point of view and from the point of view of the society's well-being evolution, in the sense that contributes with new computational methods that can be used to study healthy and diseased mechanisms using exclusively, the topology of their network-based models.

First mathematical concepts, related to graph theory, were published in the year of 1736 by Leonhard Euler, but the application of graph theory to large amounts of data was only possible after the emergence of new technologies and new computational methods. This contributed to the arise and to the great advance of the computational biology and related areas, and on the other hand generated a great amount of new data.

The graph and network theory allows the representation of interactions between objects, the network nodes, by links or edges. Those connections have a meaning and can be assigned to them weights, representing the network relevance of the edge or its strength. A network can be characterized by topological measures that are not linked to a referential system, which allows different degrees of abstraction and generalization. By representing a system through networks allows several views of the interactions between their objects, contributing to the static and dynamic study of the system.

The biological systems can be studied to understand the mechanisms associated with healthy or diseased states of the organisms. These states are not static and may vary according to several parameters, being necessary to study their dynamics. For example, cell functions require the collaboration of sets of proteins and are induced by genes, whose expression values are variable. By joining the information about the physical interaction between proteins and about genes expression, it is possible to verify when these genes are active, for a certain healthy, or not healthy process, and under what circumstances. Then it is possible to create various network-based models of protein interactions that vary according to the considered parameters. Also, by studying the inherent topology of these network models, it is possible to generalize them and obtain more precise network-based models. By looking for topological patterns or signatures, these can be related to healthy and disease mechanisms and can help to find drug targets.

The analysis and comparison of the biological networks topology, in particular PPI, allows the study of these systems using a systemic view, without the use of external information.

Therefore, it was considered relevant to explore how to characterize PPI using their topological properties, since there are still some biological systems less studied. It is recognized the importance of studying PPI, the physical connections between proteins, since they do not act alone and form complexes and functional groups.

Another research considered relevant was to consider the expression of genes, in order to find groups of genes having comparable expression patterns and represent them with a network-based model. Considering the interactions between these groups, and using biological criteria that are considered to be present in the networks, allows to relate these groups with certain risk factors of a given disease.

Since the technologies and methods used to collect expression data and protein interactions are not accurate, and because there are still biological systems not well known, it was also pertinent to explore the use of new computational methods that exclusively use networks topology, in order to obtain more accurate models of protein networks.

Finally, considering that the identification of dynamic topological patterns in PPI networks can help to predict diseases and that they may vary according to several parameters, it is explored the use of a new dynamic networks-based model that uses a supervised approach and the topology of networks, to predict cancer diseases.

7.1 Main Contributions

The first research consisted in the quantitative characterization of PPI of the oral cavity and is detailed in Section 4.2. The oral cavity protein interactome was not yet studied using a systemic view. The main contributions of this study consisted in obtaining various PPI network models according to different confidence scores and prediction methods, and to use the global topological properties to characterize and analyse these networks and compare them with the respective random networks. The experiments revealed that the largest component of these networks represented almost the whole network, and their results showed the variation of some topological measures in the various models, concluding that their topology reveals some hierarchical modularity and small world properties. Experiments also identified those network models that, regarding their node degree distributions, can be considered power-laws, and calculated their respective parameters, using maximum-likelihood fitting methods and goodness-of-fit tests based on the KS statistic. The p-values obtained showed that the power-law distribution model is consistent for the majority of network models built.

The second research consisted of the analysis of a genes co-expression network, where the WGCNA methodology was applied to the HNSCC expression data. This research is detailed in Section 4.3. The main contributions of this research were the better comprehension of the disease HNSCC, that has a high rate of incidence and in spite of having already been associated with several risk factors, still lacks a full comprehension of the genomic processes associated with it. A biological criterion was used to determine the interactions weight of the gene co-expression network model, and after the identification of modules, using unsupervised clustering techniques, was built a network of these modules representatives. It was determined the relationship between the disease and some of their risk factors and it was shown to what degree some risk factors were related with this disease. It was also possible to identify biological functions associated to some of the modules, considered more correlated to a risk factor, using gene ontology.

The third research, detailed in Chapter 5 consisted in developing a network-based method to denoise PPI networks using exclusively the topology. Main contributions of this research were the proposal of a new methodology, the OM method, which is based uniquely on the topology of the biological networks and a new NC topological measure that, when tested in three datasets of two different organisms and validated against two other methodologies showed better results. The results obtained corroborated that OM methodology could still work well in less-studied interactomes, when the subset of the interactome of interest is a representative sample of the structure of the entire interactome. The fact that OM is a network-based method, relying exclusively in the topology of biological networks, without needing extra biological knowledge, turns this methodology simpler and faster, which can be applied to other biological networks and to other domains, as long they have an inherent structure.

The last research, detailed in Chapter 6 combined information obtained from gene expression data and PPI to build series of complex networks to model the system dynamics using a supervised classification approach. This study aims to contribute with a computational method able to find signatures that distinguish between diseased biological processes and healthy ones, using the topological properties, without biological prior knowledge, in a set of networks that reflects the topological dynamics embedded in the systems. This study was applied to cancer diseases, which are considered complex genetic diseases, that affect an increasing number of citizens all over the world. Main contributions were the proposal of a new methodology, the SSN method. This method captures the topology dynamics of the system data, through a set of samples from different individuals or different cellular tissues. This is done by selecting the topological measures that give better results, among a very high number of topological properties, contributing to reduce the data and to be computationally manageable. It captures different signatures, compared to other dynamic models, that can help in the differentiation between diseased systems and healthy ones. Being grounded in the topology of networks, without external knowledge, this methodology can be scalable to other types of networks and other domains.

7.2 Future Research Directions

This research allowed the advance of the existent knowledge in this field. The topological properties can be emphasized, when developing computational methods, using network-based models, in order to capture network topological distribution trends, the dynamic of the modeled systems and their patterns and signatures. Also, the developed methods were applied in biology and medicine field, but it is worthy to explore their dissemination to other fields of knowledge.

In the area of biology and medicine, and despite the obtained results, it would be interesting to test OM methodology using as topological properties graphlets and motifs, to denoise networks. Also, it could be worth to try combine different series of networks built considering a combination of parameters, when capturing the dynamics. Another idea could be to join several gene co-expression networks created using a combination of biological criterion. Another research line could be the use of networks topology to identify overlaps in networks that model gene and protein interactions of different diseases in order to allow a study, at a meta-data level, that would contribute to a better understanding of the interdependencies of diseases and drug interactions, considering parameters such as phenotypes, prescribed drugs and environment and geographical data. Finally, other strategies can also be used for representing and studying graph/network structured data, like deep learning.

Considering the dissemination of methods to other areas of knowledge, it can be investigated the use of proposed methodologies in other areas of knowledge like in the area of Earth Sciences, where the systems are dynamic according to different parameters and already exist large amounts of parametrized data to analyze.

References

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics", *Physics reports*, vol. 424, no. 4-5, pp. 175–308, 2006.

[2] M. E. Newman, "The structure and function of complex networks", *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[3] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks", *Nature*, vol. 411, no. 6833, p. 41, 2001.

[4] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, *et al.*, "Using graph theory to analyze biological networks", *BioData mining*, vol. 4, no. 1, p. 10, 2011.

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, *et al.*, "Gene ontology: Tool for the unification of biology", *Nature genetics*, vol. 25, no. 1, p. 25, 2000.

[6] N. K. Ahmed, J. Neville, and R. Kompella, "Network sampling: From static to streaming graphs", *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 8, no. 2, p. 7, 2014.

[7] L. Lu and T. Zhou, "Link prediction in complex networks: A survey", *Physica A: statistical mechanics and its applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[8] X. Li, M. Wu, C.-K. Kwoh, and S.-K. Ng, "Computational approaches for detecting protein complexes from protein interaction networks: A survey", *BMC genomics*, vol. 11, no. 1, S3, 2010.

[9] T. Milenkovic and N. Przulj, "Uncovering biological network function via graphlet degree signatures", *Cancer informatics*, vol. 6, CIN–S680, 2008.

[10] T. Milenkovic, V. Memisevic, A. K. Ganesan, and N. Przulj, "Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data", *Journal of the Royal Society Interface*, vol. 7, no. 44, pp. 423–437, 2010.

[11] C. Guerrero, T. Milenkovic, N. Przulj, P. Kaiser, and L. Huang, “Characterization of the proteasome interaction network using a qtax-based tag-team strategy and protein interaction network analysis”, *Proceedings of the National Academy of Sciences*, vol. 105, no. 36, pp. 13 333–13 338, 2008.

[12] K. H. Rosen, *Discrete mathematics and its applications.* New York: McGraw-Hill, 2011.

[13] K. Raman, “Construction and analysis of protein–protein interaction networks”, *Automated experimentation*, vol. 2, no. 1, p. 2, 2010.

[14] D. J. Watts and S. H. Strogatz, “Collective dynamics of small-world networks”, *nature*, vol. 393, no. 6684, p. 440, 1998.

[15] V. Colizza, A. Flammini, A. Maritan, and A. Vespignani, “Characterization and modeling of protein–protein interaction networks”, *Physica A: Statistical Mechanics and its Applications*, vol. 352, no. 1, pp. 1–27, 2005.

[16] A.-L. Barabasi and Z. N. Oltvai, “Network biology: Understanding the cell’s functional organization”, *Nature reviews genetics*, vol. 5, no. 2, p. 101, 2004.

[17] J. Dong and S. Horvath, “Understanding network concepts in modules”, *BMC systems biology*, vol. 1, no. 1, p. 24, 2007.

[18] A. Platzer, P. Perco, A. Lukas, and B. Mayer, “Characterization of protein-interaction networks in tumors”, *BMC bioinformatics*, vol. 8, no. 1, p. 224, 2007.

[19] L. A. Mueller, K. G. Kugler, A. Dander, A. Graber, and M. Dehmer, “Quacn: An r package for analyzing complex biological networks quantitatively”, *Bioinformatics*, vol. 27, no. 1, pp. 140–141, 2010.

[20] J. Kim and T. Wilhelm, “What is a complex graph?”, *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 11, pp. 2637–2652, 2008.

[21] H. Wiener, “Structural determination of paraffin boiling points”, *Journal of the American Chemical Society*, vol. 69, no. 1, pp. 17–20, 1947.

[22] S. Yamaguchi, “A note on wiener index”, *MATCH Commun. Math. Comput. Chem*, vol. 60, pp. 645–648, 2008.

[23] N. Przulj, D. G. Corneil, and I. Jurisica, “Modeling interactome: Scale-free or geometric?”, *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.

[24] N. Pzulj, “Biological network comparison using graphlet degree distribution”, *Bioinformatics*, vol. 23, no. 2, e177–e183, 2007.

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

[26] N. Przulj, D. G. Corneil, and I. Jurisica, “Efficient estimation of graphlet frequency distributions in protein–protein interaction networks”, *Bioinformatics*, vol. 22, no. 8, pp. 974–980, 2006.

[27] A. Masoudi-Nejad, F. Schreiber, and Z. R. M. Kashani, “Building blocks of biological networks: A review on major network motif discovery algorithms”, *IET systems biology*, vol. 6, no. 5, pp. 164–174, 2012.

[28] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, *et al.*, “Superfamilies of evolved and designed networks”, *Science*, vol. 303, no. 5663, pp. 1538–1542, 2004.

[29] M. E. Newman, “Power laws, pareto distributions and zipf’s law”, *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[30] A. Clauset, C. R. Shalizi, and M. E. Newman, “Power-law distributions in empirical data”, *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[31] B. H. Junker and F. Schreiber, *Analysis of biological networks.* John Wiley and Sons, 2011, vol. 2.

[32] R. Albert and A.-L. Barabasi, “Statistical mechanics of complex networks”, *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[33] A. Goldenberg, A. X. Zheng, S. E. Fienberg, E. M. Airoldi, *et al.*, “A survey of statistical network models”, *Foundations and Trends® in Machine Learning*, vol. 2, no. 2, pp. 129–233, 2010.

[34] Y. Huang, I. M. Tienda-Luna, and Y. Wang, “A survey of statistical models for reverse engineering gene regulatory networks”, *IEEE Signal Processing Magazine*, vol. 26, no. 1, p. 76, 2009.

[35] F. Crick, “Central dogma of molecular biology”, *Nature*, vol. 227, no. 5258, p. 561, 1970.

[36] L. A. Allison, *Fundamental molecular biology.* Blackwell Pub., 2007, ISBN: 1405103795.

[37] M. Bizzarri, A. Palombo, and A. Cucina, “Theoretical aspects of systems biology”, *Progress in Biophysics and Molecular Biology*, vol. 112, no. 1, pp. 33–43, 2013, ISSN: 0079-6107.

[38] B. Bose, “Systems biology: A biologist’s viewpoint”, *Progress in Biophysics and Molecular Biology*, vol. 113, no. 3, pp. 358–368, 2013, ISSN: 0079-6107.

[39] H. Kitano, “Computational systems biology”, *Nature*, vol. 420, no. 6912, pp. 206–210, 2002, ISSN: 0028-0836.

[40] F. J. Bruggeman and H. V. Westerhoff, “The nature of systems biology”, *Trends in Microbiology*, vol. 15, no. 1, pp. 45–50, 2007, ISSN: 0966-842X.

[41] A. .-.-. L. Barabasi and R. A, “Emergence of scaling in random networks”, *Science*, vol. 286, 1999.

[42] A. Rzhetsky and S. M. Gomez, "Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome", *Bioinformatics*, vol. 17, no. 10, pp. 988–996, 2001.

[43] P. Crucitti, V. Latora, M. Marchiori, and A. Rapisarda, "Error and attack tolerance of complex networks", *Physica A: Statistical Mechanics and its Applications*, vol. 340, no. 1, pp. 388–394, 2004, ISSN: 0378-4371.

[44] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis.", *Statistical applications in genetics and molecular biology*, vol. 4, Article17, 2005.

[45] R. Albert and A. L. Barabasi, "Topology of evolving networks: Local events and universality", *Phys Rev Lett*, vol. 85, no. 24, pp. 5234–7, 2000.

[46] P. Csermely, T. Korcsmáros, H. J. Kiss, G. London, and R. Nussinov, "Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review", *Pharmacology and Therapeutics*, vol. 138, no. 3, pp. 333–408, 2013, ISSN: 0163-7258.

[47] H. Kitano, "Biological robustness", *Nature Reviews Genetics*, vol. 5, no. 11, pp. 826–837, 2004, ISSN: 1471-0056.

[48] ——, "Cancer as a robust system: Implications for anticancer therapy", *Nature Reviews Cancer*, vol. 4, no. 3, pp. 227–235, 2004, ISSN: 1474-175X.

[49] S. Ciliberti, O. C. Martin, and A. Wagner, "Robustness can evolve gradually in complex regulatory gene networks with varying topology", *PLOS Computational Biology*, vol. 3, no. 2, pp. 1–10, 2007.

[50] C. Luni, J. Shoemaker, K. Sanft, L. Petzold, and F. Doyle, "Confidence from uncertainty-a multi-target drug screening method from robust control theory", *BMC systems biology*, vol. 4, no. 1, p. 161, 2010, ISSN: 1752-0509.

[51] A. E. M. Sean P. Cornelius William L. Kath, "Controlling complex networks with compensatory perturbations", *ArXiv e-prints*, 2011.

[52] I. J. Farkas, T. Korcsmaros, I. A. Kovacs, A. Mihalik, R. Palotai, *et al.*, "Network-based tools for the identification of novel drug targets", *Sci Signal*, vol. 4, no. 173, pt3, 2011.

[53] J. De Las Rivas and C. Fontanillo, "Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks", *PLOS Computational Biology*, vol. 6, no. 6, pp. 1–8, 2010.

[54] B. A. Shoemaker and A. R. Panchenko, "Deciphering protein–protein interactions. part i. experimental techniques and databases", *PLoS Computational Biology*, vol. 3, pp. 1727–1736, 2007.

[55] G. A. Pavlopoulos, A.-L. Wegener, and R. Schneider, "A survey of visualization tools for biological network analysis", *BioData Mining*, vol. 1, no. 1, p. 12, 2008, ISSN: 1756-0381.

[56] W. Daniel Duanqing and H. Xiaohua, *Mining and analyzing the topological structure of protein-protein interaction networks*, Conference Paper, 2006.

[57] R. Tanaka, T.-M. Yi, and J. Doyle, "Some protein interaction data do not exhibit power law statistics", *FEBS Letters*, vol. 579, no. 23, pp. 5140–5144, 2005, ISSN: 0014-5793.

[58] V. S. Rao, K. Srinivas, G. N. Sujini, and G. N. S. Kumar, "Protein-protein interaction detection: Methods and analysis", *International Journal of Proteomics*, vol. 2014, p. 12, 2014.

[59] B. Suter, X. Zhang, C. G. Pesce, A. Mendelsohn, S. P. Dinesh-Kumar, *et al.*, "Next-generation sequencing for binary protein–protein interactions", *Frontiers in genetics*, vol. 6, p. 346, 2015.

[60] B. A. Shoemaker and A. R. Panchenko, "Deciphering protein–protein interactions. part ii. computational methods to predict protein and domain interaction partners", *PLoS Computational Biology*, vol. 3, R35–1221, 2007.

[61] F. Pazos and A. Valencia, "Protein co-evolution, co-adaptation and interactions", *The EMBO Journal*, vol. 27, no. 20, pp. 2648–2655,

[62] S. Mika and B. Rost, "Protein–protein interactions more conserved within species than across species", *PLoS Computational Biology*, vol. 2, pp. 245–246, 2006.

[63] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, *et al.*, "The string database in 2011: Functional interaction networks of proteins, globally integrated and scored", *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D561–D568,

[64] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, *et al.*, "Biomedical text mining and its applications in cancer research", *Journal of Biomedical Informatics*, vol. 46, no. 2, pp. 200–211, 2013, ISSN: 1532-0464.

[65] T. U. Consortium, "Uniprot: A hub for protein information", *Nucleic Acids Research*, vol. 43, no. D1, pp. D204–D212, 2015. DOI: `10.1093/nar/gku989`.

[66] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, *et al.*, "String v10: Protein–protein interaction networks, integrated over the tree of life", *Nucleic acids research*, gku1003, 2014, ISSN: 0305-1048.

[67] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, *et al.*, "Protein interaction data curation - the international molecular exchange consortium (imex)", *Nature methods*, vol. 9, no. 4, pp. 345–350, 2012. DOI: `10.1038/nmeth.1931`.

[68] G. Rustici, N. Kolesnikov, M. Brandizi, T. Burdett, M. Dylag, *et al.*, "Arrayexpress update—trends in database growth and links to data analysis tools", *Nucleic Acids Research*, vol. 41, no. D1, pp. D987–D990, 2013.

[69] I. Papatheodorou, N. A. Fonseca, M. Keays, Y. A. Tang, E. Barrera, *et al.*, "Expression atlas: Gene and protein expression across multiple studies and organisms", *Nucleic Acids Research*, vol. 46, no. D1, pp. D246–D251, 2018.

[70] F. C. Barbosa, J. P. Arrais, and J. L. Oliveira, "Weighted gene co-expression network analysis applied to head and neck squamous cell carcinoma data", in *The International Conference on Health Informatics: ICHI 2013, Vilamoura, Portugal on 7-9 November, 2013*, Y.-T. Zhang, Ed. Cham: Springer International Publishing, 2014, pp. 300–303, ISBN: 978-3-319-03005-0.

[71] A. L. Tarca, R. Romero, and S. Draghici, "Analysis of microarray experiments of gene expression profiling", *American journal of obstetrics and gynecology*, vol. 195, no. 2, pp. 373–388, 2006, ISSN: 0002-9378 1097-6868.

[72] H. Buermans and J. Den Dunnen, "Next generation sequencing technology: Advances and applications", *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1842, no. 10, pp. 1932–1941, 2014, ISSN: 0925-4439.

[73] J. Hardin, A. Mitani, L. Hicks, and B. VanKoten, "A robust measure of correlation between two genes on a microarray", *BMC Bioinformatics*, vol. 8, no. 1, p. 220, 2007, ISSN: 1471-2105.

[74] L. Song, P. Langfelder, and S. Horvath, "Comparison of co-expression measures: Mutual information, correlation, and model based indices", *BMC Bioinformatics*, vol. 13, no. 1, p. 328, 2012, ISSN: 1471-2105.

[75] O. S. Soyer, M. Salathe, and S. Bonhoeffer, "Signal transduction networks: Topology, response and biochemical processes", *Journal of Theoretical Biology*, vol. 238, no. 2, pp. 416–425, 2006, ISSN: 0022-5193.

[76] M. Krull, S. Pistor, N. Voss, A. E. Kel, I. Reuter, *et al.*, "Transpath®: An information resource for storing and visualizing signaling pathways and their pathological aberrations", *Nucleic Acids Research*, vol. 34, pp. D546–D551, 2006.

[77] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, *et al.*, "Data, information, knowledge and principle: Back to metabolism in kegg", *Nucleic acids research*, vol. 42, no. D1, pp. D199–D205, 2014, ISSN: 0305-1048.

[78] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Kegg as a reference resource for gene and protein annotation", *Nucleic acids research*, vol. 44 D1, pp. D457–62, 2016.

[79] A. P. Trape and A. M. Gonzalez-Angulo, "Breast cancer and metastasis: On the way toward individualized therapy.", *Cancer genomics and proteomics*, vol. 9, no. 5, pp. 297–310, 2012.

[80] L. B. Edelman, J. A. Eddy, and N. D. Price, “In silico models of cancer”, *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 2, no. 4, pp. 438–459, DOI: 10.1002/wsbm.75.

[81] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, *et al.*, “The human disease network”, *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 2007, ISSN: 0027-8424. DOI: 10.1073/pnas.0701361104.

[82] M. Vidal, M. E. Cusick, and A.-L. Barabasi, “Interactome networks and human disease”, *Cell*, vol. 144, no. 6, pp. 986–998, 2011, ISSN: 0092-8674.

[83] X. F. Liu, C. K. Tse, and M. Small, “Complex network structure of musical compositions: Algorithmic generation of appealing music”, *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 1, pp. 126–132, 2010, ISSN: 0378-4371.

[84] P. Langfelder and S. Horvath, “Eigengene networks for studying the relationships between co-expression modules”, *BMC Systems Biology*, vol. 1, no. 1, p. 54, 2007, ISSN: 1752-0509.

[85] M. C. Oldham, S. Horvath, and D. H. Geschwind, “Conservation and evolution of gene coexpression networks in human and chimpanzee brains.”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103 47, pp. 17 973–8, 2006.

[86] M. E. Newman, “Random graphs as models of networks”, *arXiv preprint cond-mat/0202208*, 2002.

[87] S. Nagi, D. K. Bhattacharyya, and J. K. Kalita, “Gene expression data clustering analysis: A survey”, in *Emerging Trends and Applications in Computer Science (NCETACS), 2011 2nd National Conference on*, IEEE, pp. 1–12, ISBN: 1424495784.

[88] F. C. Barbosa, J. P. Arrais, and J. L. Oliveira, “Quantitative characterization of protein networks of the oral cavity”, in *7th International Conference on Practical Applications of Computational Biology and Bioinformatics*, S. M. Mohamad, L. Nanni, P. M. Rocha, and F. Fdez-Riverola, Eds., ser. Advances in Intelligent Systems and Computing. Heidelberg: Springer International Publishing, 2013, vol. 222, pp. 61–68, ISBN: 978-3-319-00578-2. DOI: 10.1007/978-3-319-00578-2_9.

[89] J. P. Arrais, N. Rosa, J. Melo, E. D. Coelho, D. Amaral, *et al.*, “Oralcard: A bioinformatic tool for the study of oral proteome”, *Archives of Oral Biology*, vol. 58, no. 7, pp. 762–772, 2013, ISSN: 0003-9969.

[90] STRING, *Search tool for the retrieval of interacting genes/proteins*, Web Page. [Online]. Available: http://string-db.org/.

[91] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer, and M. Albrecht, “Computing topological parameters of biological networks”, *Bioinformatics*, vol. 24, no. 2, pp. 282–284, 2008. DOI: 10.1093/bioinformatics/btm554.

[92] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, *et al.*, “Cytoscape: A software environment for integrated models of biomolecular interaction networks”, *Genome Res*, vol. 13, 2003.

[93] A. Ghazalpour, S. Doss, B. Zhang, S. Wang, C. Plaisier, *et al.*, “Integrating genetic and network analysis to characterize genes related to mouse weight”, *PLOS Genetics*, vol. 2, no. 8, pp. 1–11, 2006. DOI: 10.1371/journal.pgen.0020130.

[94] S. M. Rothenberg and L. W. Ellisen, “The molecular pathogenesis of head and neck squamous cell carcinoma”, *The Journal of Clinical Investigation*, vol. 122, no. 6, pp. 1951–1957, 2012, ISSN: 0021-9738. DOI: 10.1172/JCI59889.

[95] M. Hashibe, P. Brennan, S.-c. Chuang, S. Boccia, X. Castellsague, *et al.*, “Interaction between tobacco and alcohol use and the risk of head and neck cancer: Pooled analysis in the international head and neck cancer epidemiology consortium”, *Cancer Epidemiology and Prevention Biomarkers*, vol. 18, no. 2, pp. 541–550, 2009, ISSN: 1055-9965. DOI: 10.1158/1055-9965.EPI-08-0347.

[96] C. Ragin, F. Modugno, and S. Gollin, “The epidemiology and risk factors of head and neck cancer: A focus on human papillomavirus”, *Journal of Dental Research*, vol. 86, no. 2, pp. 104–114, 2007. DOI: 10.1177/154405910708600202.

[97] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, *et al.*, “Arrayexpress—a public repository for microarray gene expression data at the ebi”, *Nucleic acids research*, vol. 31, no. 1, pp. 68–71, 2003, ISSN: 0305-1048.

[98] E.-E. T. E. B. Institute, *Arrayexpress – functional genomics data*, Web Page. [Online]. Available: http://www.ebi.ac.uk/arrayexpress/.

[99] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using david bioinformatics resources”, *Nature protocols*, vol. 4, no. 1, pp. 44–57, 2008, ISSN: 1754-2189.

[100] N. I. of Allergy and N. Infectious Diseases (NIAID), *David bioinformatics resources 6.7*, Web Page. [Online]. Available: https://david.ncifcrf.gov/.

[101] S. L. Carter, C. M. Brechbuhler, M. Griffin, and A. T. Bond, “Gene co-expression network topology provides a framework for molecular characterization of cellular state”, *Bioinformatics*, vol. 20, no. 14, pp. 2242–2250, 2004, ISSN: 1367-4803.

[102] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi, “The large-scale organization of metabolic networks”, *Nature*, vol. 407, no. 6804, pp. 651–4, 2000. DOI: 10.1038/35036627.

[103] S. Bergmann, J. Ihmels, and N. Barkai, “Similarities and differences in genome-wide expression data of six organisms”, *PLoS Biol*, vol. 2, no. 1, E9, 2004. DOI: 10.1371/journal.pbio.0020009.

[104] E. Rava sz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabasi, "Hierarchical organization of modularity in metabolic networks", *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002. DOI: 10.1126/science.1073374.

[105] P. Langfelder, R. Luo, M. C. Oldham, and S. Horvath, "Is my network module preserved and reproducible?", *PLOS Computational Biology*, vol. 7, no. 1, pp. 1–29, 2011. DOI: 10.1371/journal.pcbi.1001057.

[106] P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: From co-expression clustering to reverse engineering", *Bioinformatics*, vol. 16, no. 8, pp. 707–726, 2000. DOI: 10.1093/bioinformatics/16.8.707.

[107] A. M. Yip and S. Horvath, "Gene network interconnectedness and the generalized topological overlap measure", *BMC Bioinformatics*, vol. 8, pp. 22–22, 2006.

[108] M. W. Gonzalez and M. G. Kann, "Chapter 4: Protein interactions and disease", *PLOS Computational Biology*, vol. 8, no. 12, pp. 1–11, 2012. DOI: 10.1371/journal.pcbi.1002819.

[109] S. Oliver, "Guilt-by-association goes global", *Nature*, vol. 403, no. 6770, pp. 601–3, 2000. DOI: 10.1038/35001165.

[110] F. Emmert-Streib, G. Glazko, A. Gokmen, and R. De Matos Simoes, "Statistical inference and reverse engineering of gene regulatory networks from observational expression data", *Frontiers in Genetics*, vol. 3, p. 8, 2012, ISSN: 1664-8021. DOI: 10.3389/fgene.2012.00008.

[111] L. Lovasz, "Very large graphs", *arXiv preprint arXiv:0902.0132*, 2009.

[112] ——, *Large networks and graph limits.* AMS Bookstore, 2012, vol. 60, ISBN: 0821890859.

[113] G. Jurman, R. Visintainer, S. Riccadonna, M. Filosi, and C. Furlanello, "The him glocal metric and kernel for network comparison and classification", *arXiv preprint arXiv:1201.2931*, 2012.

[114] M. S. Amin, R. L. Finley Jr, and H. M. Jamil, "Top-k similar graph matching using tram in biological networks", *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 9, no. 6, pp. 1790–1804, 2012, ISSN: 1545-5963.

[115] P. Lee, L. V. S. Lakshmanan, and J. X. Yu, "On top-k structural similarity search", *2012 IEEE 28th International Conference on Data Engineering*, pp. 774–785, 2012.

[116] S. Zhang, X.-S. Zhang, and L. Chen, "Biomolecular network querying: A promising approach in systems biology", *BMC systems biology*, vol. 2, no. 1, p. 5, 2008, ISSN: 1752-0509.

[117] P. Holme and J. Saramaki, "Temporal networks", *Physics Reports*, vol. 519, no. 3, pp. 97–125, 2012, ISSN: 0370-1573.

[118] F. Ay, M. Kellis, and T. Kahveci, “Submap: Aligning metabolic pathways with subnetwork mappings”, *Journal of Computational Biology*, vol. 18, no. 3, pp. 219–235, 2011, ISSN: 1066-5277.

[119] F. Ay, M. Dang, and T. Kahveci, “Metabolic network alignment in large scale by network compression”, *BMC Bioinformatics*, vol. 13, no. Suppl 3, S2, 2012, ISSN: 1471-2105.

[120] M. Berlingerio, D. Koutra, T. Eliassi-Rad, and C. Faloutsos, “Netsimile: A scalable approach to size-independent network similarity”, *arXiv preprint arXiv:1209.2684*, 2012.

[121] G. Guelsoy, B. Gandhi, and T. Kahveci, “Topac: Alignment of gene regulatory networks using topology-aware coloring”, *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 01, 2012, ISSN: 0219-7200.

[122] Y.-X. Zhu, L. Lu, Q.-M. Zhang, and T. Zhou, “Uncovering missing links with cold ends”, *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 22, pp. 5769–5778, 2012, ISSN: 0378-4371.

[123] X. Luo, Z. Ming, Z. You, S. Li, Y. Xia, *et al.*, “Improving network topology-based protein interactome mapping via collaborative filtering”, *Knowledge-Based Systems*, vol. 90, pp. 23–32, 2015, ISSN: 0950-7051.

[124] H. N. Chua and L. Wong, “Increasing the reliability of protein interactomes”, *Drug Discov Today*, vol. 13, no. 15-16, pp. 652–8, 2008. DOI: 10.1016/j.drudis.2008.05.004.

[125] M. Varjosalo, R. Sacco, A. Stukalov, A. van Drogen, M. Planyavsky, *et al.*, “Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized ap-ms”, *Nat Meth*, vol. 10, no. 4, pp. 307–314, 2013, ISSN: 1548-7091.

[126] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, *et al.*, “Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or interologs”, *Genome Res*, vol. 11, no. 12, pp. 2120–6, 2001. DOI: 10.1101/gr.205301.

[127] M. Tarailo, S. Tarailo, and A. M. Rose, “Synthetic lethal interactions identify phenotypic “interologs” of the spindle assembly checkpoint components”, *Genetics*, vol. 177, no. 4, pp. 2525–2530, 2007, ISSN: 0016-6731. DOI: 10.1534/genetics.107.080408.

[128] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, “A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae)”, *Proceedings of the National Academy of Sciences*, vol. 100, no. 14, pp. 8348–8353, 2003. DOI: 10.1073/pnas.0832373100.

[129] A. Patil and H. Nakamura, "Filtering high-throughput protein-protein interaction data using a combination of genomic features", *BMC Bioinformatics*, vol. 6, no. 1, p. 100, 2005, ISSN: 1471-2105. DOI: 10.1186/1471-2105-6-100.

[130] G. Liu, J. Li, and L. Wong, "Assessing and predicting protein interactions using both local and global network topological metrics", *Genome Inform*, vol. 21, pp. 138–49, 2008.

[131] N. Skunca, A. Altenhoff, and C. Dessimoz, "Quality of computationally inferred gene ontology annotations", *PLOS Computational Biology*, vol. 8, no. 5, e1002533, 2012. DOI: 10.1371/journal.pcbi.1002533.

[132] J. Dutkowski, M. Kramer, M. A. Surma, R. Balakrishnan, J. M. Cherry, *et al.*, "A gene ontology inferred from molecular networks", *Nature biotechnology*, vol. 31, no. 1, 10.1038/nbt.2463, 2013, ISSN: 1087-0156 1546-1696. DOI: 10.1038/nbt.2463.

[133] R. Saito, H. Suzuki, and Y. Hayashizaki, "Interaction generality, a measurement to assess the reliability of a protein-protein interaction", *Nucleic Acids Res*, vol. 30, no. 5, pp. 1163–8, 2002, ISSN: 0305-1048.

[134] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, *et al.*, "Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network", *Genome Biology*, vol. 5, no. 1, R6, 2004, ISSN: 1465-6906 1465-6914.

[135] J. Chen, W. Hsu, M. L. Lee, and S. K. Ng, "Increasing confidence of protein interactomes using network topological metrics", *Bioinformatics*, vol. 22, no. 16, pp. 1998–2004, 2006, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl335.

[136] H. N. Chua, W. K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions", *Bioinformatics*, vol. 22, no. 13, pp. 1623–30, 2006. DOI: 10.1093/bioinformatics/btl145.

[137] T. U. Consortium, "Activities at the universal protein resource (uniprot)", *Nucleic Acids Research*, vol. 42, no. D1, pp. D191–D198, 2014. DOI: 10.1093/nar/gkt1140.

[138] L. Lu, L. Pan, T. Zhou, Y.-C. Zhang, and H. E. Stanley, "Toward link predictability of complex networks", *Proceedings of the National Academy of Sciences*, vol. 112, no. 8, pp. 2325–2330, 2015. DOI: 10.1073/pnas.1424644112.

[139] Y. Fang, M. Sun, G. Dai, and K. Ramain, "The intrinsic geometric structure of protein-protein interaction networks for protein interaction prediction", *IEEE/ACM Trans Comput Biol Bioinform*, vol. 13, no. 1, pp. 76–85, 2016, ISSN: 1545-5963. DOI: 10.1109/tcbb.2015.2456876.

[140] S. R. Collins, P. Kemmeren, X. C. Zhao, J. F. Greenblatt, F. Spencer, *et al.*, “Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae”, *Mol Cell Proteomics*, vol. 6, no. 3, pp. 439–50, 2007. DOI: 10.1074/mcp.M600381-MCP200.

[141] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, *et al.*, “The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations”, *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003, ISSN: 1432-0762. DOI: 10.1007/s00265-003-0651-y.

[142] V. E. Krebs, “Mapping networks of terrorist cells”, *Connections*, vol. 24, no. 3, pp. 43–52, 2002, ISSN: 0226-1766.

[143] O. Kuchaiev, M. Rasajski, D. J. Higham, and N. Przulj, “Geometric de-noising of protein-protein interaction networks”, *PLOS Computational Biology*, vol. 5, no. 8, e1000454, 2009. DOI: 10.1371/journal.pcbi.1000454.

[144] A. Clauset, C. Moore, and M. Newman, “Hierarchical structure and the prediction of missing links in networks”, *Nature*, vol. 453, pp. 98–101, 2008.

[145] J. I. Fuxman Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, *et al.*, “Using networks to measure similarity between genes: Association index selection”, *Nature methods*, vol. 10, no. 12, pp. 1169–1176, 2013, ISSN: 1548-7091 1548-7105. DOI: 10.1038/nmeth.2728.

[146] L. Lu, C.-H. Jin, and T. Zhou, “Similarity index based on local paths for link prediction of complex networks”, *Phys. Rev. E*, vol. 80, p. 046 122, 4 2009. DOI: 10.1103/PhysRevE.80.046122.

[147] S. Daminelli, J. M. Thomas, C. Duran, and C. V. Cannistraci, “Common neighbours and the local-community-paradigm for topological link prediction in bipartite networks”, *New Journal of Physics*, vol. 17, no. 11, p. 113 037, 2015.

[148] M.-W. Ahn and W.-S. Jung, “Accuracy test for link prediction in terms of similarity index: The case of ws and ba models”, *Physica A: Statistical Mechanics and its Applications*, vol. 429, pp. 177–183, 2015, ISSN: 0378-4371.

[149] D. Hao, C. Ren, and C. Li, “Revisiting the variation of clustering coefficient of biological networks suggests new modular structure”, in *BMC Systems Biology*, vol. 6, 2012, p. 34.

[150] A. C. Lewis, N. S. Jones, M. A. Porter, and C. M. Deane, “The function of communities in protein interaction networks at multiple scales”, *BMC Systems Biology*, vol. 4, no. 1, p. 100, 2010, ISSN: 1752-0509. DOI: 10.1186/1752-0509-4-100.

[151] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, “From molecular to modular cell biology”, *Nature*, vol. 402, no. 6761, p. C47, 1999, ISSN: 0028-0836.

[152] R. Sharan, I. Ulitsky, and R. Shamir, “Network-based prediction of protein function”, *Molecular Systems Biology*, vol. 3, no. 1, 2007. DOI: 10.1038/msb4100129.

[153] J.-D. J. Han, “Understanding biological functions through molecular networks”, *Cell research*, vol. 18, no. 2, p. 224, 2008, ISSN: 1001-0602.

[154] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, *et al.*, “String v10: Protein-protein interaction networks, integrated over the tree of life”, *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D447–52, 2015, ISSN: 0305-1048. DOI: 10.1093/nar/gku1003.

[155] B. Taboada, C. Verde, and E. Merino, “High accuracy operon prediction method based on string database scores”, *Nucleic Acids Research*, vol. 38, no. 12, e130, 2010, ISSN: 0305-1048 1362-4962. DOI: 10.1093/nar/gkq254.

[156] M. Ye, G. C. Racz, Q. Jiang, X. Zhang, and B. M. E. Moret, “Nemo: An evolutionary model with modularity for ppi networks”, in *Bioinformatics Research and Applications: 12th International Symposium, ISBRA 2016, Minsk, Belarus, June 5-8, 2016, Proceedings*, A. Bourgeois, P. Skums, X. Wan, and A. Zelikovsky, Eds. Cham: Springer International Publishing, 2016, pp. 224–236, ISBN: 978-3-319-38782-6. DOI: 10.1007/978-3-319-38782-6_19.

[157] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, “Omim.org: Online mendelian inheritance in man (omim), an online catalog of human genes and genetic disorders”, *Nucleic acids research*, vol. 43, no. D1, pp. D789–D798, 2015, ISSN: 0305-1048.

[158] A.-L. Barabasi, N. Gulbahce, and J. Loscalzo, “Network medicine: A network-based approach to human disease”, *Nature reviews. Genetics*, vol. 12, no. 1, pp. 56–68, 2011, ISSN: 1471-0056 1471-0064. DOI: 10.1038/nrg2918.

[159] J. P. Arrais and J. L. Oliveira, “Using biomedical networks to prioritize gene-disease associations”, *Open Access Bioinformatics*, vol. 1, pp. 123–130, 2011.

[160] P. F. Jonsson and P. A. Bates, “Global topological features of cancer proteins in the human interactome”, *Bioinformatics*, vol. 22, no. 18, pp. 2291–7, 2006, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl390.

[161] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, “Predicting disease genes using protein–protein interactions”, *Journal of medical genetics*, vol. 43, no. 8, pp. 691–698, 2006, ISSN: 1468-6244.

[162] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, *et al.*, “Uncovering disease-disease relationships through the incomplete interactome”, *Science*, vol. 347, no. 6224, p. 1 257 601, 2015, ISSN: 0036-8075.

[163] M. P. Pradhan, K. Nagulapalli, and M. J. Palakal, “Cliques for the identification of gene signatures for colorectal cancer across population”, *BMC systems biology*, vol. 6, no. Suppl 3, S17, 2012, ISSN: 1752-0509.

[164] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis", *Molecular systems biology*, vol. 3, no. 1, p. 140, 2007, ISSN: 1744-4292.

[165] M. Sonachalam, J. Shen, H. Huang, and X. Wu, "Systems biology approach to identify gene network signatures for colorectal cancer", *Frontiers in Genetics*, vol. 3, 2012, ISSN: 1664-8021. DOI: 10.3389/fgene.2012.00080.

[166] F. B. Correia, J. P. Arrais, and J. L. Oliveira, "Prediction of cancer using network topological features", in *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, ser. BIOSTEC 2016, Rome, Italy: SCITEPRESS - Science and Technology Publications, Lda, 2016, pp. 207–215, ISBN: 978-989-758-170-0. DOI: 10.5220/0005696202070215.

[167] I. Inza, P. Larranaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in dna microarray domains", *Artificial Intelligence in Medicine*, vol. 31, no. 2, pp. 91–103, 2004, ISSN: 0933-3657.

[168] R. G. Ramani and S. G. Jacob, "Improved classification of lung cancer tumors based on structural and physicochemical properties of proteins using data mining models", *PLoS ONE*, vol. 8, no. 3, e58772, 2013. DOI: 10.1371/journal.pone.0058772.

[169] M. Dezfuly and H. Sajedi, "Predict survival of patients with lung cancer using an ensemble feature selection algotithm and classification methods in data mining", *Journal of Information*, vol. 1, no. 1, pp. 1–11, 2015.

[170] A. Ay, D. Gong, and T. Kahveci, "Network-based prediction of cancer under genetic storm", *Cancer informatics*, vol. 13, no. Suppl 3, p. 15, 2014.

[171] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics", *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007. DOI: 10.1093/bioinformatics/btm344.

[172] V. Bolon-Canedo, N. Sanchez-Marono, A. Alonso-Betanzos, J. M. Benitez, and F. Herrera, "A review of microarray datasets and applied feature selection methods", *Information Sciences*, vol. 282, pp. 111–135, 2014, ISSN: 0020-0255.

[173] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data", in *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on.* DOI: 10.1109/IRI.2012.6303031.

[174] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques", in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, pp. 313–325, ISBN: 3540874801.

[175] I. Kononenko, “Estimating attributes: Analysis and extensions of relief”, in *Machine Learning: ECML-94*, Springer, pp. 171–182, ISBN: 3540578684.

[176] M. Robnik-Sikonja and I. Kononenko, “Theoretical and empirical analysis of relieff and rrelieff”, *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003, ISSN: 0885-6125. DOI: 10.1023/A:1025667309714.

[177] S. G. Nancy and S. A. alias Balamurugan, “A comparative study of feature selection methods for cancer classification using gene expression dataset”, *Journal of Computer Applications (JCA)*, vol. 6, no. 3, p. 2013, 2013, ISSN: 0974-1925.

[178] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, “Learning with local and global consistency”, *Advances in neural information processing systems*, vol. 16, no. 16, pp. 321–328, 2004.

[179] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics”, *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, 2015, ISSN: 1471-0056.

[180] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction”, *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015, ISSN: 2001-0370.

[181] R. Barter, S.-J. Schramm, G. Mann, and Y. H. Yang, “Network-based biomarkers enhance classical approaches to prognostic gene expression signatures”, *BMC Systems Biology*, vol. 8, no. Suppl 4, S5, 2014, ISSN: 1752-0509.

[182] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, *et al.*, “Support vector machine classification and validation of cancer tissue samples using microarray expression data”, *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000, ISSN: 1367-4803.

[183] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers”, in *Proceedings of the fifth annual workshop on Computational learning theory*, ACM, pp. 144–152, ISBN: 089791497X.

[184] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Scholkopf, and G. Ratsch, “Support vector machines and kernels for computational biology”, *PLoS Comput Biol*, vol. 4, no. 10, e1000173, 2008, ISSN: 1553-7358.

[185] W. S. Noble, “Support vector machine applications in computational biology”, *Kernel methods in computational biology*, vol. 71, p. 92, 2004.

[186] V. Vapnik and S. Mukherjee, “Support vector method for multivariate density estimation”, in *Advances in neural information processing systems*, pp. 659–665.

[187] W. S. Noble, “What is a support vector machine?”, *Nature biotechnology*, vol. 24, no. 12, p. 1565, 2006, ISSN: 1546-1696.

[188] T. Hofmann, B. Scholkopf, and A. J. Smola, “Kernel methods in machine learning”, *The annals of statistics*, pp. 1171–1220, 2008, ISSN: 0090-5364.

[189] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring", *science*, vol. 286, no. 5439, pp. 531–537, 1999, ISSN: 0036-8075.

[190] E. Moler, M. Chow, and I. Mian, "Analysis of molecular profile data using generative and discriminative methods", *Physiological Genomics*, vol. 4, no. 2, pp. 109–126, 2000, ISSN: 1094-8341.

[191] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes", *BMC Medical Informatics and Decision Making*, vol. 10, p. 16, 2010, ISSN: 1472-6947. DOI: 10.1186/1472-6947-10-16.

[192] S. Tyanova, R. Albrechtsen, P. Kronqvist, J. Cox, M. Mann, *et al.*, "Proteomic maps of breast cancer subtypes", *Nature communications*, vol. 7, p. 10 259, 2016, ISSN: 2041-1723.

[193] N. Bhatia, "Survey of nearest neighbor techniques", *arXiv preprint arXiv:1007.0085*, 2010.

[194] B. Deekshatulu and P. Chandra, "Classification of heart disease using k-nearest neighbor and genetic algorithm", *Procedia Technology*, vol. 10, pp. 85–94, 2013, ISSN: 2212-0173.

[195] A. Liaw and M. Wiener, "Classification and regression by randomforest", *R News*, vol. 2, 2002.

[196] D. R. Cutler, T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, *et al.*, "Random forests for classification in ecology", *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007, ISSN: 1939-9170.

[197] J. Gall, N. Razavi, and L. Van Gool, "An introduction to random forests for multi-class object detection", in *Outdoor and large-scale real-world scene analysis*. Springer, 2012, pp. 243–263.

[198] R. Diaz-Uriarte and S. Alvarez de Andres, "Gene selection and classification of microarray data using random forest", *BMC Bioinformatics*, vol. 7, p. 3, 2006, ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-3.

[199] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification", *BMC Bioinformatics*, vol. 9, no. 1, p. 319, 2008, ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-319.

[200] X. .-.-. W. Chen and M. Liu, "Prediction of protein–protein interactions using random decision forest framework", *Bioinformatics*, vol. 21, 2005.

[201] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. Konig, “Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012, ISSN: 1942-4795. DOI: 10.1002/widm.1072.

[202] X. Chen and H. Ishwaran, “Random forests for genomic data analysis”, *Genomics*, vol. 99, no. 6, pp. 323–329, 2012, ISSN: 0888-7543.

[203] D. Chen and H. Yang, “Comparison of gene regulatory networks of benign and malignant breast cancer samples with normal samples”, *Genetics and molecular research: GMR*, vol. 13, no. 4, p. 9453, 2014, ISSN: 1676-5680.

[204] H. Yu, C.-C. Lin, Y.-Y. Li, and Z. Zhao, “Dynamic protein interaction modules in human hepatocellular carcinoma progression”, *BMC systems biology*, vol. 7, no. Suppl 5, S2, 2013, ISSN: 1752-0509.

[205] M. Dominietto, N. Tsinoremas, and E. Capobianco, “Integrative analysis of cancer imaging readouts by networks”, *Molecular oncology*, vol. 9, no. 1, pp. 1–16, 2015, ISSN: 1574-7891.

[206] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2015”, *CA: A Cancer Journal for Clinicians*, vol. 65, no. 1, pp. 5–29, 2015, ISSN: 1542-4863. DOI: 10.3322/caac.21254.

[207] R. A. Smith, K. S. Andrews, D. Brooks, S. A. Fedewa, D. Manassaram-Baptiste, *et al.*, “Cancer screening in the united states, 2018: A review of current american cancer society guidelines and current issues in cancer screening”, *CA: A Cancer Journal for Clinicians*, vol. 68, no. 4, pp. 297–316, 2018, ISSN: 1542-4863.

[208] J. Dennis G., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, *et al.*, “David: Database for annotation, visualization, and integrated discovery”, *Genome Biol*, vol. 4, no. 5, P3, 2003.

[209] D. A. Schult and P. Swart, “Exploring network structure, dynamics, and function using networkx”, in *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, vol. 2008, pp. 11–16.

[210] L. Mueller, K. Kugler, A. Graber, F. Emmert-Streib, and M. Dehmer, “Structural measures for network biology using quacn”, *BMC Bioinformatics*, vol. 12, no. 1, p. 492, 2011, ISSN: 1471-2105.

[211] P. Ribeiro and F. Silva, “G-tries: A data structure for storing and finding subgraphs”, *Data Mining and Knowledge Discovery*, vol. 28, no. 2, pp. 337–377, 2014, ISSN: 1384-5810.

[212] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks”, *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, 2009, ISSN: 0306-4573.