UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

FABIANE CRISTINE DILLENBURG

An Approach for Analyzing and Classifying Microarray Data Using Gene Co-expression Networks Cycles

Thesis presented in partial fulfillment
of the requirements for the degree of
Doctor of Computer Science

Advisor: Prof[a]. Dr[a]. Leila Ribeiro
Coadvisor: Prof. Dr. Luigi Carro

Porto Alegre
October 2017

CIP — CATALOGING-IN-PUBLICATION

Dillenburg, Fabiane Cristine

An Approach for Analyzing and Classifying Microarray Data Using Gene Co-expression Networks Cycles / Fabiane Cristine Dillenburg. – Porto Alegre: PPGC da UFRGS, 2017.

133 f.: il.

Thesis (Ph.D.) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2017. Advisor: Leila Ribeiro; Coadvisor: Luigi Carro.

1. Bioinformatics. 2. Gene co-expression networks. 3. Cycle. 4. Negative feedback. 5. Positive feedback. 6. Systems biology. 7. Microarrays. 8. Gene expression. 9. Gliobastoma multiforme. 10. GBM. I. Ribeiro, Leila. II. Carro, Luigi. III. Título.

To my family.

"Anybody that thought the genome was going to directly provide drugs was a fool. Biological networks are not simple, and making drugs to affect them won't be simple."

— LEROY HOOD

ACKNOWLEDGEMENTS

This work would not have been possible without the guidance of my advisor Leila Ribeiro and my coadvisor Luigi Carro. I am very grateful for their support. I would also like to show my gratitude to Alfeu Zanotto-Filho and José Cláudio Fonseca Moreira for helpful comments and discussions.

I would like to express my thanks to my friend Fábio Mierlo for providing access to a good machine so that I could perform several experiments (and for lending a power adapter to my notebook); and to my friends Renato Costa and Mírian Bruckschen Motta Barros for the text revisions. They have given me great incentive since the college. Thanks also to all my friends for the emotional support.

Special thanks to my parents, Ari and Isoldi, also to my brother Luis and his fiancee Diana, for their love and care, and support in difficult moments. Thanks to all my family that was always looking forward to seeing the success of this work.

Finally, I wish to thank God for making me able to carry out this work.

ABSTRACT

One of the main research areas in Systems Biology concerns the discovery of biological networks from microarray datasets. These networks consist of a great number of genes whose expression levels affect each other in various ways. We present a new way of analyzing microarray datasets, based on the different kind of cycles found among genes of the co-expression networks constructed using quantized data obtained from the microarrays. The input of the analysis method is formed by raw data, a set of interest genes (for example, genes from a known pathway) and a function (activator or inhibitor) of these genes. The output of the method is a set of cycles. A cycle is a closed walk, in which all vertices (except the first and last) are distinct. Thanks to the new way of finding relations among genes, a more robust interpretation of gene correlations is possible, because cycles are associated with feedback mechanisms that are very common in biological networks. Our hypothesis is that negative feedbacks allow finding relations among genes that may help explaining the stability of the regulatory process within the cell. Positive feedback cycles, on the other hand, may show the amount of imbalance of a certain cell in a given time. The cycle-based analysis allows identifying the stoichiometric relationship between the genes of the network. This methodology provides a better understanding of the biology of tumors. As a consequence, it may enable the development of more effective treatment therapies. Furthermore, cycles help differentiate, measure and explain the phenomena identified in healthy and diseased tissues. Cycles may also be used as a new method for classification of samples of a microarray (cancer diagnosis). Compared to other classification methods, cycle-based classification provides a richer explanation of the proposed classification, that can give hints on the possible therapies. Therefore, the main contributions of this thesis are: (i) a new cycle-based analysis method; (ii) a new microarray samples classification method; (iii) and, finally, application and achievement of practical results. We use the proposed methodology to analyze the genes of four networks closely related with cancer - apoptosis, glucolysis, cell cycle and NFκB - in tissues of the most aggressive type of brain tumor (Gliobastoma multiforme – GBM) and in healthy tissues. Because most patients with GBMs die in less than a year, and essentially no patient has long-term survival, these tumors have drawn significant attention. Our main results show that the stoichiometric relationship between genes involved in apoptosis, glucolysis, cell cycle and NFκB pathways is unbalanced in GBM samples *versus* control samples. This dysregulation can be measured and explained by the identification of a higher percentage of positive cycles in these networks. This conclusion helps to understand more about the biology of this tumor type. The proposed cycle-based classification method achieved the same performance metrics as a neural network, a classical classification method. However, our method has a significant advantage with respect to neural networks. The proposed classification method not only classifies samples, providing diagnosis, but also explains why samples were classified in a certain way in terms of the feedback mechanisms that are present/absent. This way, the method provides hints to biochemists about possible laboratory experiments, as well as on potential drug target genes.

Keywords: Bioinformatics. gene co-expression networks. cycle. negative feedback. positive feedback. systems biology. microarrays. gene expression. gliobastoma multiforme. GBM.

Uma Abordagem para Analisar e Classificar Dados de *Microarrays* Usando Ciclos de Redes de Co-expressão Gênica

RESUMO

Uma das principais áreas de pesquisa em Biologia de Sistemas refere-se à descoberta de redes biológicas a partir de conjuntos de dados de *microarrays*. Estas redes consistem de um grande número de genes cujos níveis de expressão afetam os outros genes de vários modos. Nesta tese, apresenta-se uma nova maneira de analisar os conjuntos de dados de *microarrays*, com base nos diferentes tipos de ciclos encontrados entre os genes das redes de co-expressão construídas com dados quantificados obtidos a partir dos *microarrays*. A entrada do método de análise é formada pelos dados brutos, um conjunto de genes de interesse (por exemplo, genes de uma via conhecida) e uma função (ativador ou inibidor) destes genes. A saída do método é um conjunto de ciclos. Um ciclo é um caminho fechado com todos os vértices (exceto o primeiro e o último) distintos. Graças à nova forma de encontrar relações entre os genes, é possível uma interpretação mais robusta das correlações dos genes, porque os ciclos estão associados a mecanismos de *feedback*, que são muito comuns em redes biológicas. A hipótese é que *feedbacks* negativos permitem encontrar relações entre os genes que podem ajudar a explicar a estabilidade do processo regulatório dentro da célula. Ciclos de *feedback* positivo, por outro lado, podem mostrar a quantidade de desequilíbrio de uma determinada célula em um determinado momento. A análise baseada em ciclos permite identificar a relação estequiométrica entre os genes da rede. Esta metodologia proporciona uma melhor compreensão da biologia do tumor. Como consequência, pode permitir o desenvolvimento de terapias de tratamento mais eficazes. Além disso, os ciclos ajudam a diferenciar, medir e explicar os fenômenos identificados em tecidos saudáveis e doentes. Os ciclos também podem ser usados como um novo método para a classificação de amostras de um *microarray* (diagnóstico de câncer). Em comparação com outros métodos de classificação, a classificação baseada em ciclos fornece uma explicação mais rica da classificação proposta, que pode dar pistas sobre as possíveis terapias. Portanto, as principais contribuições desta tese são: (i) um novo método de análise baseada em ciclos; (ii) um novo método de classificação; (iii) e, finalmente, aplicação dos métodos e a obtenção de resultados práticos. A metodologia proposta foi utilizada para analisar os genes de quatro redes fortemente relacionadas com o câncer - apoptose, glicólise, ciclo celular e NFκB - em tecidos do tipo mais agressivo de tumor cerebral (Gliobastoma multiforme - GBM) e em tecidos cerebrais saudáveis. A maioria dos pacientes com GBM morrem em menos de um ano, essencialmente nenhum paciente tem sobrevivência a longo prazo, por isso estes tumores têm atraído atenção significativa. Os principais resultados nesta tese mostram que a relação estequiométrica entre genes envolvidos na apoptose, glicólise, ciclo celular e NFκB está desequilibrada em amostras de GBM em comparação as amostras de controle. Este desequilíbrio pode ser medido e explicado pela identificação de um percentual maior de ciclos positivos nas redes das primeiras amostras. Esta conclusão ajuda a entender mais sobre a biologia deste tipo de tumor. O método de classificação baseado no ciclo proposto obteve as mesmas métricas de desempenho como uma rede neural, um método clássico de classificação. No entanto, o método proposto tem uma vantagem significativa em relação às redes neurais. O método de classificação proposto não só classifica as amostras, fornecendo diagnóstico, mas também explica porque as amostras foram classificadas de uma certa maneira em termos dos mecanismos de *feedback* que estão presentes/ausentes. Desta forma, o método fornece dicas para bioquímicos sobre possíveis experiências laboratoriais, bem como sobre potenciais genes alvo de terapias.

Palavras-chave: Bioinformática, Redes de Co-expressão Gênica, Ciclo, *Feedback* Negativo, *Feedback* Positivo, Biologia de Sistemas, *Microarrays*, Expressão Gênica, Apoptose, Glicólise, Ciclo Celular, NFκB, Gliobastoma multiforme, GBM, Análise, Classificação.

LIST OF ABBREVIATIONS AND ACRONYMS

AC	Accuracy
cDNA	Complementary DNA
DNA	Deoxyribonucleic Acid
CDK	Cyclin-dependent Kinase
GCN	Gene Co-expression Network
GRN	Gene Regulatory Networks
GBM	Glioblastoma multiforme
GTS	Graph Transformation System
mRNA	Messenger RNA
RNA	Ribonucleic Acid
SAGE	Serial Analysis of Gene Expression
RT-PCR	Reverse Transcription Polymerase Chain Reaction
ARACNE	Algorithm for the Reconstruction of Accurate Cellular Networks
CLR	Context Likelihood of Relatedness
MRNET	Maximum Relevance/Minimum Redundancy Network
OLS	Ordinary Least Squares
SE	Sensitivity
SP	Specificity
ODE	Ordinary Differential Equations

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Bioinformatics aims to treat biological data and answer questions related to the functioning of biological systems. In this sense, the construction of models to analyze and predict the behavior of biological systems is of extreme importance. These models can, along with laboratory experiments, reduce the number, cost and time of experiments needed to discover new information. Computational methods, besides enabling the manipulation of a large amount of biological data, also allow the inference of models that can quantify and explain the data produced by the biological system, enabling classification and analysis techniques.

Gene expression provides information for building models of biological systems. Gene expression analysis comparing normal and neoplastic tissues have been used to identify genes associated with tumor genesis and potential therapeutic targets (PARMIGIANI et al., 2003). Genomic high-throughput technologies, such as microarrays, may considerably facilitate the molecular profiling of human tumors. Thousands of genes can now be analyzed using a single microarray hybridization chip (STEKEL, 2003). The expression profile from a single tumor reflects the state of events of an individual malignancy at a certain time point. To generalize the findings and provide conclusive evidence for the involvement of a molecular alteration, it is often necessary to analyze several hundred tumors. Using traditional molecular pathology, such verification could take several months, or even years, to reach completion. To facilitate translational research in a large-scale manner, new techniques are needed.

1.1 Motivation

One of the main research areas in systems biology concerns the analyzes of biological pathways or networks discovered from microarray datasets. A gene co-expression network (GCN) can be inferred from microarrays experiments (DAS et al., 2009). A GCN is an undirected graph, where nodes correspond to genes and undirected edges between nodes represent significant co-expression relationships. In a GCN, two genes are connected by an undirected edge if their activities have significant association (usually quantified by correlation or mutual information) considering a series of gene expression measurements in many different conditions or at different times (BOCCALETTI; LATORA; MORENO, 2010). Compared to gene regulatory networks, a GCN does not attempt to represent direct causal relationships among the participating genes in the form of directed edges.

GCNs can be analyzed in several ways (SERIN et al., 2016; RUAN; DEAN; ZHANG, 2010), such as topology analysis (HSU; JUAN; HUANG, 2015; VIALANEIX et al., 2013; XULVI-BRUNET; LI, 2010), modularity analysis and gene ontology analysis (GAITERI et al., 2014; TEJERA; BERNARDES; REBELO, 2013; FULLER et al., 2007). GCNs can be used to address questions related to modules (clusters of interconnected nodes), for example, to summarize the node profiles of a given module by a representative (highly connected) hub node (LANGFELDER; HORVATH, 2008). These networks allow the simultaneous investigation of multiple gene co-expression patterns across a wide range of conditions; and the two main aims of co-expression network analysis are to find new genes involved in the biological process under investigation and to suggest the biological processes a gene is involved in (SERIN et al., 2016).

The aforementioned analysis techniques allow to inspect different aspects of GCN's structure. Although these approaches have led to many valuable insights, they tend to focus attention on a few genes (for instance, hub genes) and ignore many others that may be used to generate and translate systems biology insights into testable predictions. Thus, understanding the network remains a challenge: *how to model and explain biological process?* Also *how to extract relevant information from biological data (specifically, from microarrays)?*

1.2 Objectives and Thesis Contributions

The main objective of our research is to define a new way of analyzing microarray datasets, based on the different kind of cycles found among genes of the GCN constructed using quantized data obtained from microarrays. A cycle is a closed walk with all vertices being distinct (except the first and last ones) (DIESTEL, 2012). We follow the hypothesis that cycles are associated with feedback mechanisms very common in biological networks. Considering this objective and hypothesis, our contributions are the following:

- We propose a new way of finding relations among genes, that allows a different interpretation of gene correlations. The cycle-based analysis allows one to identify the stoichiometric relationship between the genes of the network providing, for example, a better understanding of tumor biology. Thus, cycle analysis may significantly contribute to the development of more appropriate treatment therapies that act effectively on problematic components. Furthermore, cycles help differentiate, measure and explain the phenomena observed in healthy and diseased tissues.
- We propose using cycles as the basis of a new method for classification of samples of a microarray (cancer diagnosis). The main advantage of our approach is embedded biolog-

ical significance. Our classification method achieved the same performance than a neural network, which is one of the most common classification method. Thus, our method and a neural network have parity in the diagnosis of the disease. However, while the neural network is a black box, our method provides biological information regarding the response of classification. If the analyzed pathway is related to the disease, our method may explain the action of the disease through cycles, providing hints to biochemists about possible laboratory experiments, as well as on potential drug target genes.

- We use the proposed methodology to analyze the genes of three networks closely related with cancer - apoptosis, glucolysis and cell cycle - in tissues of brain tumor Gliobastoma multiforme (GBM) and in healthy tissues.
  GBM is the most aggressive type of brain tumor. This collection of tumors arise from glia or their pre-cursors within the central nervous system. Because most patients with GBMs die in less than a year, and essentially no patient has long-term survival, these tumors have drawn significant attention. In fact, GBMs have evaded increasingly clever and intricate attempts at therapy over the last half-century. Since the prognosis of GBMs is still extremely poor, the discovery of novel molecular therapeutic targets can be important to optimize treatment strategies (MRUGALA, 2013). Our main results show that the stoichiometric relationship between genes involved in apoptosis, glucolysis, cell cycle pathways is unbalanced in GBM samples *versus* control samples. This dysregulation can be measured and explained by the identification of a higher percentage of positive cycles in these networks. This conclusion helps to understand more about the biology of this tumor type. Furthermore, the results show that our method provided a faithful classification of microarray samples based on the cycles of the three networks.
- We present a NFκB pathway analysis in tissues of the GBM tumor and in healthy tissues using our method. GBM patients carry aberrant NFκB activation, but the pathway mechanisms are not completely understood. The genes of the NFκB pathway are involved in the control of a plethora of biological processes ranging from inhibition of apoptosis to metastasis in cancer. In GBM samples, we show that the stoichiometric relationship between genes involved in NFκB pathway regulation is unbalanced. This can be measured and explained by the identification of a positive cycle. This conclusion helps to understand more about the NFκB pathway and the biology of this tumor type.

1.3 Research Context

The present work is part of the multidisciplinary project that combines knowledge from the fields of Computer Science and Biochemistry, called *BioModelos* (*Modelagem Computacional de Sistemas Biológicos* – Computational Modeling of Biological Systems). This way, we have a close collaboration with researchers from the Department of Biochemistry of the Universidade do Rio Grande do Sul (UFRGS), so that all experiments with models are supported (or not) with experimental analysis made in biochemistry laboratories.

The aim of this project is the construction and validation of different models of biological processes, allowing the analysis and prediction of the reactions that occur in biological machines at different levels of abstraction. This will allow to interfere in a much more controlled and predictable manner, as nowadays occurs when handling models silicon-based machines. In addition, with the availability of models at adequate degrees of abstraction and accuracy, pathway simulation and reactions allow an acceleration in the development of biochemical research, with obvious productivity gains.

The challenge in terms of Computer Science research is provide computational models to reflect complex biological processes reliable and faithfully, that can be used to analyze and simulate in reasonable time the multiple reactions of the huge biological machines. Not only the chain of reactions linked to a physical-chemical excitation is important, but relationships between the different chains are also key, and many details are still unknown to researchers from the Biochemistry area. Even to describe simple pathways, a lot of knowledge about the agents involved is necessary. Each human cell contains about 20,000 to 25,000 genes that can potentially generate the corresponding proteins, which in turn trigger the reactions in a pathway. The same gene may influence various cellular pathways and in different ways (for example, by promoting, inhibiting, accelerating). In addition to genes, other aspects, such as environmental factors, affect pathways. Therefore, the construction of these models requires a lot of interdisciplinary interaction, combining knowledge of Computer Science and Biochemistry areas.

1.4 Thesis Outline

The remaining chapters of this thesis are organized as follows:

- Chapter 2 presents background information on the main issues addressed in this thesis. The concept of gene expression and measurement methods are introduced. We present concepts of the abstraction used to represent networks (graphs) and describe two types of gene networks: (i) gene regulatory networks and (ii) gene co-expression networks. The pathways used as case studies of our methodology are also presented.

- Chapter 3 revises the related literature. The first part of the chapter addresses different methods of analysis of GCNs. The second part, in turn, discusses a set of methods used for classification based on microarray data.
- Chapter 4 details our microarray data analysis methodology. In addition, we present as case studies the analysis of GCNs that represent apoptosis, glucolysis and cell cycle pathways in GBM tumor and in healthy tissues. We close the chapter with a discussion of the results.
- Chapter 5 describes our microarray samples classification methodology. We use the proposed method for classification based on microarray data in tissues of GBM tumor and in healthy tissues using the information obtained in the analysis of GCNs that represents apoptosis, glucolysis and cell cycle pathways. We discuss the results at the end of the chapter.
- Chapter 6 presents a NFκB pathway analysis in tissues of the GBM tumor and in healthy tissues using our method based on the different kind of cycles found among genes of a GCN constructed with quantized data obtained from the microarrays. First, the NFκB pathway is presented. Next, the results are described and discussed.
- Concluding remarks and research perspectives are presented in Chapter 7.

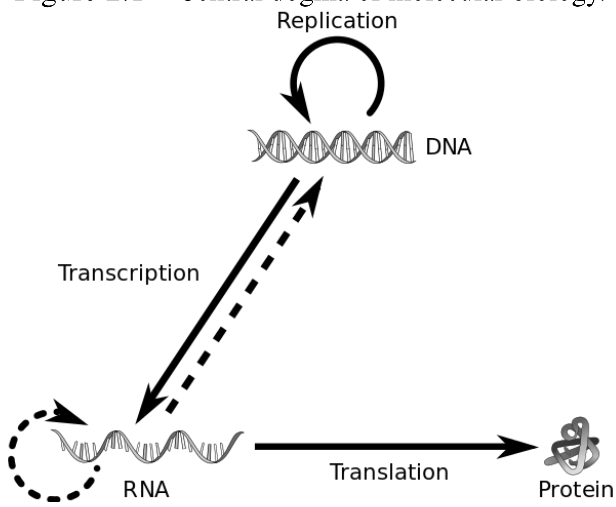
2 BACKGROUND

In this chapter, we present background information on the main issues addressed in this thesis. Section 2.1 introduces the concept of gene expression and presents some measurement methods. In Section 2.2, we provide some concepts of the abstraction used to represent the gene networks: graph. Section 2.3 presents two types of gene networks: (i) gene regulatory networks and (ii) gene co-expression networks. In Section 2.4, we present the apoptosis, glucolysis and cell cycle pathways, which are used as case studies of our methodology.

2.1 Gene Expression

The genetic information of an individual (the *genome*) is encoded in double-stranded deoxyribonucleic acid (DNA) molecules, which are arranged into chromosomes. The DNA describes all the proteins that are potentially present in every cell of a living organism. The *central dogma of molecular biology* says "DNA makes RNA, RNA makes protein, and proteins make the cell" (CRICK, 1970); so it explain how the cell converts the information contained in the DNA in proteins or enzymes. Besides that, a separate process, called *replication*, occurs more rarely, and only when a cell is ready to divide, and results in the DNA duplication, one copy to each of the two daughter cells. See Figure 2.1.

Figure 2.1 – Central dogma of molecular biology.



Source: Barillot et al. (2012).

Gene expression refers to the process by which genetic information gets transformed into working proteins. The main steps are transcription from DNA to RNA, translation from

RNA to linear amino acid sequences, and folding of these into functional proteins. In any given cell of an organism, at any time, thousands of genes and their products (RNA, proteins) actively participate in a synchronized process. Thus gene expression is one of the main determinants of a cell's state, or phenotype. One can, for example, investigate the differences between a normal cell and a cancer cell by examining their relative gene expression profiles.

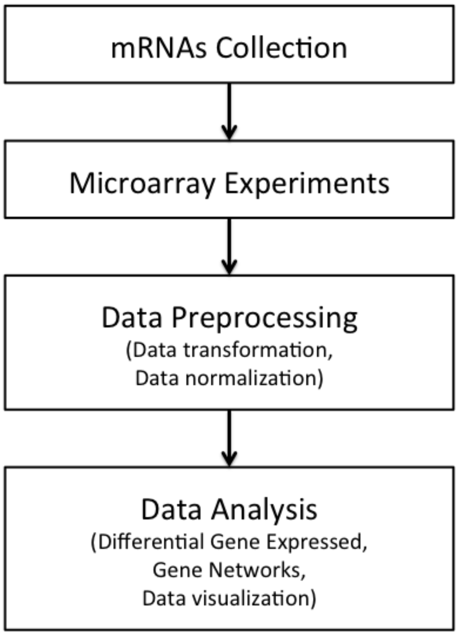
Therefore gene expression provides information for building models of biological systems. When genes are expressed, the genetic information (base sequence) on DNA is first transcribed (copied) to a molecule of messenger RNA (mRNA). The mRNA molecules then leave the cell nucleus and enter the cytoplasm, where they participate in protein synthesis by specifying the particular amino acids that make up individual proteins (National Center for Biotechnology Information - NCBI, 2014).

Gene expression analysis is the determination of the pattern of genes expressed at the level of genetic transcription, under specific circumstances or in a specific cell. Gene expression analysis is used to study regulatory gene defects in cancer and other devastating diseases, cellular responses to the environment, cell cycle variation. Several techniques are available for measuring gene expression, including serial analysis of gene expression (SAGE), cDNA library sequencing, differential display, cDNA subtraction, multiplex quantitative RT–PCR, gene expression microarrays c and RNA-sequencing (RNA-seq) (MANTIONE et al., 2014).

The high-throughput technique known as microarray (or DNA chips) is particularly powerful in providing a global view of gene expression patterns in biological samples. Thousands of genes can now be analyzed using a single microarray hybridization chip (STEKEL, 2003). This technique allows to increase the current knowledge about the causes and mechanisms involved in various complex disorders. Researchers can compare the molecular behavior of different types of cells lines or specific tissues that have been exposed to pathological or experimental conditions. The method may provide insights into physiological processes and facilitate the identification of novel biological markers for diagnostic, prognostic and pharmacological treatments for a number of diseases.

The typical microarray data processing and analysis is presented in Figure 2.2. The process consists of three major steps: (i) determination of the biological problem and sample preparation, (ii) array generation and (iii) data analysis (ZHANG, 2006). In the first step, the RNA sources are collected from the tissues of model systems or diseased/normal patients or from cultivated homogeneous population as appropriate to the particular problem being investigated. Then, RNAs are extracted from these cells.

Figure 2.2 – Microarray analysis flow.

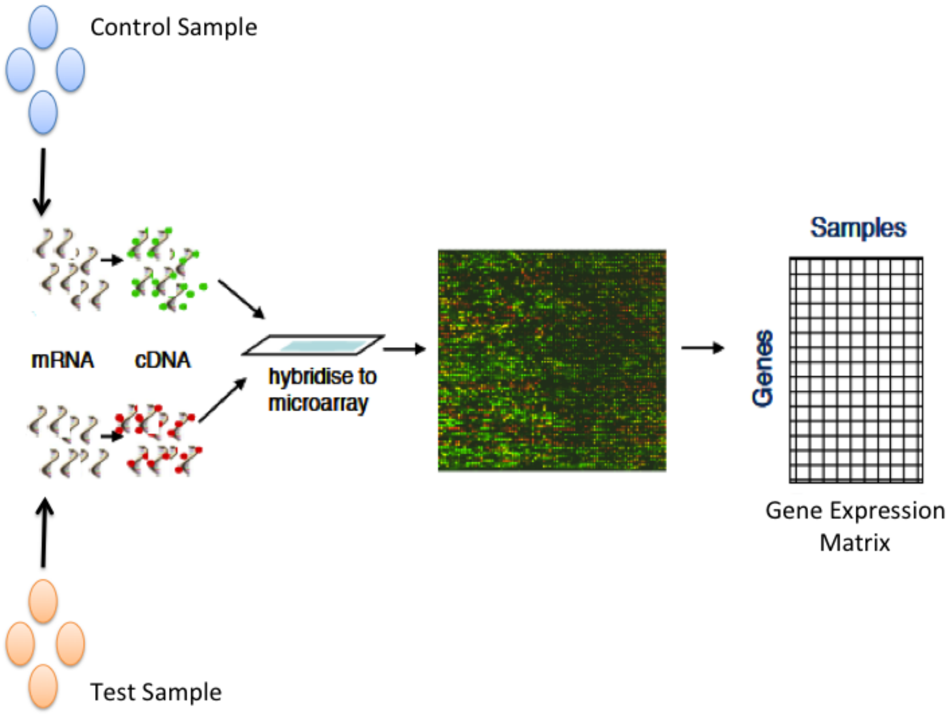


Source: Author (2016).

In the second step, a microarray experiment is performed. There are different types of microarrays. Two different approaches are prevalent, and they differ on the way probes are placed on the slide (PARMIGIANI et al., 2003): (i) Spotted or cDNA microarrays, the DNA molecules (also called *probes*) are synthesized apart and printed mechanically on the chip; and (ii) oligonucleotide chips, whose main representatives are Affymetrix GeneChip, the probes are directly synthesized on the surface. In the latter, a gene is not represented by one probe but by a set of them (a probe set). All types of microarrays follow common basic procedures (Figure 2.3):

- *Chip manufacture*: a microarray is a small chip made of chemically-coated glass, nylon membrane or silicon in which thousand of probes are attached in fixed grids. Each grid cell is related to a DNA sequence.
- mRNA preparation, labeling and hybridization: usually, two mRNA samples (a test sample and a control sample) are reverse-transcribed into complementary DNA strand (cDNAs) targets, labeled using either fluorescent dyes or radioactive isotopics, and then hybridized with the cloned sequences on the surface of the chip.
- Chip scanning: chips are scanned to read the signal intensity that is emitted from the labeled and hybridized targets. The scanned image consists in a grid of shined spots, each one corresponding to a probe. Finally, this image is processed and transformed into numbers, consisting the basis of the gene expression analysis.

Figure 2.3 – Microarray experiment.



Source: Adapted from supplementary material provided by Schlitt and Brazma (2007).

When theses procedures are completed, the raw microarray data are obtained, then data preprocessing may need to be executed prior to any data analysis. The preprocessing include data transformation and data normalization. After this, the microarray data can typically be represented by a two-dimensional matrix $X = x_{ij}$, where each row i in the data matrix correspond to one gene, each column j corresponds to each biological sample (experimental condition: disease or normal tissue; or different time points), and each cell x_{ij} is a real value recording the expression level of gene i from sample j (or under condition j). Finally, the last step can be performed: data analysis and visualization from the preprocessed data.

Gene expression measuring technologies are in continuous progress. Despite the advances in experiments technology, data availability is still an obstacle to be overcome in the process of model biological process. Microarray data presents some challenges such as the *dimensionality problem* due to the scarcity of biological samples (instances). Microarray analysis typically measure tens of thousands of genes in only tens of samples, which increases the risk of detecting spurious relationships. Another important problem concerning biological data is that the gene expression data provided by microarray quantify the concentration of mRNA and ignores information about possible interventions and environmental changes after the transcription phase.

2.2 Graphs

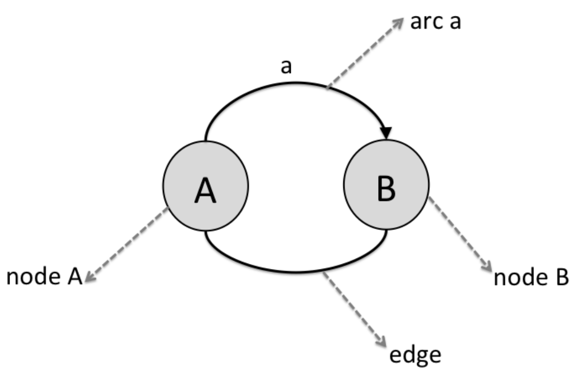
Systems biology (IDEKER; GALITSKI; HOOD, 2001) is the study of an organism seen as a *network* of biological components (genes, proteins, and chemical reactions). Instead of analyzing individual components or aspects of the organism, the focus are all components and the interaction between them, all as part of a single system, as indeed happens in reality. These interactions are responsible for the shape and the functions of a body. Systems Biology seeks to understand these complex interactions, as they are the key to understanding life.

A *network* is a collection of connected objects. Usually, the objects are called nodes or vertices and drawn as points (or circles); the connections between the nodes are called edges and drawn as lines between nodes (points). In mathematics, networks are often called as *graphs* and its study is referred as *graph teory* (BORNHOLDT; SCHUSTER, 2003). Graphs can represent all sorts of systems in the real world. One could, for example, describe the World Wide Web as a graph where the pages are nodes and links are the edges. Social networks are other examples of graphs.

Formally, a graph G is a pair (V, E), where V is the set of nodes (also called vertices) and E is the set of edges (CORMEN et al., 2001). An edge $e_{ij} \in E$ is a pair $e_{ij} = (v_i, v_j)$, where v_i and v_j are nodes ($v_i \in V$ and $v_j \in V$) to be connected, linked or adjacent to each other. Nodes and edges can have labels (i.e., character strings, name, words attached to them) and weights (i.e., numerical values). Nodes with a large number of connections are hubs. To model the 'real world', nodes are used to represent entities and edges various relationships between these entities. Usually, nodes are depicted as dots or circles and edges as lines connecting nodes; arrows represent arcs (directed edges), where the arrowheads indicate the directionality. Figure 2.4 presents basic elements of a graph representation.

Graphs can be directed or undirected (CORMEN et al., 2001). In the first case, edges are ordered pairs of nodes and are often called arcs: (v_i, v_j) means an arc from v_i to v_j and (v_j, v_i) means an arc from v_j to v_i. Directed graphs can be effectively used to represent causal influences or communication between the nodes. In these graphs, we call nodes with outgoing edges source nodes and nodes with incoming edges target nodes; for each source node we define the target set as the set of all its target nodes. In an undirected graph, the edges do not have a particular direction: if $e_1 \in E$ and $e_1 = (v_i, v_j)$, where $v_i \in V$ and $v_j \in V$, and $e_2 \in E$ and $e_2 = (v_j, v_i)$, where $v_j \in V$ and $v_i \in V$, then $e_1 = e_2$.

Figure 2.4 – Basic elements of a graph representation – Nodes are represented as circles and are named as A and B; they, for example, represent gene A and gene B. Connections between nodes represent an undirected relationship (edge) and a relationship where directionality is important (arc named a). The source node A have an outgoing arc; and the target node have and incoming arc. The number of incoming and outgoing arcs (and edges) is the degree of a node (node A and node B have degree 2).



Source: Author (2016).

A mixed graph G is a graph in which some edges may be directed and some may be undirected. It is written as an ordered triple (V, E, D), where V is the set of nodes, E is the set of edges and D is the set of arcs. With an edge an and arc defined as above. Directed and undirected graphs are special cases of such mixed graphs. These graphs can thus represent associations as well as causal influences between the nodes (DAS et al., 2009).

A graph where all nodes are connected to each other by an edge is called a complete graph (CORMEN et al., 2001). A subgraph of a graph G is another graph formed from a subset of the nodes and edges of G. A clique is a subset of the nodes in a graph such that every pair of nodes are connected by an edge (i.e., a clique is complete subgraph) (CORMEN et al., 2001).

A walk is an alternating sequence of nodes and edges, starting and ending at a vertex, in which each edge is adjacent in the sequence to its two endpoints. In a directed graph, the ordering of the endpoints of each edge in the sequence must be consistent with the direction of the edge. A closed walk is one that starts and ends at the same node. A cycle is a closed walk with all nodes distinct (except the first and last nodes) (CORMEN et al., 2001).

The degree of a node is defined as a number of connections (edges) adjacent to this node. For a node in a directed graph we can distinguish between the number of incoming arcs (indegree) and outgoing arcs (outdegree). In an undirected network only the degree is defined as the number of edges of a node (CORMEN et al., 2001). The average degree of a graph G is a measure of how many edges are in set E compared to number of nodes in set V. A diameter of a graph is the largest number of nodes which must be traversed in order to travel from one node to another when backtrack, detour, or loop are excluded from consideration (CORMEN et al., 2001). The density of a graph G measures how many edges are in set E compared to the maximum possible number of edges between nodes in set V (CORMEN et al., 2001).

2.3 Gene Networks

Genomic high-throughput technologies are generating a great amount of biological data and represent a fertile source of knowledge. These data allowed to get inside main features of gene expression and its regulation and, at the same time, to discover a more complex level of organization (DAS et al., 2009). Networks offer a theoretical picture that can be used to explain and analyze the structure of these data, and so from biological systems and their evolution. Many theoretical studies on networks have demonstrated their application to model biological networks, as, for example, metabolic networks, neuronal networks and gene networks.

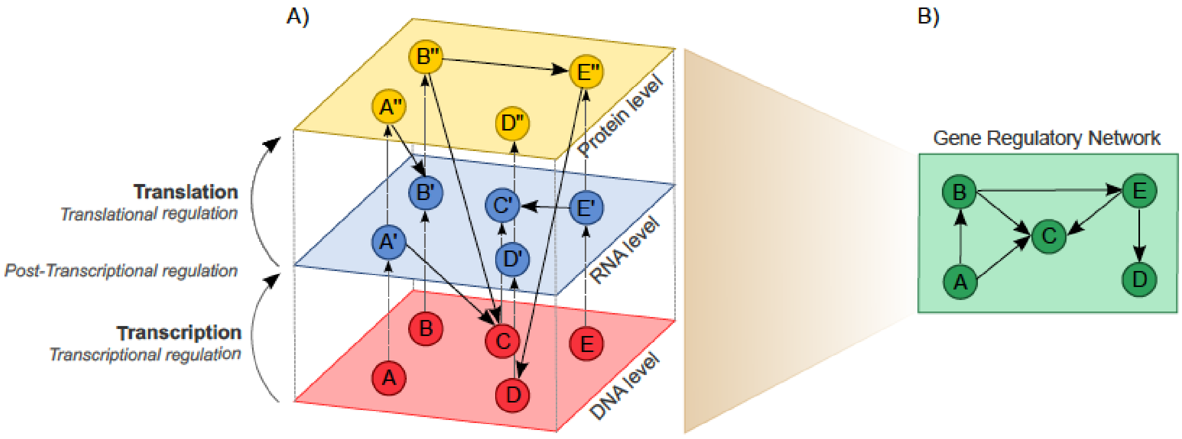
Many different kinds of gene networks can be obtained, depending on which particular biological target is considered. The Subsection 2.3.1 and 2.3.2 present, respectively, two types of gene networks: (i) gene regulatory networks and (ii) gene co-expression networks. Often, in the literature, the two terms are used as synonyms, but actually they are not.

2.3.1 Gene Regulatory Networks

A gene regulatory network (GRN) lies at the core of intracellular signal transduction. A GRN can be defined as a collection of DNA segments in a cell which interact with each other indirectly through their RNA and protein products (and with other substances in the cell), thereby governing the rates at which genes in the network are transcribed into mRNA (STURROCK, 2013). In other words, these networks consist of a great number of genes whose expression levels affect each other in various ways. Computational models of GRNs can take a variety of forms. A graphical representation of a GRN is present in Figure 2.5.

A GRN is a mixed graph $G = (V, E, D)$ over a set V of nodes, corresponding to gene-activities (gene expression levels or RNA concentrations), with unordered pairs E, the undirected edges, and ordered pairs D, the directed edges (DAS et al., 2009). A directed edge d_{ij} from v_i to v_j is present if and only if a causal effect runs from node v_i to v_j and there exist no nodes or subsets of nodes in V that are intermediating the causal influence (it may be mediated by hidden variables, i.e. variables not in V). An undirected edge e_{ij} between nodes v_i and v_j is present if and only if gene-activities v_i and v_j are associated by other means than a direct causal

Figure 2.5 – Gene regulatory network graphical representation – (A) A hypothetical instance of the multilayered regulatory machinery underlies organisms functioning. The genetic elements are organized in three levels: DNA, RNA and protein. Regulatory interactions, in turn, are distributed in transcriptional layer, the post-transcriptional layer and the translational layer. These interaction could occur both intra an inter-level, increasing the complexity of the system. (B) A simplified representation of a GRN, which is given as a graph model, where the regulatory layers are no longer distinguishable and the type of interactions covered by the model depends on the experiment goal, available data and the biological knowledge.



Source: Mendoza (2014).

influence, and there exist no nodes or subsets of nodes in V that explain that association (it is caused by a variable hidden to V).

The directed edges in GRNs correspond to causal influences between gene-activities. More accurately, these could include regulation of transcription by transcription factors, but also less intuitive causal effects between genes involving signal-transduction or metabolism. When GRNs are inferring from gene-expression data alone, the metabolites and proteins act as hidden variables. These variables mediate communication between genes, but since they are not included explicitly in the GRNs, only their effects appear as edges between the observed variables. Only cause-effect relations between observed quantities can be established. Thus these networks describe communication between genes implicitly including all regulatory processes inside living cells and therefore give a complete description of cellular regulation projected on the gene activities. The undirected edges in GRNs, in turn, correspond to associations (for instance, correlations) between gene expression levels or RNA concentrations. These edges should not be confused with reciprocal effects, i.e. two nodes that are connected by directed edges in both directions (DAS et al., 2009).

2.3.2 Gene Co-expression Networks

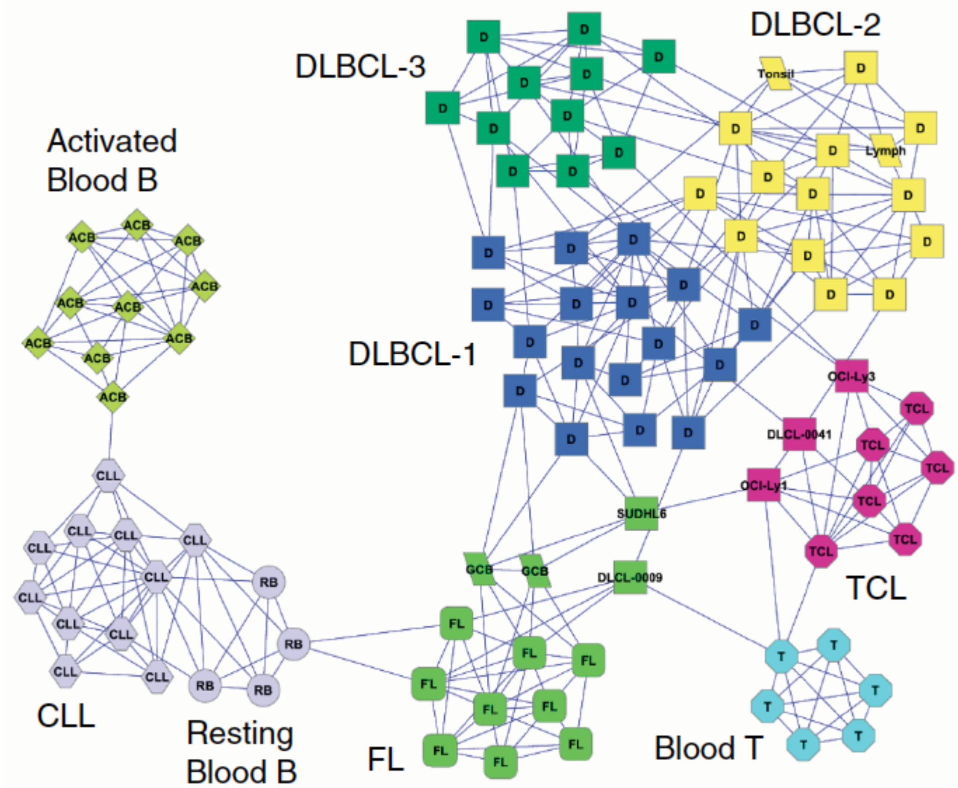
A gene co-expression network (GCN) is an undirected graph $G = (V, E)$ over a set V of nodes and a set of the undirected edges, unordered pairs, E (DAS et al., 2009). The abstraction from biological data to the mathematical model (graph) is realized by mapping genes to nodes and putting edges representing similarity of gene expression according to a given quantitative notion of similarity (or dissimilarity). Thus, GCNs are inferred from gene expression data. Figure 2.6 shows an example of a GCN.

Of particular interest is the problem of analyzing GCNs construct by using gene expression data from DNA microarray experiments. Microarrays quantify gene expression by measuring the hybridization, or matching, of DNA immobilized on a small matrix (array) to mRNA representation from the sample under study. Arrays can currently have hundreds of thousands of spots, so can measure simultaneously a large proportion of the genes on the genome allows the investigation of the interactions among the genes on a large scale, the discovery of the role of the vast number of genes whose function is not adequately understood and the characterization of how pathways are changed under varying conditions (PARMIGIANI et al., 2003). Thus, GCNs provide a conceptual framework to study gene interactions. However, their static representation does not capture all possible gene relationships as these do not operate simultaneously due to spatial and temporal variation in gene expression (SERIN et al., 2016). Other applications of GCNs include the gene functional annotation and the comparison of GCNs across tissues and/or species. These different analysis will further contribute to the elucidation of important biological processes and provide a valuable predictive tool for molecular breeding (SERIN et al., 2016).

The network construction is conceptually straightforward: nodes represent genes and nodes are connected if the corresponding genes have significantly similarity of gene expression according to an used quantitative notion. Hence a co-expression link exists when the similarity measure between two genes is higher than a defined cutoff. There are several questions associated with the choice of a cutoff (ZHANG; HORVATH, 2005). The first is “how to pick a cutoff?”. One strategie is picking a number based on the notion of statistical significance. Its drawbacks include loss of information (if the threshold has been set to 0.8, there will be no connection between two nodes if their similarity equals 0.79) and sensitivity to the threshold.

Usually, a networks can be specified with the following adjacency matrix: $A = [a_{ij}]$ is symmetric with entries in $[0, 1]$. By convention, the diagonal elements are assumed to be zero.

Figure 2.6 – An example of gene co-expression network.



Source: Ruan, Dean and Zhang (2010).

For unweighted networks, the adjacency matrix contains binary information (one if connected and zero if unconnected). In weighted networks, the adjacency matrix contains weights.

Given two genes it is possible to use different quantitative measures of co-expression to construct different GCNs. Thus, two genes are connected by an undirected edge if their activities have significant association over a series of gene expression measurements. Usually, Pearson correlation, Spearman correlation or Mutual Information are used to quantify similarity of gene expression (DAS et al., 2009).

It is important to emphasize the difference between GCNs and GRNs, since the first has also been incorrectly called GRNs in the literature by several authors (DAS et al., 2009). Gene activities (gene expression levels or RNA concentrations) can be correlated due to different causal relationships: (i) direct effects, (ii) indirect effects (correlation is transitive) and (iii) confounding. Several algorithms have been proposed to eliminate edges corresponding to the cases (ii) and (iii), thus resulting in a network which is the undirected version of the GRN. A correlation does not imply causation and many of the undirected edges may be due to hidden confounding factors. Only gene expression data obtained through a strategy of gene perturbations, or other targeted disturbances to the system, allow for inferring causal relationships. While it has been shown that under certain assumptions it is possible to infer causality without making experimental interventions, such assumptions are unfortunately not justified in this context. The strongest assumption is that there are no hidden variables with confounding effects on the observed variables. Given the fact that gene expression levels are generally the only observed quantities in the data used to infer GCNs or GRNs, and that all variables mediating the causal effects between them, i.e. the proteins and metabolites are hidden, such assumption can not be justified under any circumstance. Gene perturbations are thus necessary to infer causality and thus GRNs. Such perturbations could be experimentally created by knocking-out or over-expressing genes. Also natural genetic polymorphisms could be used to infer causal relationships between gene-activities. Therefore, the major difference between the two types of network is the semantics of undirected edges.

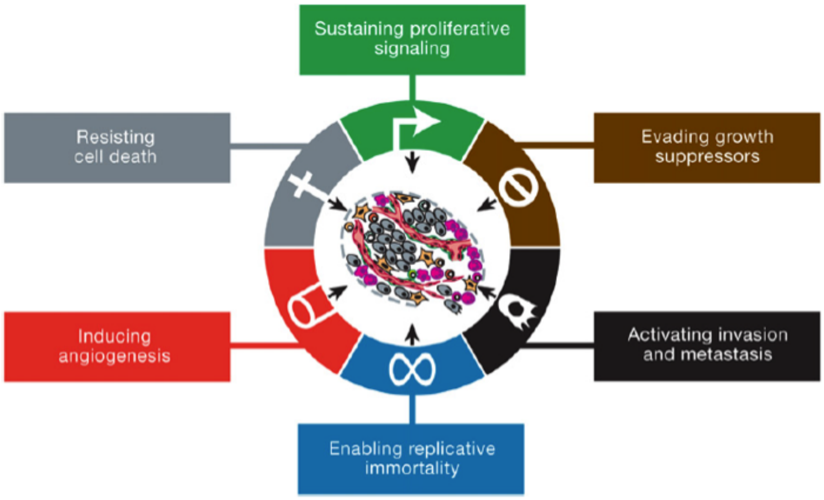
2.4 Pathways

Hanahan and Weinberg (2011) proposed six hallmarks of cancer that together constitute an organizing principle for rationalizing the complexities of neoplastic disease. The hallmarks of cancer comprise biological capabilities acquired during the multistep development of human tumors. Implicit in their discussion was the notion that as normal cells evolve progressively to a neoplastic state, they acquire a succession of these hallmark capabilities. The multistep process of human tumor pathogenesis could be rationalized by the need of incipient cancer cells to acquire the traits that enable them to become tumorigenic and ultimately malignant.

The hallmarks include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis (HANAHAN; WEINBERG, 2011). Underlying these capabilities are genome instability, which generates the genetic diversity that expedites their acquisition, and inflammation, which promotes multiple hallmark functions. Advances in the last decade has added two emerging hallmarks of potential generality to the original list: (i) reprogramming of energy metabolism and (ii) evading immune destruction.

In this work, we have a close collaboration with researchers from the Department of Biochemistry of the Universidade do Rio Grande do Sul (UFRGS), so that all experiments with models are supported (or not) with experimental analysis made in biochemistry laboratories. To this end, we chose, together with the biochemistry group, as case studies three networks closely related with these hallmarks: apoptosis, glucolysis and cell cycle. These three pathways are discussed in the following subsections.

Figure 2.7 – Hallmarks of cancer - the first six hallmark capabilities proposed in 2000.



Source: Hanahan and Weinberg (2011).

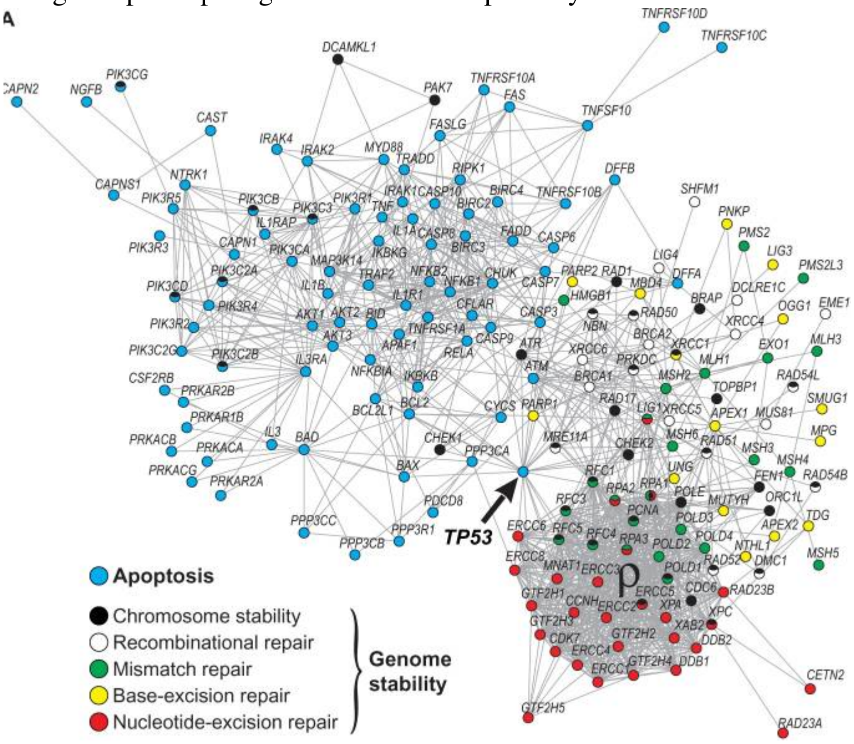
2.4.1 Apoptosis

Apoptosis is the process of programmed cell death. It is considered a vital component of various processes including normal cell turnover, proper development and functioning of the immune system, hormone-dependent atrophy, embryonic development and chemical-induced cell death (ELMORE, 2007). The apoptosis is associated with the maintenance of tissue homeostasis: the programmed cell death guarantees the substitution of old and/or dysfunctional cells in multicellular organisms, which are impaired by the accumulation of cellular damages due to environmental insults, as well as participates directly in tissue development (CASTRO et al., 2008).

There are more then 100 genes (Figure 2.8) working coordinately in apoptosis (KANEHISA et al., 2006). Removing one of these components affects several others and affect the whole pathway (CASTRO et al., 2008). Inappropriate apoptosis is a factor in many human conditions including neurodegenerative diseases, ischemic damage, autoimmune disorders and many types of cancer (ELMORE, 2007). Suppressed apoptosis, which is associated with uncontrolled cell growth is typically found in neoplastic diseases.

The apoptotic machinery is composed of both upstream regulators and downstream effector components. The first group is divided into two major circuits, one receiving and processing extracellular death-inducing signals and the other sensing and integrating a variety of

Figure 2.8 – Human apoptosis and genome-stability gene network. Gene network from apoptosis and DNA repair pathways. Different pathways are represented in different colors; and nodes with more than one color represent genes participating in more than one pathway.



Source: Castro et al. (2008).

signals of intracellular (intrinsic) origin. Currently, the intrinsic apoptotic program is more widely implicated as a barrier to cancer pathogenesis (HANAHAN; WEINBERG, 2011). The "apoptotic trigger" that transmit signals between the regulators and effectors is controlled by counterbalancing pro- and antiapoptotic members of the BCL2 family of regulatory protein. This family are inhibitors of apoptosis, acting in largepart by binding to and thereby suppressing two proapoptotic triggering proteins (BAX and BAK). When the proapoptotic signaling proteins are release (cytochrome c), a cascade of caspases, that act via their proteolytic activities to induce the multiple cellular changes associated with the apoptotic program, is activated. Between several abnormality sensors, that play key roles in tumor development, is a DNA damage sensor, that functions via the TP53 tumor suppressor; TP53 induces apoptosis by upregulating expression of the Noxa and Puma BH3-only proteins. Alternatively, insufficient survival factor signaling can elicit apoptosis through a BH3-only protein called Bim. Hyperactive signaling by certain oncoproteins, such as Myc, also triggers apoptosis unless counterbalanced by antiapoptotic factors. Tumor cells evolve a variety of strategies to limit or bypass apoptosis (HANAHAN; WEINBERG, 2011). Most common is the loss of TP53 tumor suppressor function, which eliminates this critical damage sensor from the apoptosis-inducing circuitry. Tumors also may achieve similar results by increasing expression of antiapoptotic regulators or of survival signals, by downregulating proapoptotic factors, or by short-circuiting the extrinsic ligand-induced death pathway. This mechanisms for avoiding apoptosis presumably reflects the diversity of apoptosis-inducing signals that cancer cell populations encounter during their evolution to the malignant state.

The programmed cell death is recognized for its immense therapeutic potential, as a barrier to cancer. However more research is necessary for elucidated how are the mechanisms of action or inaction of the proteins from the apoptosis machinery.

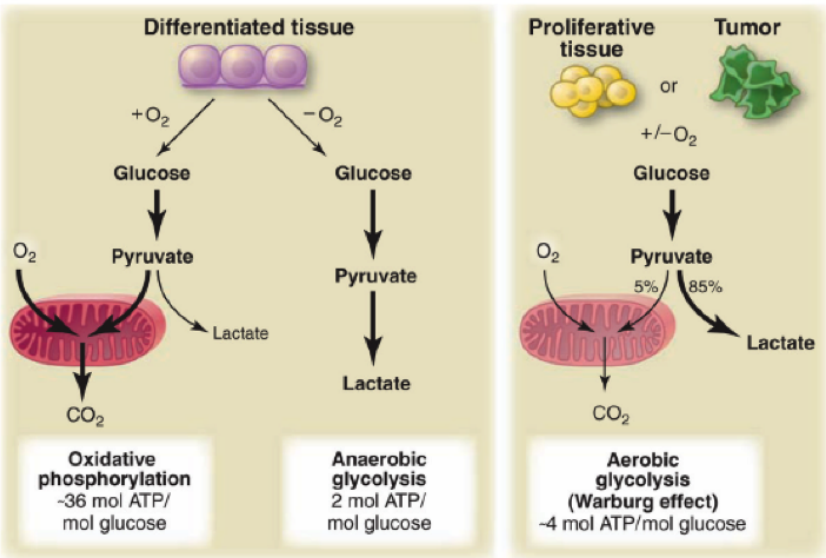
2.4.2 Glucolysis

Glucolysis (glycolysis) is the metabolic pathway through which glucose is converted into pyruvate (LEMAIGRE; ROUSSEAU, 1994). It is stimulated when the energy charge falls. In yeast, glucose facilitates its own use by inducing expression of genes involved in its metabolism while repressing that of those involved in the utilization of alternative carbon sources. In mammals, the response to the glucose is more complex because it combines effects related to glucose metabolism itself and effects secondary to glucose-dependent hormonal modifications, mainly pancreatic stimulation of insulin secretion and inhibition of glucagon secretion (VAULONT; VASSEUR-COGNET; KAHN, 2000).

Glucolysis occurs in every tissue and its main function is to provide substrates for anabolic processes. The pathway flux must be adjusted according to conditions both inside and outside the cell. The rate of conversion of glucose into pyruvate is regulated to meet two major cellular needs (BERG; TYMOCZKO; STRYER, 2002): (i) the production of ATP (energy), generated by the degradation of glucose, and (ii) the provision of building blocks for synthetic reactions, such as the formation of fatty acids. Glucolysis can occur with oxygen (aerobic) or without oxygen (anaerobic). In the former, glucolysis is the first stage of cellular respiration. In the later, glucolysis allows cells to make small amounts of ATP: this process is called fermentation, that produce lactic acid.

Second Heiden, Cantley and Thompson (2009), in the presence of oxygen, nonproliferating (differentiated) tissues first metabolize glucose to pyruvate via glucolysis and then completely oxidize most of that pyruvate in the mitochondria to CO2 during the process of oxidative phosphorylation. Oxygen is required as the final electron acceptor to completely oxidize the glucose. Thus oxygen is essential for this process. When oxygen is limiting, cells can redirect the pyruvate generated by glucolysis away from mitochondrial oxidative phosphorylation by generating lactate (anaerobic glucolysis). This process of lactate produce during anaerobic glycolysis allows glycolysis to continue (by cycling NADH back to NAD+ – LDHA is the gene which encodes these (National Center for Biotechnology Information - NCBI, 2016)), but results in minimal ATP production when compared with oxidative phosphorylation. Cancer cells, in turn, make adjustments of energy metabolism in order to fuel cell growth and division caused by deregulated control of cell proliferation, that represents the essence of neoplastic disease. So these cells reprogram their energy production, by limiting their energy metabolism largely to glycolysis (HANAHAN; WEINBERG, 2011), and tend to convert most glucose to lactate regardless of whether oxygen is present (aerobic glycolysis or Warburg Effect) (GILLIES; ROBEY; GATENBY, 2008). This property is shared by normal proliferative tissues. Mitochondria remains functional and some oxidative phosphorylation continues in both cancer cells and normal proliferating cells. However, aerobic glycolysis is less efficient than oxidative phosphorylation for generating ATP. Figure 2.9 shows a schematic representation of the differences between oxidative phosphorylation, anaerobic glucolysis and aerobic glucolysis.

Figure 2.9 – Difference between oxidative phosphorylation, aerobic glucolysis and anaerobic glucolysis.



Source: Heiden, Cantley and Thompson (2009).

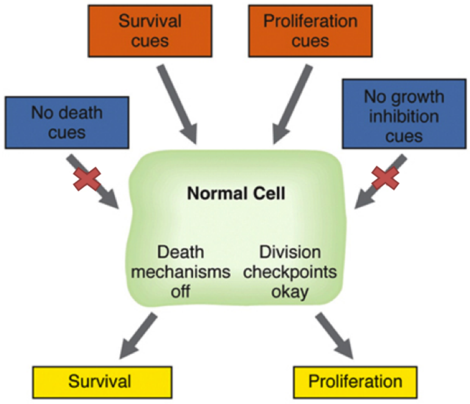
The reprogrammed energy metabolism as an emerging hallmark of cancer (HANAHAN; WEINBERG, 2011). This reprogramming of energy metabolism is seemingly counterintuitive, once cancer cells must compensate for the lower efficiency of ATP production provided by glycolysis relative to mitochondrial oxidative phosphorylation. However, this cells do so in part by upregulating glucose transporters, which substantially increases glucose import into the cytoplasm. Heiden, Cantley and Thompson (2009) suggest that increased glycolysis allows the diversion of glycolytic intermediates into various biosynthetic pathways, including those generating nucleosides and amino acids; this facilitates, in turn, the biosynthesis of the macromolecules and organelles required for assembling new cells.

2.4.3 Cell cycle

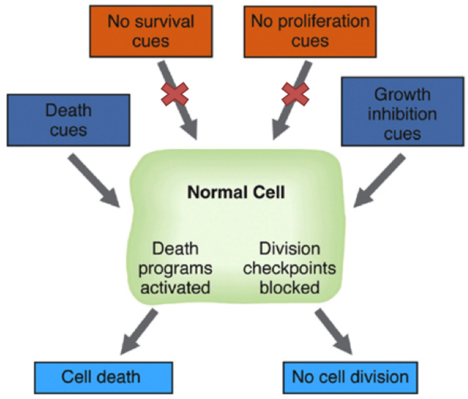
The cell cycle pathway governs cell growth-and-division. It controls the transition from quiescence to cell proliferation, and through its checkpoints, ensures the fidelity of the genetic transcript (SCHWARTZ; SHAH, 2005). The pathway mechanism is divided into four phases: (i) chromosomes are replicated once (DNA synthesis or S-phase) and (ii) segregated to create two genetically identical daughter cells (mitosis or M-phase). These events are spaced by intervals of growth and reorganization, gap phases (iii) G1 and (iv) G2. Cells can stop cycling after division, entering a state of quiescence. In G1, cells assume the commitment to traverse an entire cycle.

Crucial parts of the cell cycle machinery are the cyclin-dependent kinases (CDKs), which, when activated, drive the cell forward from one phase of the cell cycle to the next. The CDKs are regulated positively by cyclins and regulated negatively by naturally occurring CDK inhibitors (CDKIs). The pattern of cyclin expression varies with a progression of the cell through the cell cycle, and this specific cyclin expression pattern defines the relative position of the cell within the cell cycle. Cancer cells present a dysregulation of the cell cycle such that cyclins are overexpressed or the CDKIs are not expressed, such that the cell growth become unregulated (SCHWARTZ; SHAH, 2005).

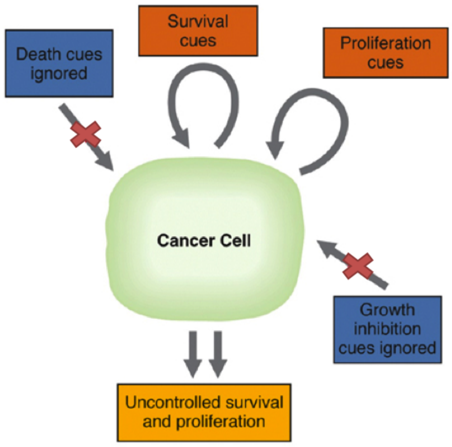
Cell proliferation is necessary for replacement of destroyed cells, growth and development (SCHWARTZ; SHAH, 2005; HANAHAN; WEINBERG, 2011). Survival and proliferation controls are highly integrated and dependent on inter-cellular communications: cells constantly evaluate their own condition via continuous communication among neighboring cells and tissues. Proliferation is regulated at the level of mitosis. Cells may be triggered into an apoptotic cycle if they are damaged, dangerously abnormal, or needed only transiently during embryo development. Normal tissues control the production and release of growth-and-division cycle, thereby ensuring a homeostasis of cell number and thus maintenance of normal tissue architecture and function. In cancer cells, in opposite, proliferation and apoptosis mechanisms have failed due to mutations in tumor-suppressing genes, so the signals of these pathways are deregulating and may influence yet other cell-biological properties, such as cell survival and



(a) Normal cell – proliferation context.



(b) Normal cell – death context.



(c) Cancer cell – proliferation context.

Figure 2.10 – Cell proliferation – In this schematic representation, one can see the differences between normal cell and cancer cell proliferation. The normal cells are sensitive to cellular death and their proliferation is dependent of external stimuli. The cancer cells, in turn, are resistant to cellular death and their proliferation is independent of external stimuli.
Source: Krempels (2016)

energy metabolism. Figure 2.10 shows a schematic representation of the differences between normal cell and cancer cell proliferation.

Thus the cell cycle machinery controls cell proliferation. An inappropriate cell proliferation is a cancer characteristic. So this pathway is related to a cancer hallmark: sustaining proliferative signaling (HANAHAN; WEINBERG, 2011). The cell cycle also serves to protect the cell from DNA damage. Therefore, cell cycle also represents a survival mechanism that provides the cancer cell the opportunity to repair its own damaged DNA. So cell cycle checkpoints, before DNA repair is complete, can activate the apoptotis pathway, leading to cell death. An understanding of the cell cycle is critical to understanding how best to develop drugs that directly inhibit the CDKs, inhibit unrestricted cell growth, induce growth stop and make the tumor cell susceptible to apoptosis (SCHWARTZ; SHAH, 2005).

3 RELATED WORK

The present chapter revises the related literature. The first part, Section 3.1, addresses different methods of GCNs analysis. The second part, Section 3.2, in turn, discusses a set of methods used for microarray data classification. Finally, Section 3.3 presents a summary of the chapter.

3.1 Gene Co-expression Networks Analysis

The availability of omics data allowed the system-wide study of the flow of biological information underlying. However, dealing with the large datasets represents a challenging effort that requires the development of powerful bioinformatics methods. In this context, a common approach is the construction and analysis of gene networks (SERIN et al., 2016). One option of network is the GCN, based on gene expression similarity. One of its main applications is the functional annotation of unknown genes.

GCNs analysis allows the simultaneous identification, clustering and exploration of thousands of genes with similar expression patterns across multiple conditions. GCNs contains gene neighborhood relations that are usually overlooked in tradicional cluster analysis (RUAN; DEAN; ZHANG, 2010). These relations permit an interesting geometric interpretation of GCNs. A principle called guilt-by-association, extensively validated, states that genes which are associated or interacting (hence present similar expression profiles) sharing the same function or are involved in the same pathway and tend to form clusters or modules in the network (SERIN et al., 2016). In this way, genes of known function can be used to predict the function of co-expressed unknown genes within the same module.

Although the GCN construction is generally straightforward (commented in Chapter 2, section 2.3.2), the resulting GCN can become very complex, with thousands of nodes that difficult the visualization and comprehension, and limits its biological interpretation (SERIN et al., 2016). Several strategies can be employed to improve the analysis of the resulting network. It is necessary to establish a coherent strategy with the biological question addressed to obtain reliable information from the network. The elucidation of relationships between genes in the GCN can be enhanced by using prior knowledge and data integration (for example, metabolome and proteome). Thus, these networks can provide more than the simple visualization of co-expressed genes.

A meaningful GCN analysis depend on a combination of factors involved in the network inference process: the quality, type and availability of the input data, the correlation coefficient, the prior knowledge, the experimental and computational resources, any negligence can result in unreliable networks and subsequent wrong biological interpretations (SERIN et al., 2016). GCNs analysis allows the development of biological relevant hypotheses and have assisted in the design of data-driven hypothesis experiments and gene prioritization for those experiments. Also, it provides novel insights into the system-level understanding of cellular processes (AOKI; OGATA; SHIBATA, 2007). Thus, it can help many researchers. In this study, we reviewed different GCNs analysis approaches.

3.1.1 Topology analysis

The different network properties can provide valuable insight into the internal organization of a GCN. Several studies have analyzed the topological properties of GCNs (HSU; JUAN; HUANG, 2015; VIALANEIX et al., 2013; XULVI-BRUNET; LI, 2010). Different graph properties are commonly analyzed in networks (RUAN; DEAN; ZHANG, 2010; PAVLOPOULOS et al., 2011; XULVI-BRUNET; LI, 2010):

- *Singletons* - is a node with zero connection;
- *Average degree* - is the average number of connections per node;
- *Clustering coefficient* - is the measurement that shows the tendency of a graph to be divided into clusters;
- *Average path length* - is the average shortest path, or the smallest number of edges needed to connect two nodes, between any two reachable nodes in the network;
- *Diameter* - is the longest path length between any two reachable nodes in the network;
- *Density* - shows how sparse or dense a graph is according to the number of connections per node set.

Previous studies have analyzed the topological properties of GCNs and have shown that these networks have small-world or scale-free properties, similar to many other biological networks and real world networks. These studies have also reported that GCNs differ from other types of biological networks in node degree and hierarchical organization (RUAN; DEAN; ZHANG, 2010). A small-world network has a small diameter and a large clustering coefficient, usually related to an efficient and controlled flow of information (RUAN; DEAN; ZHANG, 2010). A scale-free network is characterized by a few nodes highly connected, acting as hubs, and most nodes with low degrees. These networks are considered robust to random failures, but vulnerable to deliberate attacks (RUAN; DEAN; ZHANG, 2010).

3.1.2 Centrality analysis

Network centrality is the method of ranking network elements used to identify interesting elements of a network (KOSCHÜTZKI; SCHREIBER, 2008). It is particularly useful to identify key players in biological processes (SERIN et al., 2016; KOSCHÜTZKI; SCHREIBER, 2008). That method consists of analyze parameters derived from network topology. Thus, this analysis provided also structural information of the networks. Different centrality measures indicate the importance of individual elements (nodes) of the network. Common centrality measures result in different valuations of the elements (nodes) (KOSCHÜTZKI; SCHREIBER, 2008).

Formally, a network centrality is a function C that assigns every vertex v of a graph a numeric value $C(v)$. Usually, a vertex u is more important than another vertex v if and only if $C(u) > C(v)$ (KOSCHÜTZKI; SCHREIBER, 2008). Some network centralities often used (KOSCHÜTZKI; SCHREIBER, 2008; PAVLOPOULOS et al., 2011):

- *Degree Centrality* - is defined by the vertices degree. Thus, an important node has a large number of interactions. Nodes with highest degree centrality are called *hubs* because they are connected to many neighbors. This is a local centrality measure, since only the immediate neighborhood of the vertex is considered. However, the removal of such central nodes has great impact on the topology of the network. It is common to analyze hub genes, since they are almost always disease-correlated, as putative mediators of pathology (GAITERI et al., 2014). This approach allows valuable insights. However, it tends to focus attention on a few hub genes, so ignoring the many other ways in the networks can be used to generate and translate systems biology insights into testable hypotheses.
- *Eigenvector Centrality* - is a measure of the influence of a node in a network: a node is important if it is linked to by other important nodes. It is a natural extension of degree centrality. In degree centrality, for every interaction, a node awarded one centrality point. However, not all vertices are equivalent: some are more relevant than others, and, reasonably, links from important nodes count more. This is a global centrality measure, since ranks higher the nodes that are connected to important neighbors.
- *Closeness Centrality* - is defined by the sum of the minimal distances of a vertex to all other vertices. Thus, an important node can communicates quickly with other nodes of the

network. This is a global centrality measure, since uses information about the length of the shortest paths within a network. These nodes can help to identify network bottlenecks.

- *Betweenness Centrality* - is defined as the number of the shortest paths that go through an vertex. Thus, an important node is intermediate between neighborhoods, it lies on a high proportion of paths between other nodes in the network. Without the node, there would be no way for two neighborhoods to communicate with each other. This is a global centrality measure, since vertices with high betweenness may have considerable influence within a network by virtue of their control over information passing between others. They are also the ones whose removal from the network will most disrupt communications between other vertices because they lie on the largest number of paths taken by messages. These nodes can also help to identify network bottlenecks.

3.1.3 Modularity analysis

The identification of the modular structure of a network is the main aim of the GCNs modularity analysis. Modules detection strategies depend only on the topological properties of the networks. Modules are also called “functional modules” since they are associated with specific biological processes. Modules are usually defined as groups of highly connected genes (SERIN et al., 2016). Also, modules can be considered as subnetworks, where the nodes within each subnetwork are relatively densely connected to one another but have fewer connections to the other subnetworks (RUAN; DEAN; ZHANG, 2010). Genes within each subnetwork are mutually co-expressed, while co-expression between genes in different subnetworks are sparse.

A gene of a module commonly reflects the module behavior when compared with genes that mainly work alone, hence it may be more informative with respect to the biological process (AOKI; OGATA; SHIBATA, 2007). Modules are many times the starting point for more detailed studies because they minimize the global complexity, since they consider only part of the whole network. A variety of tools can be used to explore these modules, make functional annotation of genes (nodes) and modules and to elucidate the nature of the gene relationships (SERIN et al., 2016).

Two practical protocols of modularity analysis can be performed (AOKI; OGATA; SHIBATA, 2007):

(i) Guide-gene: this approach is characterized by selecting an appropriate set of genes relating to the biological problem based on experimental knowledge and literature information before to calculate the correlation coefficient and to construct the network. The pre-selected set of genes are called *guide genes*. Correlation coefficients between the guide genes are calculated from gene expression data. Next, modules are extracted among the guide genes. Thus, genes of interest are present in the modules.

(ii) Non-targeted: this approach is characterized by a knowledge-independent module search within entire network based on the topology of the network. This approach is necessary to achieve one of the main aims of GCNs: to infer gene function. If an unknown gene is part of a densely connected module in which other member genes were known to be involved in a certain biological process, it would be hypothesized that the unknown gene had functional relationships with that process. The discovery of novel modules that may not be obtained using the first approach is expected, because the modules are detected from the entire network, according to the topology of the connections.

There are multiple methods to find modules in GCNs. Zhang and Horvath (2005) in WGCNA define modules as groups of nodes with high topological overlap. The topological overlap of two nodes reflects their relative interconnectedness. The topological overlap matrix (TOM) provides a similarity measure, which has been found useful in biological networks. A node has high TOM-based connectivity if it has high overlap with many other nodes. They use average linkage hierarchical clustering coupled with the TOM-based dissimilarity to group genes with coherent expression profiles into modules. Thus, WGCNA identifies modules without the use of a priori guide gene, since it uses unsupervised clustering (LANGFELDER; HORVATH, 2008). Hierarchical clustering algorithms identify clusters by iteratively assigning nodes (genes) to clusters. In a first step, weights are assigned to the network vertices (for example, using the calculated correlation coefficient). Clusters are then built from high weight vertices and progressively expanded by including neighboring vertices. The number of final clusters varies, for instance depending on a chosen threshold (cutoff). They have also used the TOM-based dissimilarity in conjunction with partitioning around medoid clustering. A *Medoid* is the most centrally located object of the cluster, with minimum sum of distances to other points (JIN; HAN, 2010). Therefore, they consider that gene modules correspond to branches of the hierarchical clustering tree (dendrogram). The choice of a height cutoff to cut branches off the tree is the simplest method is to define the resulting branches correspond to gene modules. It remains an open research question how to choose optimal cutting parameters or how to estimate the number os cluster in a dataset.

Gaiteri et al. (2014) highlighted some emerging directions for GCN research beyond modules and hubs that are especially relevant to complex brain disease. The authors present an important criticism of GCNs analysis focus on patterns of networks, since the search for modules and hub genes is often treated as a black box that obscured other aspects of the networks. Usually, the identification of hub genes is disease-correlated, as putative mediators of pathology. This approach tends to focus attention on a few hub genes and ignores the many other ways in which networks can be used to develop systems biology insights into testable hypotheses. Various biological activities can influence the expression of two or more genes and yield correlated expression patterns. Thus, a network edge reflect the converging influences of genetic, biochemical and environmental factors, so are potentially source of information about the biological state of an individual. However, interpreting these networks composed of many connections is challenging because they can arise from several biological and non-biological sources that are mathematically indistinguishable. Studies suggest a possible aggregation of multi-omic data to GCNs, so it is expected the proposal of new procedures to generate novel hypotheses for disease mechanisms and changes in the interpretation of these networks.

3.1.4 Network Motifs

Shen-Orr et al. (2002) generalize the notion of *motifs*, widely used for sequence analysis, to the level of networks to uncover their structural design principles. They define *network motifs* as patterns of interconnections that recur in many different parts of a network at frequencies much higher than those found in randomized networks. The motifs allow a representation of the network in a compact, modular form: by using symbols to represent the different motifs, the network is broken down to its basic building blocks (MILO et al., 2002). Thus, the motif structure also allows an easily interpretable view of the entire known network (SHEN-ORR et al., 2002). This approach may help define the basic elements of networks which could be used for computational construction of networks. Detect and understand network motifs gains insights into their dynamical behavior and allows to define classes of networks and network homologies (MILO et al., 2002).

Milo et al. (2002) found motifs in networks from biochemistry, neurobiology, ecology, and engineering. Similar motifs were found in networks that perform information processing, even though they describe elements as different as biomolecules within a cell and synaptic connections between neurons in Caenorhabditis elegans. These motifs may have specific functions (for example, as elementary computational circuits) and they may be interpreted as structures that arise because of the special constraints under which the network has evolved. The motifs shared by ecological networks were distinct from the motifs shared by the genetic networks of

Escherichia coli and Saccharomyces cerevisiae or from those found in the World Wide Web. Motifs may thus define universal classes of networks.

Motif determination provides lots of information concerning the properties and the characteristics of a network, however it does not necessarily reveal evidence about its function and the function of its components (ALON, 2007). The fact that these patterns appear at frequencies much higher than expected at random suggests that they may have specific functions in the information processing performed by the network (SHEN-ORR et al., 2002). Their possible function could be inferred by common functions of the systems in which they appear. Besides, a mathematical analysis of their dynamics could suggest additional insights. The concept of homology between genes based on sequence motifs has been crucial for understanding the function of uncharacterized genes. Moreover, the notion of similarity between connectivity patterns in networks, based on network motifs, may be helpful in gaining insight into the dynamic behavior of newly identified gene circuits. Some motifs have been found to be associated with optimized biological functions, like in the case of positive and negative feedback loops, oscillators or bifans.

There are several studies which investigate network motifs in GCNs. Ma et al. (2013) describe a bottom-up strategy to identify gene expression modules from GCNs that are regulated by known promoter motifs. For any given motif, every gene in the network was first analyzed to calculate its probability of belonging to an expression module regulated by that motif. Then, all the top ranked genes were used to extract a subnetwork from the original GCN. From this sub-network, the modular structures will extracted, thus enabling discovery of novel signaling pathways. This approach identified novel expression modules for four well studied motifs from an Arabidopsis thaliana GCN based on the graphical Gaussian model.

Peng et al. (2007) developed novel algorithms to decompose clusters of genes into smaller ones by integrating protein domain information into the clustering algorithm. They characterized a network motif as a completely connected subgraph (i.e., a clique) in a GCN with genes encode proteins with the same combination of protein domains. They expect that a motif may be reused multiple times within organisms. They have developed algorithms for constructing labeled GCNs with corresponding protein sequence domain combinations and then detected recurring network motifs with similar protein domain memberships within these labeled networks. The motifs statistical significance was evaluated by comparing the results with those from randomized networks. The motifs biological relevance was evaluated by Gene Ontology annotations on biological processes. Their approach was applied to the malaria transcriptome. They found many motifs with three, four, and five members.

Choobdar, Ribeiro and Silva (2015) propose a analysis that combined network edges weights and network motifs. They characterized a subgraph as a motif if the weights of edges inside the subgraph hold a significantly different distribution than what would be found in a random distribution. They use the Kolmogorov-Smirnov test to calculate the significance score of the subgraph, avoiding the time consuming generation of random networks to determine statistic significance. Their approach was applied to GCNs related to three different types of cancer (lung, breast and neuroblastoma) and also to two healthy datasets. The structure of the networks is compared using weighted motif profiles (each profile is a feature vector containing all the individual weighted scores found): the results show that they are able to clearly distinguish the networks and separate them by type. They also compare the biological relevance of their approach to an unweighted motif profile. Their results of gene enrichment analysis show that weighted motifs are biologically more significant than the unweighted.

3.1.5 Differential Network Analysis

The differential network analysis combined two common approaches to the analysis of gene expression data: differential expression and networks (FUENTE, 2010). Usually, the differential network analysis involve, either explicitly or implicitly, the construction of GCNs for healthy and disease samples, followed by the comparison of the structure of the two networks providing insights into disease-specific alterations in the biological processes underlying the correlation patterns. Thus, the main goal is to identify differences in the network patterns in healthy and disease-affected samples instead of evaluating the differences in mean gene expression levels (FUENTE, 2010). Pairwise relationships between network nodes result from relationships among the genes in the biological process of interest, so identifying which of these are altered in disease-affected tissue when compared to healthy tissue is a first step in recognize dysfunctional systems.

The comparison of the structure of the two networks can be performed in several ways. Fuente (2010) says that the simplest way to accomplish this task is to look at the degree (or connectivity) of each vertex (gene) in the two networks. Genes that have a strongly altered connectivity are thought to play an important role in the disease phenotype. According to Fuente (2010), the main difficulty of this approach is establishing a threshold for each edge to be included in the network. For example, a very high correlation threshold indeed guarantees the exclusion of many spurious edges, but obviously will also exclude many relevant ones. Ideally, the threshold level is defined such that the resulting networks include as many biologically relevant edges while keeping the spurious edges low. Use the global network topology of the inferred GCNs to guide the choice is a potential effective way to select the threshold (FUENTE, 2010). Different global network topological properties have been observed in biological networks, for example the degree distribution and clustering coefficient. This analysis quantities describing the relationships between the genes such as intramodular connectivity. Differential analysis of intramodular connectivity was used, for example, to identify key differences in expression networks of human and chimpanzee brains (FULLER et al., 2007).

Fuller et al. (2007) used differential network analysis to uncover differences in the modules and connectivity between different datasets. They use body weight to arrive at two distinct datasets: lean and obese mice. Each dataset was used to construct a network. Next, the networks were contrasted to find nonpreserved modules, differentially expressed genes, and differentially connected genes. Thus, their main goal was to relate differences in gene expression profiles to phenotypic differences across different conditions (two different groups of mice). They identified the 30 mice at both extremes of the weight spectrum and constructed the first network using the 30 leanest mice and the second network using the 30 heaviest mice. They defined as a differential connectivity measure the difference between the connectivity each gene (divide by the maximum network connectivity) of the two networks, but other measures of differential connectivity could also be considered. They used the absolute value of the Student t-test statistic to measure differential gene expression between the lean and the obese mice. Their differential network analysis revealed differences in connectivity and module structure between two networks based on the liver expression data of lean and obese mice.

Iancu et al. (2013) used differential network analysis to investigate haloperidol response and non-response in three mouse populations of increasingly complex genetic structure. They goal was to detect significant changes in network structure. They used an adapted method (inspired in WGCNA method) for evaluation the preservation of network properties: they create separate networks corresponding to the two biological conditions; for any collection of network nodes of interest (module), preservation statistics are created by comparing network/module statistics against changes that could occur by chance. They computed network preservation statistics for pair of networks. The network statistics used include intramodular connectivity, total network eigengene connectivity, module eigengene connectivity, clustering coefficient and maximum adjacency ratio. These measures were applied to individual nodes and for a given module the values for all nodes are arranged in a vector. Vectors originating from two different networks are correlated. High correlation values correspond to strong preservation. Their analysis found that gene coexpression patterns changed significantly. A number of these changes were concordant across genetic backgrounds.

It is important to note that module changes and module preservation are related and complementary concepts and they can both hold for a given module (IANCU et al., 2013). Even though modules might be highly preserved across biological conditions, this does not preclude the emergence of subtle changes in network structure that are not enough to render the module non-preserved, but nevertheless are statistically significant and, potentially, biologically meaningful. Several studies focus cross-species comparison of GCNs for understanding the interplay between regulatory function and evolution and are interested in module preservation (also called consensus modules analysis (LANGFELDER; HORVATH, 2008)).

Movahedi et al. (2012) study comparative co-expression analysis in plant biology. They say that a major objective in this context is the systematic comparison of gene clusters across species using homologous or orthologous genes. Defining sequence-based of these genes is a powerful approach to link expression datasets across species and to identify genes with conserved gene functions or conserved modules that participate in similar biological processes. As most approaches use gene homology or orthology information to connect GCNs between different species, larger co-expression clusters will logically also yield a higher number of shared orthologs. The probability to have shared orthologs between co-expression clusters is also higher for genes involved in many-to-many orthology relationships compared with small families with one-to-one orthology relationships.

Serin et al. (2016) list several advantages of cross-species network comparisons in a study of different approaches applied to analyze plants. Networks of well-studied plants can enrich sparse networks, reducing the need of extensive functional genomic and phenomic resources. Besides, cross-species comparisons can accelerate the functional annotation of genes and the discovery of gene-gene interactions, consequently accelerating the gene prioritization process for targeted mutational studies. GCNs comparisons can be used also to identify functionally conserved network patterns and to study their evolution. These comparisons can be done at the global scale or focused on specific gene modules.

Hansen et al. (2014) also discuss comparative analyses. When GCNs from different species are highly similar, a whole biological pathway are conserved across species. Thus, the analysis can transfer gene function annotation from well-studied plants, to other uncharacterized plant species. Functionally equivalent genes can be also identified, since the analysis could find genes that have similar sequence and similar expression pattern across different organisms. Furthermore, a comparative analysis should have higher performance, because it eliminates the noise from coexpression analysis, since only parts of GCNs are conserved and tend to be functionally relevant. Biologically relevant associations are likely to be independently observed in the different species, whereas false associations are less likely to be repeatedly observed. Sets of genes that are conserved among multiple species are expected to play a key role in biological responses. Thus, the analysis can be thought as biologically meaningful approach to remove false positives (present due to noise in the data) and false negatives (due to missing data in one of the species).

Lee et al. (2004) present a comparative analysis of 60 large human datasets containing a total of 3924 microarrays. They look for pairs of genes that were reliably coexpressed in multiple datasets, establishing a high-confidence network that are observed in at least three datasets. They have confirmed positive correlations between genes much more often than negative correlations. The confirmation of correlated expression provides a useful way to improve the confidence in any particular correlated expression pattern and these patterns are more likely to be functionally relevant. Several criteria must be met for a link to be confirmed: (i) the pairs of genes must be present and detectably expressed in multiple datasets, so a gene that is only represented in one dataset will never have any confirmed links; (ii) a link might be sample-type specific, even if the genes are expressed in all cases. When a level of confirmation of a link increases, it is more likely that the link is between two genes that are already known to have a functional relationship. They evaluated this by examining the overlap of Gene Ontology annotations for each pair of linked genes.

Yang et al. (2014) performed a analysis of the properties of prognostic genes (key molecules informative for cancer prognosis and treatment) in networks across four cancer types (GBM, ovarian serous cystadenocarcinoma, breast invasive carcinoma, and kidney renal clear cell carcinoma) from TCGA. They used the GCNs constructed from a single type of microarray, reducing various confounding factors in the data analysis, such as prior knowledge bias. Their study revealed some distinct properties of prognostic genes. They found that prognostic mRNA genes tend not to be hub genes and this pattern is unique to the corresponding cancer-type specific network. The prognostic genes are enriched in modules, especially in module genes conserved across different cancer co-expression networks. The target genes of prognostic miRNA genes show similar patterns. Some modules enriched in various prognostic genes show cross-tumor conservation. Their study provides a systems-level understanding of the behaviors of cancer prognostic genes, thereby laying a foundation for how to incorporate the co-expression network information into prognostic modeling. The analysis across tumor types allows elucidate the common/distinct biological processes involved in different cancer types, which may facilitate novel subtype classifications.

3.2 Microarray Classification

The use of microarrays to find groups of genes that can be used to determine a diagnosis of a disease that an individual is suffering from or predict the prognosis of a course of a therapy or the result of an experiment is one of the most exciting areas of research microarray (STEKEL, 2003). The samples are taken from several groups of individuals with known diseases, outcomes or phenotypes and hybridized to microarrays. The objective then is to find a small number of genes that can predict to which group belong each individual. These genes can then be used in the future as part of a molecular test for certain individuals (using microarray or other method) (STEKEL, 2003).

Different methods can be used for partitioning space and predicting the group of a new sample. Some commonly used methods:

- N-nearest neighbors – we have a number of samples with known class membership, to classify a new sample with unknown class membership there are three steps: (i) we look at the gene expression measurements for the sample we are trying to classify; (ii) we find the nearest of the known samples as measured by an appropriate distance measure (typically Euclidean distance); (iii) the class of the sample is the class of the nearest sample (STEKEL, 2003; PARMIGIANI et al., 2003). The method is intuitive and easy to understand, there is no training time, but is not robust to outliers;
- Neural networks – are based on a model of the working of the brain – the network is organized as a series of nodes (simulating neurons), which have inputs and outputs; the output of the nodes depends on the input into the nodes; the inputs importance are determined by a set of parameters known as weights; the neural network learns by adjusting the weights (STEKEL, 2003; PARMIGIANI et al., 2003). It is trained by giving it examples of samples to be classified; the network adjusts the weights on the input of the nodes so that produces the correct output; the network is trained until it shows no improvements in predicting the classes of the training set. After the train using the samples with known classes, apply the neural network to the new individual to determine its class. There is an important training time and a complex unknown architecture involved with the nodes; neural networks are not transparent (usually called black box approach) (ASYALI et al., 2006);
- Support vector machines – they work by separating space into two regions by a straight line or hyperplane in higher dimensions, which is chosen so as to minimize the misclassification error of the method. There are three steps in applying the method: (i) project the

data from the known classes into a suitable high-dimensional space; (ii) identify a hyperplane that separates the two classes; (iii) the class of the new individual is determined by the side of the hyperplane on which the sample lines (STEKEL, 2003; PARMIGIANI et al., 2003). There is training time and does not extend naturally to analyze more than two classes.

There are many works that apply classical machine learning for classification of microarray data methods. Statnikov et al. (2005) evaluate the use of methods widely used in the classification of genomic domains. The authors conclude that random forests, support vector machines, kernel ridge regression, Bayesian logistic and regression with Laplace priors are the most effective machine learning techniques for performing accurate classification from microbiomic date.

Liu et al. (2005) present a novel method based on network topology for cancer classification. In this method, first, are built basic GRNs using the Pearson correlation. For each test sample, they add the sample into the dataset of each class and reconstruct all of the networks. Cancer type was classified according to the correlation of topological quantity between the basic and the reconstructed networks. Although Carter et al. (2004) have established that global topological properties are conserved in expression correlation and physical interaction networks, the full significance of this finding is not well understood.

Rapaport et al. (2007) construct a classifier in which the predictor variables are grouped according to their neighborhood relations in the network. They assume that the genes close on the network are more likely to contribute to the prediction function. Their hypothesis is that the genes close on the network should have similar expression profiles. However, this is only a trend, valid when one takes the average large scale, so it is possible to find many local exceptions. Thus, this method is very coarse-grained, it does not allow inferring a precise network logic but rather detects average predictors of relatively big network modules.

3.3 Summary

Topology analysis allows the characterization of networks. The centralities analysis allows one to identify the most important elements, which can provide many insights, but the focus on only a few elements can ignore other elements and their relationships that could be turned into precious insights for biologists (GAITERI et al., 2014). Some articles focus on how to detect biologically meaningful modules (ZHANG; HORVATH, 2005) and recurring patterns called motifs (ALON, 2007) in networks. Langfelder and Horvath (LANGFELDER; HORVATH, 2008), for example, do not assume prior pathway information and construct modules in an unsupervised fashion. They relate a handful of modules to the clinical trait to find clinically interesting modules, using intramodular connectivity along with gene significance to screen for significant hub genes. The differential analysis, in turn, allows the comparison of the structure of networks providing insights into specific alterations in the processes under study.

Our analysis is focused on the genes of a pathway so the goal is not to identify modules, pathways or motifs, but rather to better understand the relationships among genes of the pathway of interest and their variations on samples of diseased and control tissues. The overall goal is to get insights on how alterations in the levels of expression may affect the activation of the pathway based on target genes evaluation.

Our approach is complementary to the related works. We use also topological properties to characterize our networks. We propose the use of cycles to identify central elements, implementing centrality analysis. The genes of interest used in the application of our methodology can be obtained in a module detection. In addition, we compared the networks of healthy and diseased individuals, performing differential network analysis.

Finally, about the classification related works, although there are already many methods for classifying microarrays, most of these still lack a biological explanation associated with the samples classification, so there is still scope for defining new methods to close this gap, thus the development of our method is justified, since it includes the biological information in the classification.

4 ANALYSIS METHODOLOGY

In this chapter, we present our approach of analyzing microarray datasets, based on the different kind of cycles found among genes of the GCN constructed using quantized data obtained from the microarrays. As mentioned before, a cycle is a closed walk with all vertices distinct (except the first and last vertices) (DIESTEL, 2012) and is associated with feedback mechanisms very common in biological networks.

Our challenge with the proposal of this new method is to understand the built network. We look for the cycles, and we classify them in positive and negative, so we seek to know the structure of the network and to be able to understand the relationships established by the genes, and to identify and quantify the differences between different types of tissues. A greater understanding of the structure and the differences may help to predict models that solve or prevent the studied problem.

The proposed approach innovates by using the existing cycles in the network for analysis, instead of using the connectivity of the whole network or the intramodular connectivity as the measure of node importance as other approaches do (MA et al., 2010), thus providing a different and potentially fruitful strategy to analyze complex interactions in pathways.

Section 4.1 describes our analysis methodology. Section 4.2 relates our main results. Finally, section 4.3 presents a discussion about the method and the results.

4.1 Analysis Method

This section is divided into two parts. First, section 4.1 explains our method of analysis divided into two phases: (i) section 4.1.1 describes our approach to construct GCNs using microarray data and to obtain cycles; and (ii) section 4.1.2 details the cycle-based analysis.

4.1.1 Phase I – Graphs and Cycles

Our analysis method starts with the following steps, which are depicted in Figure 4.1, for the GCNs construction and obtaining cycles:

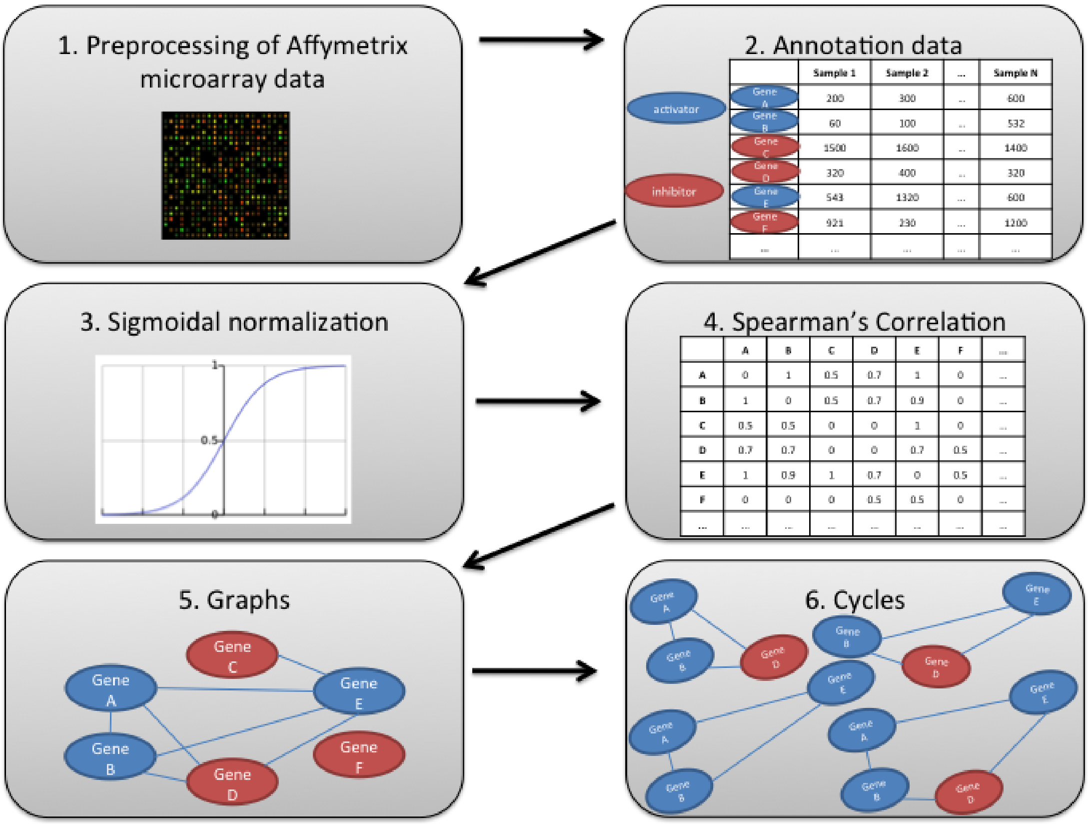


Figure 4.1 – Overview of co-expression network analysis using cycles.
Source: Author(2016).

1) Preprocessing of Affymetrix microarray data. This step consists in importing the raw microarray data and defining expression values per probe set. These values are obtained by (a) background correction - adjusting hybridization effects; (b) normalization - removing systematic errors and biases to allow comparisons among arrays; (c) summarization - combining multiple probe intensities to obtain a single value for each gene. All these operations are supported by the Bioconductor package *affy*. We used the function *ReadAffy* to import the data and the function *mas5* for defining corresponding expression values. Lim, Wang, Lefebvre and Califano (LIM et al., 2007) suggest that MAS5 provides the most faithful network reconstruction. It is worth mentioning that data from high-throughput gene expression measurements are affected by a relatively high level of noise (BOCCALETTI; LATORA; MORENO, 2010).

2) Data annotation. The purpose of the annotation is to provide detailed information about the data. These operations are performed by the Bioconductor package *annotate* and *hgu133plus2.db*. We extracted the feature names with the function *featureNames* and the genes symbols with the function *getSYMBOL*. For each of the main genes involved in the pathway of interest a data record is created, containing the feature name, the gene symbol, the gene role in the pathway (activator or inhibitor) and the expression value (calculated in step 1). The generated data records are classified in GBM samples and control samples.

3) Sigmoidal normalization. This step reduces the influence of extreme values or outliers in the data without removing them from the dataset. Expression values are nonlinearly transformed by using a sigmoidal function (PRIDDY; KELLER, 2005) and the normalized values range from 0 to 1.

4) Calculation of Spearman's correlation. Correlation is used to discover sets of genes with similar expression profiles, and have been widely used to analyze gene expression data. Genes with correlated expression profiles may correspond to protein complexes, pathways, or participate in regulatory and signaling circuits (HORVATH, 2011). Spearman's rank correlation coefficient is non-parametric and allows to identify whether two variables (genes) are related by a monotonic function. We used the function *cor* (specifying the parameter for the type of correlation as *method = spearman*) for creating the correlation matrix of expression values between the selected genes.

It is worth pointing out that GCNs are effectively only able to identify correlations (see more in Section 2.3.2). Hence, GCNs indicate which genes are active simultaneously, which often indicates they are active in the same biological processes, but do not normally confer information about causality or distinguish between regulatory and regulated genes (DAM et al., 2017). Correlation is the measurement of relationship occurring between two things. On the other hand, causation means that one thing will cause the other. Causation can also be termed as causality or cause-effect feature. GRNs are able to identify causality (see more in Section 2.3.1).

5) Generate graphs. The undirected graphs (representing the GCNs) are constructed by computing a correlation coefficient for each pair of genes. If the coefficient is above a certain threshold and is statistically significant ($p < 0.05$), the gene pair gets connected in the graph. If not, it remains unconnected. Ideally, the correlation coefficient used would be 1, but the graph generated with data from GBM samples does not show cycles with such a high coefficient, so we decided to use the highest correlation coefficient in which the GBM samples graph present at least one cycle, thus way allowing our algorithm to obtain results. We consider weighted undirected graphs that can be represented by a symmetric adjacency matrix $A = [a_{ij}]$, where i and j are nodes and represent genes, and the pairwise adjacency (connection strength) a_{ij} takes values in the unit interval, i.e., $0 < a_{ij} \leq 1$. The adjacency value of a_{ij} represents the coefficient of correlation between i and j; it is greater than zero if nodes i and j are connected and 0 otherwise. For notational convenience, we set the diagonal elements to 0. We used the R package *igraph* for obtaining graphs statistics and we used the function *graph.adjacency* for creating the graphs.

6) Cycles. In order to seek the biological explanation of the observed gene associations, we look for cycles in the gene network. A *cycle* is a path from a vertice back to itself (so the first and last vertices are not distinct) (DIESTEL, 2012). A *path*, in turn, is a walk with all vertices (and hence all edges) distinct (DIESTEL, 2012). Finally, given a graph $G = (V, E)$, where V is the set of vertices and E is the set of edges, a *walk* of length k from $v_0 \in V$ to $v_k \in V$ is a sequence of vertices $v_0 v_1 v_2 \ldots v_{k-1} v_k$ such that the adjacent pairs $v_0 v_1, v_1 v_2, \ldots, v_{k-1} v_k$ are all edges (DIESTEL, 2012). We used a C++ implementation of Johnson's algorithm (JOHNSON, 1975) to find the cycles in the graphs. After the search algorithm is completed, the repeated cycles are excluded from the analysis. Feedback mechanisms are very common in biological networks. Our hypothesis is that negative feedback induces relations among genes that may help explaining the stability of the regulatory process within the cell. Positive feedback cycles, on the other hand, reflect the amount of imbalance of a cell. The genes of interest are of two types: activators and inhibitors. We assume that a cycle of a graph is positive when the number of inhibitors in the cycle is zero. Similarly, it is said to be negative when the number of inhibitors in the cycle is greater than or equal to one.

Identifying these cycles of the graph manually (from a visual perspective) may not be an obvious or even a feasible task. Enumerating cycles is a particular case of combinatorial patterns in graphs. One of the defining features of the problem of listing combinatorial patterns is that there frequently exists an exponential number of patterns in the input graph. Other feature is explicitly generating each pattern found in the input graph. The number of cycles in a complete undirected graph with n vertices is given by equation 4.1:

$$\sum_{k=3}^{n} \frac{n!}{(n-k)!k!} \tag{4.1}$$

As n becomes big, the resulting combination can only be dealt with using the computer. In this way, an automated tool that identifies the cycles and that can analyze them is justified.

The identification of cycles is a computationally costly process. This way, it may be necessary to limit the size of the cycles to make the method computationally feasible. The maximum size to be considered depends on the size of the network being analyzed and the pathway under investigation. In our case studies, cycles of small sizes (up to size ten) were sufficient to show significant differences between the control and GBM samples. We will discuss more about this in the conclusion.

4.1.2 Phase II – Analysis

Once the graphs and cycles are obtained, we start our analysis.

1) Generating graph statistics. We use topological metrics that provide an overview of the network structure: (i) average degree – which defines the average number of edges connected to a node (a measure of how many edges are in the graph compared to the number of nodes); (ii) density – the measure of the level of edges connected in the network in relation to the total of possible connections; density is defined by a decimal value between zero and one; graphs with values close to one are typically dense graphs, while near zero are called sparse graphs; (iii) diameter – this concept refers to the measurement of the distance between the two most distant nodes in the network; it may be an indicative of how complex the network is; for example, a graph with a diameter three is usually less complex than one with a diameter seven. These measures can provide tangible numbers that support a visual assessment of the network as well as can help explore facets of the network that are less obvious from a visual perspective.

2) Analyzing cycles. Next, we look at the cycles. We perform a subgroup analysis, which refers to the search for patterns in subsets of the samples. This type of analysis aims to compare the results of each particular subgroup of samples. Our analysis considers two subsets of samples: GBM and control. Therefore, the execution of the previous section steps results in a set of cycles for each subset of samples. Each cycle is represented by a data record, containing the symbol of the genes that compose the cycle (i.e. the nodes), the cycle length (determined by the number of nodes), the absolute correlation (i.e. the product of the edges weight) and the cycle type (positive or negative). The generated data records are summarized by cycle type as well as by cycle length. Our hypothesis is that the number of positive and negative cycles (i.e., positive and negative feedback mechanisms) are related to the functionality and complexity of a given pathway machinery. To increase the confidence in the relevance of the found cycles, it is possible to consider data records from samples of different experiments. In our study we considered data records for the GBM samples in three different experiments: one available in the GEO and two in the TCGA.

3) Find target genes. Finally, we check which genes appear most frequently in cycles. We count the positive and negative cycles where each gene appears in each subset of samples and compare these values. Subsequently, we compared the means of the expression level of these genes among the subgroups using the t-test to verify if they have a statistically significant difference. The identification of these genes provides an alternative way to highlight important nodes for the network, as well as allows an information gain in relation to analysis of the gene expression level in cases in which the average difference is not statistically significant.

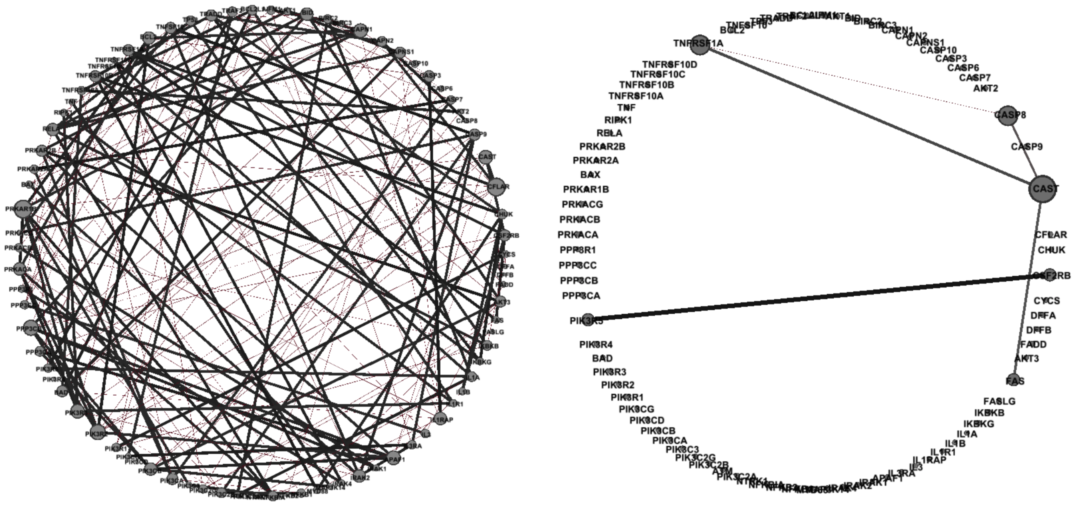
The raw data of 276 GBM samples of all histology and eight normal adult brain samples used in this study is available as experiment number GSE16011 (GRAVENDEEL et al., 2009) in the Gene Expression Omnibus (GEO)[1]. Furthermore, we analyzed GBM samples of two independent studies available on TCGA[2]: 220 samples characterized in (The Cancer Genome Atlas (TCGA) Research Network, 2008) and 520 samples characterized in (BRENNAN et al., 2013). Experimental data used in the analysis are available in AffymetrixGeneChip Human Genome U133 Plus 2.0 Array format. The analyses of Affymetrix microarray data were performed using R[3] and Bioconductor[4].

[1] http://www.ncbi.nlm.nih.gov/geo/
[2] http://cancergenome.nih.gov/
[3] http://www.r-project.org/
[4] http://www.bioconductor.org/

4.2 Results

We use the proposed methodology to analyze the genes of three networks closely related with cancer: apoptosis, glucolysis and cell cycle. The gene expression analysis comparing normal and GBM tissues was performed from previously published and characterized database comprising 8 brain samples of non-neoplastic white matter tissue compared to 276 GBM samples of all histology (GRAVENDEEL et al., 2009). The difference in the number of samples from both groups stems from the difficulty of obtaining normal control data (GRIZZLE; BELL; SEXTON, 2010). The control samples were characterized as belonging to a single cluster of gene expression level similarity, while the GBM samples were divided into twenty-four different clusters (GRAVENDEEL et al., 2009). Therefore, a higher correlation coefficient is expected between the genes of the control samples than in the GBM samples. However, one can not infer from this characterization which genes are correlated (and constitute a network), as well as one can not infer the classification of the cycles from a network in positive and negative.

We constructed several graphs for each network with different correlation coefficients for control samples and GBM samples. Figure 4.2 shows the graphs created for both control and GBM samples of the experiment available in GEO (GSE16011 (GRAVENDEEL et al., 2009)) for the apoptosis pathway with correlation coefficient of at least 0.75. We selected 86 genes as main involved in this pathway: 32 activators and 54 inhibitors. The graph of control samples has 84 nodes and 180 edges, its average degree is 4.286, its diameter is 8 and its density is 0.052. The graph of GBM samples has 84 nodes and 5 edges, its average degree is 0.119, its diameter is 2 and its density is 0.001. We found 201,942 cycles in the graph of control samples (with the search limited to cycles of no more than 10 nodes), among which three positive cycles, i.e., only three cycles have no inhibitors genes in their formation. In the graph of GBM samples, three edges connect the nodes CASP8, TNFRSF1A and CAST, forming a positive cycle, comprising activating genes without the presence of inhibitors genes. This cycle was also found in the graphs built with the data of the two experiments provided in TCGA: with the samples of the experiment described in (BRENNAN et al., 2013), the cycle was found in the graph with threshold correlation 0.45; with samples of the experiment described in (The Cancer Genome Atlas (TCGA) Research Network, 2008), the cycle was found in the graph with threshold correlation 0.35.



(a) Control samples (b) GBM samples

Figure 4.2 – Apoptosis graphs (threshold correlation 0.75). Nodes are represented as circles and correspond to genes. Connections between nodes represent a weighted undirected relationship (edge), its value represents the coeficient of correlation between genes.

Source: Author(2016).

Table 4.1 – Number of cycles of the most common gene in cycles of apoptosis pathway.

Gene	GBM Positive Cycles	GBM Negative Cycles	GBM Total Cycles	Control Positive Cycles	Control Negative Cycles	Control Total Cycles
CAST	1	0	1	1	114015	114016
TNFRSF1A	1	0	1	0	27404	27404
CASP8	1	0	1	0	0	0
PIK3R2	0	0	0	0	127453	127453
CFLAR	0	0	0	1	118662	118663
PRKAR1B	0	0	0	0	118623	118623
PPP3CB	0	0	0	0	92458	92458
BCL2	0	0	0	0	90421	90421
APAF1	0	0	0	1	83347	83348
PRKACA	0	0	0	0	79830	79830
AKT3	0	0	0	0	76269	76269
BID	0	0	0	0	74875	74875

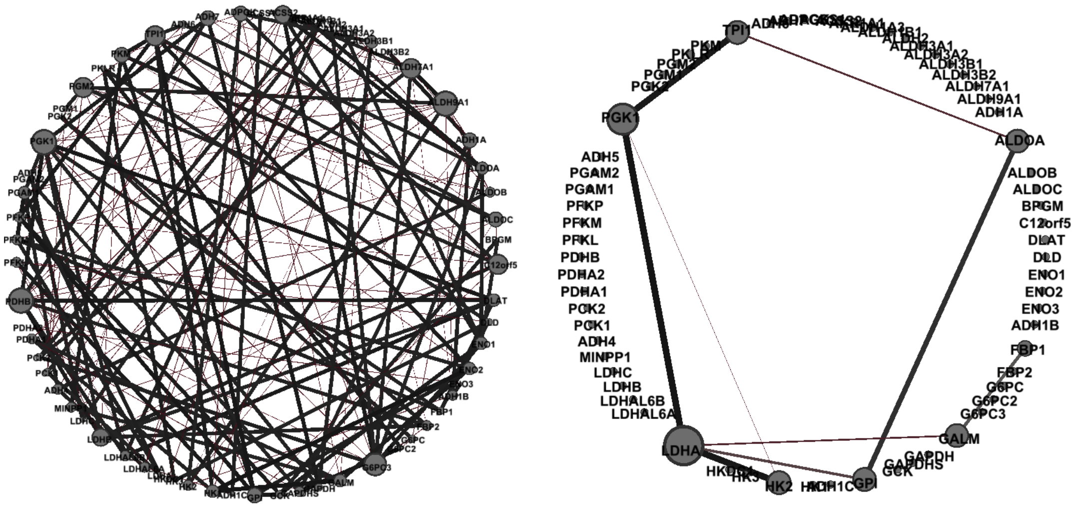
Source: Author(2016).

We list the genes that appear more often in cycles of the apoptosis network (with threshold correlation 0.75) in the Table 4.1 (sorted by frequency in the GBM graph first and, in sequence, by frequency in the control graph). One can observe that the node CASP8, which form the single cycle of the GBM graph, does not participate in the formation of any control cycle. Still about the genes of the table, we observed that a t-test with 0.05 significance level indicated that there is no significant difference between the average of the gene expression level of the control samples and the GBM samples of the genes PIK3R2 and BCL2.

Figure 4.3 shows the graphs created for both control and GBM samples of the experiment available in GEO (GSE16011 (GRAVENDEEL et al., 2009)) for the glucolysis pathway with correlation coefficient of at least 0.65. We selected 68 genes as main involved in this pathway: 67 activators and only one inhibitor gene. The graph of control samples has 67 nodes and 168 edges, its average degree is 5.015, its diameter is 7 and its density is 0.076. The graph of GBM samples has 67 nodes and 9 edges, its average degree is 0.269, its diameter is 4 and its density is 0.004. We found 460,160 cycles in the graph of control samples (with the search limited to cycles of no more than 10 nodes), among which 255,991 positive cycles and 204,169 negative cycles. In the graph of GBM samples, we found three positive cycles: (i) a cycle with three nodes; (ii) a cycle with five nodes; and (iii) a cycle with six nodes. The cycle formed by the nodes GPI, TPI1 and PKM is found in the control graph with threshold correlation 0.65 and in the GBM graph with threshold correlation 0.55. In relation to the cycles of GBM samples, the cycle formed by the nodes LDHA, TPI1 and PGK1 appear in the analysis of the three experiments: GEO (GSE16011 (GRAVENDEEL et al., 2009)) and TCGA (The Cancer Genome Atlas (TCGA) Research Network, 2008) in the graph with threshold correlation 0.60 and TCGA (BRENNAN et al., 2013) in the graph with threshold correlation 0.55.

We also list the genes that appear more often in cycles of the glucolysis network (with threshold correlation 0.65) in the Table 4.2 (sorted by frequency in the GBM graph first and, in sequence, by frequency in the control graph). One can observe that the node C12orf5, only inhibitor gene from this pathway, does not participate in the formation of any GBM cycle. Still about the genes of the table, we observed that the t-test with 0.05 significance level indicated that there is no significant difference between the average of the gene expression level of the control samples and the GBM samples of the genes TPI1, ALDOA, PKM and GAPDH.

Figure 4.4 shows the graphs created for both control and GBM samples of the experiment available in GEO (GSE16011 (GRAVENDEEL et al., 2009)) for the cell cycle pathway with correlation coefficient of at least 0.85. We selected 133 genes as main involved in this



(a) Control samples (b) GBM samples

Figure 4.3 – Glucolysis graphs (threshold correlation 0.65). Nodes are represented as circles and correspond to genes. Connections between nodes represent a weighted undirected relationship (edge), its value represents the coeficient of correlation between genes.

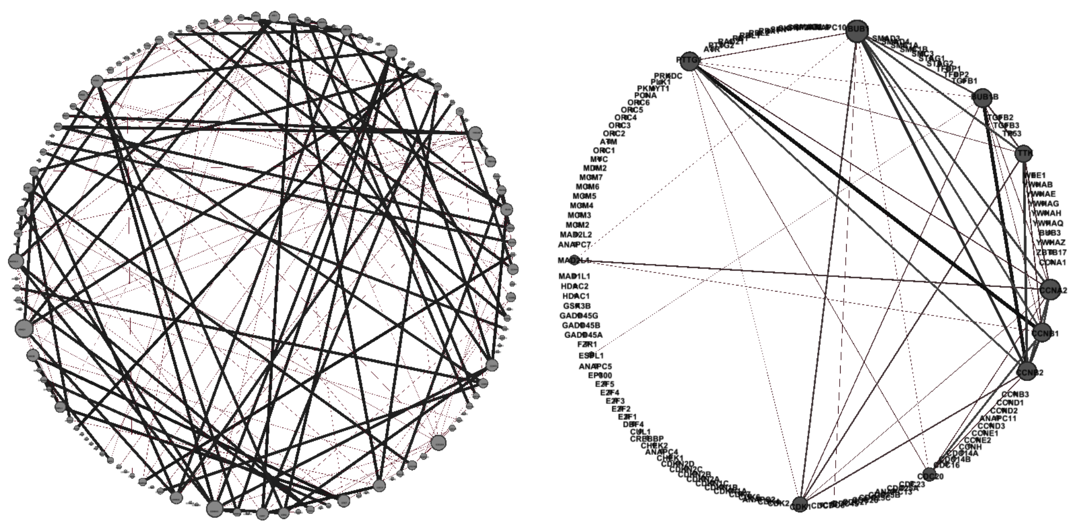
Source: Author(2016).

Table 4.2 – Number of cycles of the most common gene in cycles of glucolysis pathway.

Gene	GBM Positive Cycles	GBM Negative Cycles	GBM Total Cycles	Control Positive Cycles	Control Negative Cycles	Control Total Cycles
PGK1	3	0	3	171596	93845	265441
LDHA	3	0	3	9440	4220	13660
TPI1	2	0	2	75397	99482	174879
GPI	2	0	2	40506	75450	115956
ALDOA	2	0	2	53808	19205	73013
HK2	2	0	2	33985	26689	60674
PDHB	0	0	0	148509	81465	229974
ALDH9A1	0	0	0	141844	84857	226701
LDHB	0	0	0	97470	111253	208723
ALDOC	0	0	0	94589	112999	207588
C12orf5	0	0	0	0	204169	204169
ALDH7A1	0	0	0	143223	54655	197878
G6PC3	0	0	0	130742	50482	181224
PGM2	0	0	0	66300	111277	177577
ADH7	0	0	0	107314	58695	166009

Source: Author(2016).

pathway: 96 activators and 37 inhibitors. Nine genes have dual function (are activators and inhibitors). The graph of control samples has 122 nodes and 140 edges, its average degree is 2.295, its diameter is 17 and its density is 0.019. The graph of GBM samples has 122 nodes and 37 edges, its average degree is 0.607, its diameter is 3 and its density is 0.005. We found 27,863 cycles in the graph of control samples, among which 152 positive cycles and 27,711 negative cycles. In the graph of GBM samples, we found 39,508 cycles, among which 2,679 positive cycles and 36,829 negative cycles. Genes with dual function were found only in cycles of the control graph, which may suggest a greater level of control. We noticed that the percentage of positive cycles is greater in GBM graph than in the control graph, suggesting that cell cycle pathway genes are showing an unbalanced behavior, without inhibitors genes there is no negative feedback and activators genes are unregulated. Still in relation to the cycles of GBM samples, 240 cycles appear in the analysis of the three experiments.



(a) Control samples (b) GBM samples

Figure 4.4 – Cell cycle graphs (threshold correlation 0.85). Nodes are represented as circles and correspond to genes. Connections between nodes represent a weighted undirected relationship (edge), its value represents the coeficient of correlation between genes.

Source: Author(2016).

4.3 Discussion

In this study we aimed to detect stable cycles with the basic genes of three networks closely related with cancer – apoptosis, glucolysis and cell cycle – in GBM patients – which are known to carry aberrant activation in these pathways but not completely understood mechanisms – comparing them with controls to further understand its physiopathology. Our main strategy is based on the idea that a network must present stable feedback connections. Hence, we look for these stable connections among genes after pre-processing. Our main results show that the stoichiometric relationship between genes involved in apoptosis, glucolysis and cell cycle pathways is unbalanced in GBM samples. This dysregulation can be measured and explained by the identification of a higher percentage of positive cycles in these networks, comprising activators genes without the presence of inhibitors genes. This conclusion helps to understand more about the biology of this tumor type.

The results evidence the differences between the GCNs of the control and GBM samples. The cycles of the control graphs use about all the genes of each network, while the cycles of the GBM graphs use a small group of genes of each network. In apoptosis, only a few cycles were found in the GBM graph, which would indicate that the cell cannot die (HANAHAN; WEINBERG, 2011). In glucolysis, LDHA is the gene which encodes an enzyme essential in the process of glycolysis metabolism for formation of lactic acid. In the control graph, this gene appears in negative cycles (more regulated, since this cycles have inhibitor). In the GBM graph, it appears in positive cycles (dysregulated, without inhibitor), which may be associated with accelerated metabolism of glucolysis in the tumor and, besides, generating acidification in the tumor, which is well-known to accelerate tumor growth (HANAHAN; WEINBERG, 2011). In the cell cycle, a greater number of cycles has been found in GBM, which might indicate that the tumor has more active cell cycle mechanisms, since it is more proliferated (HANAHAN; WEINBERG, 2011). Analyzing the most common genes found in the cycles, we observed that the t-test with 0.05 significance level indicated that there is no significant difference between the average of the gene expression level of the control samples and the GBM samples of some genes of the three pathways. Thus, there is a new valuable information available with the analysis using cycle with respect to analysis of the gene expression level, since cycles highlight the difference between the control and the GBM samples.

Unfortunately, we have a small control sample size, so the results may be underpowered, and as a consequence some potentially important relationships may be overlooked. Further studies should include a larger number of samples to confirm these results. The formation of this group of samples raises the ethical issue of sampling normal brain tissue. Access to normal live brain tissue is very limited, because typically it is not removed during routine surgery (GRIZZLE; BELL; SEXTON, 2010). Hence, this tissue is not available to be used as control sample under all circumstances, consequently the control sample commonly used in GBM studies include brain tissue obtained during surgery for intractable epilepsy or from the margins of the resection cavity during GBM surgery, with the informed consent of the patient (LEMéE et al., 2013). However, the use of tissue adjacent to the tumor as control should be avoided, since there is the possibility of tumor cell infiltration. On the other hand, GBM specimens are easily obtained, since they were obtained from excess tissues collected from patients undergoing surgical resection of their tumors (GRAVENDEEL et al., 2009).

Several studies have even described GBM heterogeneity at the cellular and molecular level (LEMÉE; CLAVREUL; MENEI, 2015). Its heterogeneity is a hot topic in neuro-oncology. Glioblastoma is multiforme as the name implies (HOLLAND, 2000). It is multiforme in many levels: grossly, showing regions of necrosis and hemorrhage; microscopically, with regions of pseudopalisading necrosis, pleomorphic nuclei and cells, and microvascular proliferation; and genetically, with various deletions, amplifications, and point mutations leading to activation of signal transduction pathways downstream of receptors (HOLLAND, 2000). These tumors also show intratumoral genetic heterogeneity (LEMÉE; CLAVREUL; MENEI, 2015). GBM heterogeneity is one of the reasons for its resistance to therapeutic intervention (HOLLAND, 2000). This heterogeneity also explains the lower correlation coefficients observed between the genes in the GBM samples. Therefore, an interesting possibility of further work is to analyze the cycles in GBM samples subdivided into histological and molecular subgroups defined in (GRAVENDEEL et al., 2009) and (BRENNAN et al., 2013) . It is expected that the most aggressive groups have lower correlation coefficients than the less aggressive, so possibly present less cycles as well.

Tumor diversity is also a big problem in analysis based on classical samples generated from large number of cells using gene-expression profiling. Samples are generally prepared from a mixture of different cells that are present in unknown proportions. Tumors are unique in cell composition and they contain different tumor cells, all with differences in growth, migration and survival capacity and this heterogeneity confounds the expression profile (STåHLBERG; KUBISTA; ÅMAN, 2011). Bulk measurements destroy crucial information about cell-to-cell variability by averaging signals from individual cells together. These measurements are fundamentally constrained by averaging. The result of this averaging may mask information about mRNA that are present only in a subset of cells in the tissue, as their impact will be diluted. Thus accurately defining the cell types and states in our bodies and explaining how they arise in development and disease demands single cell measurements. The capability to select and analyze single cells constitutes a major step forward in tumor analysis. Today, gene expression profiles of single cells can be measured with high precision and accuracy, identifying different cell types as well as revealing heterogeneity among cells of the same kind. For GBM and for normal live brain tissues this has huge implications. Single cell would make possible to assess a cell type overall contributions to the transcriptome and how transcriptome of a certain cell type varies across different region of a tumor or brain (HOLLAND, 2000). It will also provide individual RNA information about each cell type. This is particularly important for brain studies, where cell types of interest are often present in low numbers and have a huge intra cell type gene expression variation (TRAPNELL, 2015). The use of single cell data in our method is a very exciting further work possibility.

This thesis does not describe a new software or method for constructing networks. Our challenge is not to find the network. Many studies have been developed in this direction and today the network can be found easily (LANGFELDER; HORVATH, 2008; LEE; TZOU, 2009; HACHE; LEHRACH; HERWIG, 2009; ALTAY; EMMERT-STREIB, 2010; BUTTE et al., 2000; MARGOLIN et al., 2006; FAITH et al., 2007; MEYER et al., 2007). Instead, our challenge is to understand the network: we explore the structure of the network to be able to understand better the relationships established among genes and identify and quantify differences between GBM and control networks. This greater understanding of the structure and the differences can help predict models that solve or prevent the problem (the tumor).

The most important step of our analysis method is the identification of cycles and its classification into positive and negative cycles. There is an intuitive relationship between a cycle (structure in the graph) and a biologically important concept (feedback). The cycle component is responsible for most of the dynamical properties of the whole network. Cyclic dependencies are associated with many fundamental properties of living systems, such as homeostasis, robustness, excitability, multistationarity and biological rhythms (THIEFFRY; THOMAS, 1998; DAS et al., 2009). Although our method analyze a static model, the analysis of the cycles related to the function (activation and inhibition) of the genes - positive and negative feedback - suggests the dynamical operation of the network.

A full enumeration of cycles is computationally costly. Johnson's algorithm (JOHNSON, 1975) is the most efficient variant and has a time complexity that is proportional to the number of cycles in the graph where the proportionality constant is the number of nodes and edges. The algorithm is linear in the output size, but exponential in the input size because the number of cycles can increase exponentially with network size (especially with the number of edges). Although the identification of the cycles can be costly, their association with feedback mechanisms could allow a greater understanding of the stability or imbalance of the process in the cell in a given state, thus justifying the use of the method. Depending on the pathways under investigation, it is possible to obtain significant differences between disease and control samples even considering only cycles of small sizes (up to size ten), which could be computed in feasible time (depending on the size of the network).

Our method allows to contrast one network with another network (for example, control *versus* GBM). This differential network analysis can be used to identify changes in cycles patterns in different conditions. Moreover, one can find shared cycles between two or more networks (consensus cycles analysis). Consensus cycles may represent fundamental structural properties of the network. Our method of analysis and classification is not dependent on network construction method, so another method could be used for the network construction. Thus, our methodology could be used as a complement to modularity analysis. After identifying modules of GCNs, each module can be taken as a new network, so one can search for cycles in these networks to identify the relationship between the genes and better understand the role of a given gene in a module of interest.

It is noteworthy that the proposed methodology is generic and can be used to study other networks. Concentrations of key intracellular proteins that are prevalent in a diverse range of important cellular processes are controlled by negative feedback loops (STURROCK, 2013). Inflammation, meiosis and NFκB are examples of these. Thus, our approach for analyzing microarray data using cycles of networks can be used for studying, for example, these processes.

5 CLASSIFICATION METHODOLOGY

In this chapter, we present our approach for classifying microarray samples in tissues of the most agressive type of brain tumor (GBM) and in healthy tissues, based on the different kind of cycles found among genes of the GCN constructed using quantized data obtained from the microarrays. Diestel (2012) defines a cycle as a closed walk with all vertices distinct (except the first and last vertices). We hypothesize that a cycle is associated with feedback mechanisms very common in biological networks.

Our challenge in proposing a new classification method is to associate a biological explanation with the classification of a sample. We use the cycles found in the networks, as well as their classification in negative and positive, to classify a sample. Hence, if a sample is classified in a given group of individuals, it means that this sample has the cycles that represent the group, and is thus associated with the feedback mechanisms related with those cycles. Therefore, we associate an biological information gain with our classification.

Section 5.1 describes our classification methodology. Section 5.2 relates our main results. Finally, section 5.3 presents a discussion about the method and the results.

5.1 Method

Our approach for sample classification of microarray datasets use the different kind of cycles found among genes of the co-expression networks constructed using quantized data obtained from the microarrays. Initially, we need to apply the methodology described in the previous chapter (Chapter 4). These steps are required to build the network (graph) and to identify the cycles that can be used for classifying the samples of the input microarray. Our classification method continues with the following steps, which are depicted in Figure 5.1:

1) **Cycles selection.** This step selects the cycles with the highest absolute correlation in the control samples graph and in the GBM samples graph (we use at most ten cycles, because we did not observe that more cycles decrease the absolute error; as future work, the ideal number of cycles can be set according to the obtained error).

2) **Cycle characterization by an equation system.** Each selected cycle is associated to an equation system. Each vertex (gene) of the cycle is represented by an equation, that is defined as a regression function, because each edge indicates the existence of a significant correlation between two vertices (genes). The correlation determines that there is a relationship

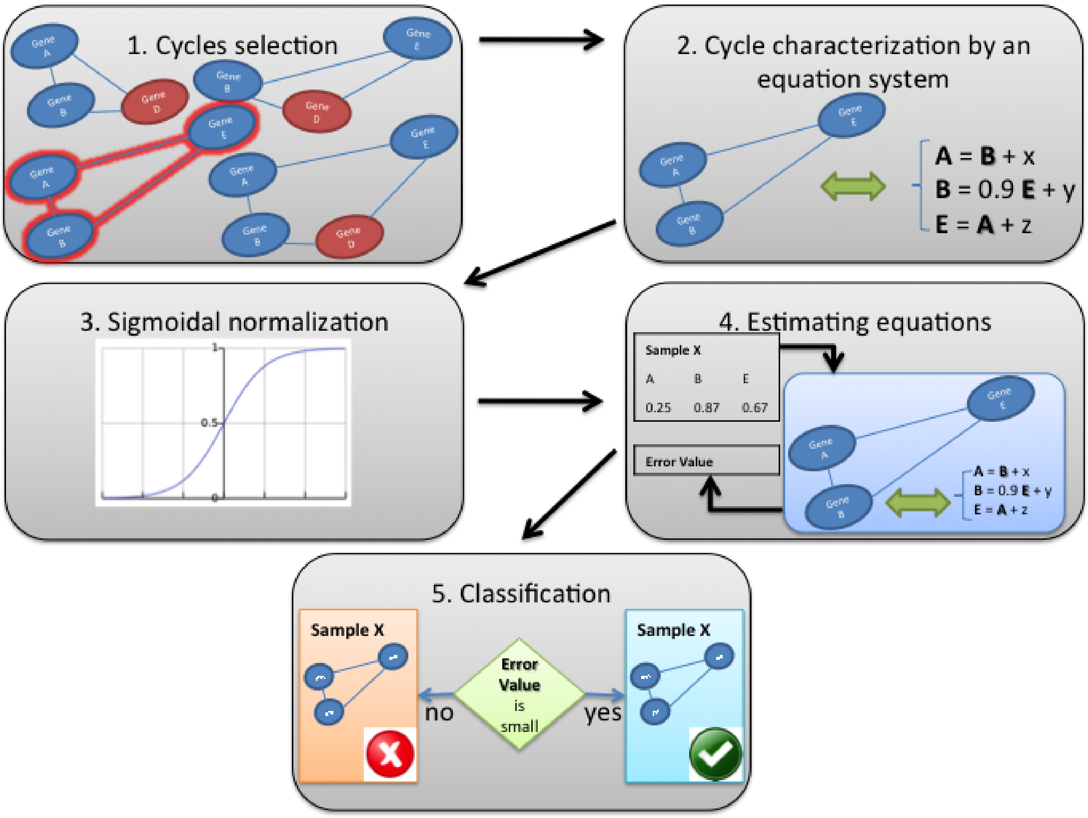


Figure 5.1 – Overview of microarray samples classification using cycles.
Source: Author (2016).

between the behaviors of the two genes; the regression determines how strong the relationship is.

The simple linear regression model for the variables X and Y is written

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{5.1}$$

where y_i is the response variable for the i^{th} observation of Y; x_i is the i^{th} observation of independent variable X; ϵ_i is the error for the i^{th} observation, that is, the effect of factors that are affecting the Y observation randomly; and β_0 and β_1 are coefficients that must be estimated. In our methodology, Y and X are variables that represent genes (vertices) of the cycle; y_i is the i^{th} observation of gene expression of the gene Y and x_i is the i^{th} observation of gene expression of the gene X. The goal is to estimate the coefficients β_0 and β_1. The most well-known method of estimating these coefficients is to use ordinary least squares (OLS). OLS provides estimates of β_0 and β_1 by minimizing the sum of the squared deviations of the y_i for all possible i. Specifically, the sum of the squared residuals is minimized when the OLS estimators of β_0 and β_1 are

$$b_0 = \bar{y} - b_1\bar{x} \tag{5.2}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sum_{i=1}^{n}(x_i - \bar{x})} \tag{5.3}$$

respectively. Note that the estimated regression function is written as

$$\hat{y}_i = b_0 + b_1 x_i \tag{5.4}$$

We use an adaptation of the method of least squares to define the coefficients:

$$b_0 = \bar{y} - b_1\bar{x} \tag{5.5}$$

$$b_1 = \text{coefficient of correlation between Y and X} \tag{5.6}$$

Note that Y and X are genes (vertices) of the cycle and the coefficient of correlation is associated to the edge between Y and X. $\bar{y}$ and $\bar{x}$ indicate the mean of all samples values of the gene expression, so the coefficient b_0 is estimated based on the gene expression values from the cycle of the normalized samples; and i denotes the sample we want to classify.

3) Sigmoidal normalization. For the classification the data needs also to be normalized, because the microarray data was normalized for the construction of the networks. The data is normalized in two ways: (i) considering the mean and standard deviation of the control samples and (ii) considering the mean and standard deviation of GBM samples. This is necessary because we do not know *a priori* if the sample being classified is from healthy individual or patient, the goal of this method is to make this differentiation.

4) Estimating equations. The estimation of the equation system of all selected cycles is performed using the normalized data generated in the previous step. As the system of equations is cyclical, some iteractions are made to achieve system convergence (we define the number of iteractions as 7, because we observed a decrease in the absolute error with this number; as future work, the number of iterations can be set according to the obtained error). At each iteraction the percentage error is computed. After the last iteration, we calculate the mean of the absolute error of the equations that compose the system.

5) Classification. This last step is based on the error computed in the previous step. If the error is small (different thresholds should be tested to obtain more robust results; the user can set this threshold), we consider that the sample has the cycle. But to evaluate the error, we must also consider that the data were normalized in two ways, so we have the possibilities shown in Table 5.1.

Table 5.1 – Classification possibilities.

Sample normalized as	Small errors in	Large errors in	Sample classification
control	control cycles	GBM cycles	control
control	GBM cycles	control cycles	GBM
GBM	control cycles	GBM cycles	control
GBM	GBM cycles	control cycles	GBM

Source: Author (2016).

As presented in the previous chapter, the raw data of 276 GBM samples of all histology and eight normal adult brain samples used in this study is available as experiment number GSE16011 (GRAVENDEEL et al., 2009) in the Gene Expression Omnibus (GEO)[1]. Furthermore, we analyzed GBM samples of two independent studies available on TCGA[2]: 220 samples characterized in (The Cancer Genome Atlas (TCGA) Research Network, 2008) and 520 samples characterized in (BRENNAN et al., 2013). Experimental data used in the analysis are

[1] http://www.ncbi.nlm.nih.gov/geo/
[2] http://cancergenome.nih.gov/

available in AffymetrixGeneChip Human Genome U133 Plus 2.0 Array format. The analyses of Affymetrix microarray data were performed using R[3] and Bioconductor[4].

5.2 Results

We use the proposed methodology to classify the samples of a microarray in tissues of the most aggressive type of brain tumor (GBM) and in healthy tissues. We use the GCNs construct to analyze the genes of the three pathways present in the previous section. The samples classification in normal and GBM tissues was performed from previously published and characterized database comprising 276 GBM samples of all histology compared to 8 brain samples of non-neoplastic white matter tissue.

We compared the performance of our method to a neural network, since it is a commonly algorithm used for classification in data science. Our approach uses the cycles as an *a priori* knowledge. The neural network, in other hand, does not use network information. The neural network was implemented in R using the *nnet* package, which lets one to construct neural networks with three layers: an input layer where features are fed in, an output layer with one neuron per class, and one hidden layer of sigmoid function neurons. We implement the neural network using the *nnet* function. We performed the classification of the microarray samples using the predict function. For our training set we use 70% of the dataset (284 elements): 199 samples. The validation set is composed by the remaining 30%: 85 samples. Our method achieves the same performance metrics that the implemented neural network.

The performance of the methods was evaluated by computing the percentages of Sensitivity (SE), Specificity (SP) and Accuracy (AC). We used the same definitions presented in (SARHAN, 2009). The definitions are as follows: sensitivity – is the fraction of real events that are correctly detected among all real events; specificity – is the fraction of nonevents that has been correctly rejected. Sensitivity, specificity and accuracy of prediction have been calculated according to the following formulas, where FP is control samples predicts as GBM; TP is GBM samples predicts as GBM; FN is GBM samples predicts as control; and TN is control samples predicts as control:

$$SE = \frac{TP \times 100}{(TP + FN)} \tag{5.7}$$

[3]http://www.r-project.org/
[4]http://www.bioconductor.org/

$$SP = \frac{TN \times 100}{(TP + FN)} \quad (5.8)$$

$$AC = \frac{(TP + TN) \times 100}{(TP + FP + TN + FN)} \quad (5.9)$$

The metrics of performance of each method are shown in Table 5.2. The total number of samples considered for both methods was 85 (30% of the total samples of the microarray). One can observe that the metrics of our method and the metrics of the implemented neural network are the same.

Table 5.2 – Performance metrics of the classification methods.

Method	Pathway	FP	TP	FN	TN	SE	SP	AC
Our Method	Apoptosis	3	82	0	0	100	0	96.47
Neural Network	Apoptosis	3	82	0	0	100	0	96.47
Our Method	Glucolysis	0	85	0	0	100	0	100
Neural Network	Glucolysis	0	85	0	0	100	0	100
Our Method	Cell cycle	1	84	0	0	100	0	98.82
Neural Network	Cell cycle	1	84	0	0	100	0	98.82

Source: Author (2016).

5.3 Discussion

Regarding the result of the classification method, we conclude that the metrics of our method and of the implemented neural network are the same. As a major advantage of the method over the neural network, we can point the information associated with the classification. In our method, when an individual is placed in a certain group, we know that it has the cycles used for the classification of this group, so there is biological information included in the classification. In neural networks, on the other hand, the classification is performed using formulas that are like black boxes. In addition, despite the parity in the diagnosis of the disease, our method is superior to a neural network, because it provides information gain in the data analysis. As mentioned before, our method allows classification and further helps to explain the tumor biology: if the analyzed set of genes is related to the disease, the found cycles may help to explain the action of the disease, since they are associated with feedback mechanisms. Thus, the cycles could provide information to develop laboratory experiments as well as to identify potential target genes for novel medications.

Our method can also be used to analyze and classify samples of GBM with different histological types. The conventional method of cancer diagnosis uses a combination of clinical and examination of the morphological appearance of stained tissue specimens under light microscopy. However, this method is subjective and depends on highly trained pathologists. Microarrays datasets may allow a more objective and highly accurate cancer classification. The use of microarray datasets to determine a diagnosis of a disease that an individual is suffering from or predict the prognosis of a course of a therapy or the result of an experiment is one of the most challenging areas of microarray research (STEKEL, 2003).

6 NFκB PATHWAY ANALYSIS

In this work, as mentioned previously, we have a close collaboration with researchers from the Department of Biochemistry of the Universidade do Rio Grande do Sul (UFRGS). This group studied extensively the NFκB pathway, which is topic of a series of their publications: (ZANOTTO-FILHO et al., 2009; ZANOTTO-FILHO et al., 2010; ZANOTTO-FILHO et al., 2011; ZANOTTO-FILHO et al., 2012a; ZANOTTO-FILHO et al., 2012b). So we chose, together with the biochemistry group, the NFκB pathway to apply our analysis methodology.

The genes of the NFκB pathway are involved in the control of a large number of biological processes ranging from inhibition of apoptosis to metastasis in cancer. GBM patients carry aberrant NFκB activation, but the pathway mechanisms are not completely understood.

We present a NFκB pathway analysis in tissues of the GBM tumor and in healthy tissues using our method. Section 6.1 describes the NFκB pathway. Section 6.2 presents an overview of the analysis. Section 6.3 relates our analysis results. Finally, section 6.4 presents a discussion about these results.

6.1 NFκB

NF-κB is a transcription factor formed by a group of 5 genes (NFKB1, NFKB2, RELA, RELB, REL) codifying for the REL family of proteins (NFκB1, NFκB2, p65\RelA, RelB and c-Rel, respectively), which are involved in the control of many of biological processes ranging from inhibition of apoptosis to pro-apoptotic effects, as well as controlling other important processes as inflammation, invasiveness and metastasis in cancer. These so diverse branches of action are due to the ability of the NFκB family proteins to form homo and heterodimers, which can modulate different group of genes by exerting inducer or repressor activities. While p65, c-Rel and RelB proteins are synthetized in the mature form, the NFκB1 and NFκB2 proteins are synthesized as large precursors, p105 and p100, which undergo processing by proteasome to generate the mature NF-κB subunits, p50 and p52, respectively (ORLOWSKI; JR, 2002; AGGARWAL, 2004; BAUD; KARIN, 2009).

The most well-described and studied NFκB dimer is p65/p50, which is frequently described as a promoter of gene transcriptional activation. Unlike RelA, RelB, and c-Rel, the p50 and p52 NFκB subunits do not contain C-terminal transactivation domains, but play pivotal functions in modulating the specificity of NFκB function, since both participate in target gene transactivation by forming heterodimers with RelA, RelB, or c-Rel. In addition, p50 and p52 homodimers are also able to bind to the nuclear protein Bcl-3, and such complexes may act as transcriptional repressors or activators in different cell contexts. Although the mechanisms by which p65/p50 dimers drive gene expression are well described, the cellular functions of other NFκB complexes remains to be completely elucidated (BARKETT; GILMORE, 1999; BASAK; SHIH; HOFFMANN, 2008; BAUD; KARIN, 2009).

There are two well-defined mechanisms for NFκB activation (shown in Figure 6.1): a canonical and a non-canonical. In the canonical mechanism, NFκB is sequestered in the cytosol as an inactive ternary complex formed by p65/p50 proteins bounded by a family of inhibitory proteins named IκBs (Inhibitor of κB). Although the IκB family consists of IκBα (NFKBIA), IkBβ (NFKBIB), and IKBϵ (NFKBIE) and Bcl-3 (BCL3 gene), the most-studied is IκBα. Activation of the NFκB is initiated by degradation of IκB proteins, which occurs via activation of IKK (IκB kinase). IKK complex is composed of a heterodimer of the catalytic IKKα and IKKβ subunits (CHUK and IKBKB genes, respectively) and a key regulatory protein named NEMO (NFκB essential modulator) or IKKγ (IKBKG gene). Stimuli as TNFα, cytokines, AGE (Advanced Glycation End-products) and lipopolisacaride (LPS) as well as intracellular stimuli as reactive oxygen species can induce phosphorylation of IKK proteins. Once activated, IKKα phosphorylates IkBα at ser32 and 36 residues promoting its ubiquitination and subsequent proteasomal degradation. This process releases active NFκB, which migrates into nuclei and regulates gene expression. Active p65/p50 NFκB can modulate a variety of anti/proapoptotic (bcl-2, bcl-xL, XIAP, survivin, cIAP1/2), inflammatory (IL1β, IL-8, TNFα) and invasiveness/angiogenic (VEGF, MMP2, MMP9, TWIST) mediators ultimately assigning to NFκB an important role in several normal and pathological conditions as inflammation, cancer and diabetes. The pathway shutdown comes from p65/p50-dependent induction of IκBα, which binds to nuclear/active NFκB and drives it back to cytosol (ORLOWSKI; JR, 2002; AGGARWAL, 2004; BAUD; KARIN, 2009).

Non-canonical NFκB signaling involves the NIK (NFκB inducing kinase)-mediated activation of RelB:p52 heterodimers. In this pathway, activation of the NIK upon receptor activation led to the phosphorylation and proteasomal processing of NFκB2/p100 into p52 in an IKKα-dependent manner. Then p52 dimerizes with RelB to pop up as a nuclear RelB:p52complex, which harbor DNA binding activity and regulate a specific subset of genes (NAKANISHI; TOI, 2005; BASAK; SHIH; HOFFMANN, 2008; TCHOGHANDJIAN et al., 2013). Although much has to be discovered on the non-canonical pathway, a select set of stimuli as lymphotoxin-α, RANKL and BAFF are described to activate this route. Though RelB proteins expression was

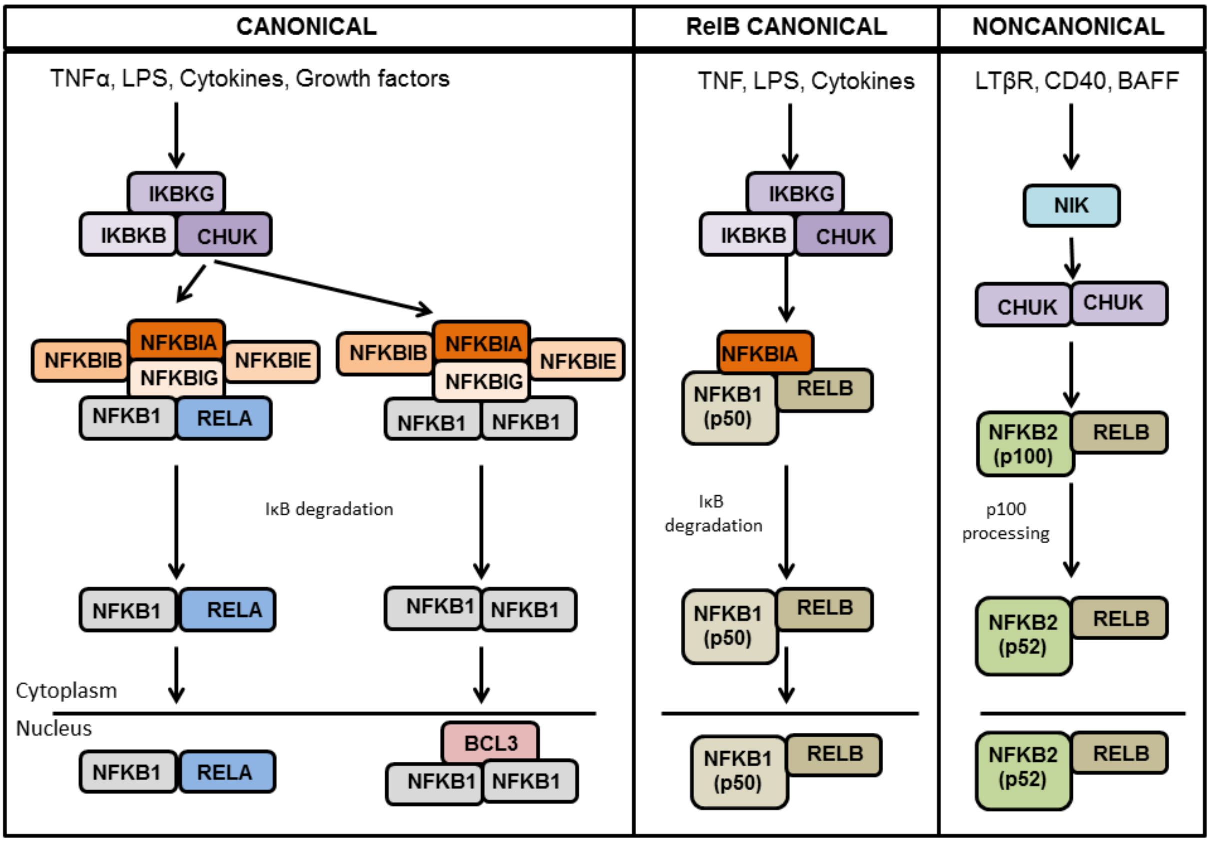


Figure 6.1 – Canonical, RelB-mediated canonical and Non-canonical mechanisms of NFκB activation. In the canonical mechanisms, NFκB dimers (NFKB1:RELA; NFKB1:NFKB1 or NFKB1:RelB) are constitutively inhibited by binding of inhibitory proteins (IκB family genes: NFKBIA, NFKBIB, NFKBIE, NFKBIG). Upstream stimulation by LPS, cytokines among others causes IKK family proteins activation (genes: IKBKB, CHUK and IKBKG), which phosphorylates IκB members to release NFκB dimers. These dimers enter the nuclei to modulate a wide range of genes involved in inflammation, invasion and antiapoptotic defenses. Non-canonical pathways is better described during immune system development, and it involves NIK (NFκB inducing kinase)-dependent activation of CHUK homodimers, which induce the processing of NFKB2 (p100) to form p52:RelB active transcription factor.

Source: Author (2016).

initially described to be exclusive of lymphoid tissues and lymphoid organ development, recent evidence has pointed out an involvement of RelB in the physiopathology of certain types of cancers as GBMs (TCHOGHANDJIAN et al., 2013). In contrast to previous beliefs on the functional independence of canonical and non-canonical pathways, recent studies have weighed on the evidence that expression of constituents of the non-canonical pathway (RelB and p52) is controlled by the canonical p65/p50 signaling. It suggests that an integrated NFκB system network relies on activation of both p65 and RelB containing dimers and that a dysregulation of canonical pathway will affect cellular responses through the non-canonical pathway (TCHOGHANDJIAN et al., 2013).

Interestingly, dysregulation of NFκB system has long been reported in several cancer cells types, suggesting that the classical mechanism controlling activation of this pathway is impaired in tumors (AGGARWAL, 2004). Mutations in IKK proteins have been described but whether these phenomena are sufficient to cause NFκB dysregulation is questionable, once many tumors display NFκB overstimulation even in the absence of IKK mutations. It was recently described that GBM cells overexpress miRNA-30e* which binds to and promotes degradation of IKBα mRNA, thus sustaining a prolonged NFκB activation in these tumors (JIANG et al., 2012). To complete the myriad of described mechanisms affecting NFκB activation, MAPKs (ERK, p38 and JNK1/2), PKC- and Akt-mediated phosphorylation of IKK and/or p65 as well as p65 acetylation also modulate its activation, nuclear translocation and the transcriptional activity of p65/p50 NFκB driving different levels of transcriptional activity for NFκB as well as making the study of NFκB pathway an extremely complex task, which is frequently incomplete (ORLOWSKI; JR, 2002; BASAK; SHIH; HOFFMANN, 2008; BAUD; KARIN, 2009).

When the transcription factor NFκB is active, there is a group of classical target genes that also increase and exert inflammatory, metastatic and antiapoptotic responses in normal and cancer cells (RAYCHAUDHURI et al., 2007; ZANOTTO-FILHO et al., 2011; JIANG et al., 2012; TCHOGHANDJIAN et al., 2013). The target genes are, in fact, an indirect measure of the activation of the transcription factor. Among the typical target genes of the NFκB transcription factor are CCL2, ICAM1, IL1A, IL1B, CXCL8 (IL8), MMP2, MMP9, SELE, TNFα and VCAM1, which are genes responsible for inflammation and cell invasion in GBM.

6.2 Analysis Overview

We present a NFκB pathway analysis in tissues of the most aggressive type of brain tumor (GBM) and in healthy tissues, based on the different kind of cycles found among genes of a gene co-expression network (GCN) constructed using quantized data obtained from the microarrays. The input of the analysis method is formed by the raw data, the basic set of NFκB pathway genes and the function (activator or inhibitor) of these genes. The output of the method is a set of cycles formed with the NFκB pathway genes and present in the built networks.

We aimed to detect stable regulatory cycles with the basic NFκB gene network in GBM patients - which are known to carry aberrant NFκB activation but not completely understood mechanisms - comparing them with controls to further understand physiopathology. Our hypothesis is that negative feedbacks allow finding relations among genes that may help explaining the stability of the regulatory process within the cell. Positive feedback cycles, on the other hand, may show the amount of imbalance of a certain cell in a given time.

We analyzed the target genes of the NFκB transcription factor, which are an indirect measure of the activation, to complement the cycle-based analysis. We adopted a gene-by-gene approach to discover which of the inflammatory genes of interest are differently expressed. Several alternatives could be used to investigate genes differentially expressed in lists generated from microarray data, including t-test (JEANMOUGIN et al., 2010), Significance Analysis of Microarrays (SAM) (TUSHER; TIBSHIRANI; CHU, 2001) and Linear Models for Microarray Data (*limma*) (RITCHIE et al., 2015). We used Welch two samples t-test to have an indication of the differential expression of inflammatory genes between normal and tumoral brain tissues. T-test is a widely used method to identify differentially expressed genes between two conditions, presumably because of its simplicity and interpretability (JEANMOUGIN et al., 2010). We calculated the adjusted p-values (the smallest family wise significance level at which a particular comparison will be declared statistically significant as part of the multiple comparison testing) using the function *p.adjust* supported by the R package *stats*. This function calculates a variety of different approaches for multiple comparisons given a vector of p-values. We specified the parameter for the type of correction as *method = "fdr"*. This method is a re-interpretation of Benjamini and Hochberg (BENJAMINI; HOCHBERG, 1995) procedure in terms of adjusted p-values; it controls the false discovery rate, the expected proportion of false discoveries amongst the rejected hypothesis.

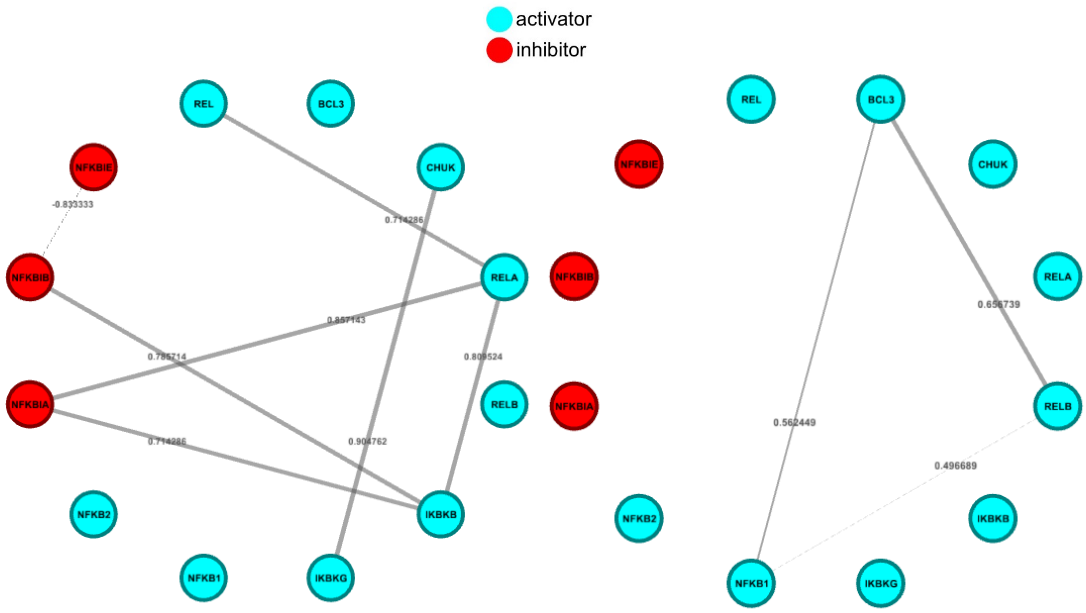
Our main results show that the stoichiometric relationship between genes involved in NFκB pathway regulation is unbalanced in GBM samples in contrast to control samples. This unregulation can be measured and explained by the identification of a positive cycle, comprising activators genes without the presence of inhibitors genes. This conclusion helps understanding more about the biology of this tumor type. As a consequence, it may enable the development of more appropriate treatment therapies that act effectively on problematic components. Furthermore, cycles help differentiate, measure and explain the phenomena identified in healthy and diseased tissue.

As presented in previous chapter, the raw data of 276 GBM samples of all histology and eight normal adult brain samples used in this study is available as experiment number GSE16011 (GRAVENDEEL et al., 2009) in the Gene Expression Omnibus (GEO)[1]. Furthermore, we analyzed GBM samples of two independent studies available on TCGA[2]: 220 samples characterized in (The Cancer Genome Atlas (TCGA) Research Network, 2008) and 520 samples characterized in (BRENNAN et al., 2013). Experimental data used in the analysis are available in AffymetrixGeneChip Human Genome U133 Plus 2.0 Array format. The analyses of Affymetrix microarray data were performed using R[3] and Bioconductor[4].

6.3 Results

The herein proposed methodology (presented in Chapter 4) analyzed genes of the NFκB network in non-tumor compared to GBM tissues from previously published and characterized database comprising 8 brain samples of non-neoplastic white matter tissue compared to 276 GBM samples of all histology (GRAVENDEEL et al., 2009). As observed previous, the difference in the number of samples from both groups stems from the difficulty of obtaining normal control data (GRIZZLE; BELL; SEXTON, 2010). The control samples were characterized as belonging to a single cluster of gene expression level similarity, while the GBM samples were divided into twenty-four different clusters (GRAVENDEEL et al., 2009). Therefore, a higher correlation coefficient is expected between the genes of the control samples than in the GBM samples. However, one can not infer from this characterization which genes are correlated (and thus constitute a network), as well as one can not infer the classification of the cycles from a network in positive and negative.

[1]http://www.ncbi.nlm.nih.gov/geo/
[2]http://cancergenome.nih.gov/
[3]http://www.r-project.org/
[4]http://www.bioconductor.org/



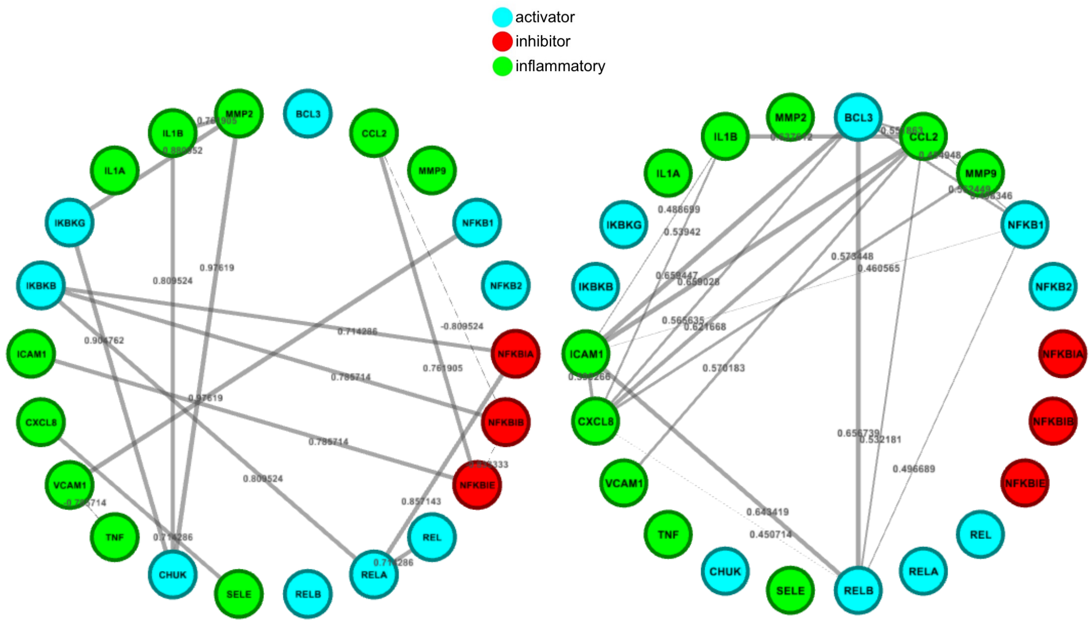
(a) Control samples (b) GBM samples

Figure 6.2 – NFκB graphs (threshold correlation 0.45). Nodes are represented as circles and correspond to genes. Connections between nodes represent a weighted undirected relationship (edge), its value represents the coefficient of correlation between genes.

Source: Author (2017).

We constructed several graphs for the network with different correlation coefficients for control samples and GBM samples. Tables 6.1 and 6.2 show, respectively, Spearman correlation coefficients and p-values for control and GBM samples. Figure 6.2 shows the graphs created for both control and GBM samples of the experiment available in GEO (GSE16011 (GRAVENDEEL et al., 2009)) for the NFκ pathway with correlation coefficient of at least 0.45. We selected 12 genes as main involved in this pathway: 9 activators (CHUK, IKBKB, IKBKG, BCL3, NFKB1, NFKB2, RELA, RELB and REL) and 3 inhibitors (NFKBIA, NFKBIB, NFKBIE). The graph from the control samples has 12 nodes and 7 edges, its average degree is 1.167, its diameter is 4 and its density is 0.106. The graph from the GBM samples has 12 nodes and 3 edges, its average degree is 0.5, its diameter is 1 and its density is 0.045. We found one cycle in the graph of control samples: three edges connect the nodes IKBKB, RELA and NFKBIA, forming a negative cycle, comprising activating genes and one inhibitor gene. In the graph of GBM samples, three edges connect the nodes BCL3, RELB and NFKB1, forming a positive cycle, comprising activating genes without the presence of inhibitors genes. This cycle was also found in the graphs built with the data of the two experiments provided in TCGA: with the samples of the experiment described in (BRENNAN et al., 2013), the cycle was found in the graph with threshold correlation 0.35; with samples of the experiment described in (The Cancer Genome Atlas (TCGA) Research Network, 2008), the cycle was also found in the graph with threshold correlation 0.35. Tables 6.3 and 6.4 show, respectively, Spearman correlation coefficients and p-values for GBM samples of TCGA experiments (The Cancer Genome Atlas (TCGA) Research Network, 2008) and (BRENNAN et al., 2013). These graphs suggest that NFκB pathway genes are in fact working under a more regulated balance between activators and inhibitors expression in normal samples, agreeing with the concept of regulated system in normal tissues.

Table 6.5 shown that most genes of the classical target genes were up-regulate in GBM samples compared to normal tissue counterparts when the transcription factor NFκB is active, agreeing with our previously published data from other datasets (ZANOTTO-FILHO et al., 2011). We observed that the T-test with significance level of 0.01 indicated that only the mean of the gene expression level of the control samples and the GBM samples of SELE and TFNα genes showed no significant difference. Furthermore, we observed that the t-test with significance level of 0.001 indicated that only IL1A, SELE and TFNα genes showed no significant difference.



(a) Control samples (b) GBM samples

Figure 6.3 – NFκB and inflammatory genes graphs (threshold correlation 0.45) of experiment GSE16011 (GRAVENDEEL et al., 2009). Nodes are represented as circles and correspond to genes. Connections between nodes represent a weighted undirected relationship (edge), its value represents the coeficient of correlation between genes.

Source: Author (2017).

Table 6.1 – Spearman correlation coefficients for the genes in control samples of GSE16011 (GRAVENDEEL et al., 2009). P-values are given in parenthesis.

Genes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1. BCL3	-																					
2. CCL2	0.11	-																				
	(0.77)																					
3. CHUK	0.45	0.38	-																			
	(0.26)	(0.35)																				
4. CXCL8	-0.09	-0.19	0.19	-																		
	(0.82)	(0.65)	(0.65)																			
5. ICAM1	0.52	0.54	0.59	0.04	-																	
	(0.18)	(0.16)	(0.11)	(0.91)																		
6. IKBKB	0.42	-0.54	0.40	-0.04	-0.14	-																
	(0.28)	(0.16)	(0.31)	(0.91)	(0.73)																	
7. IKBKG	0.52	0.11	0.90	0.28	0.57	0.47	-															
	(0.18)	(0.77)	(0.00)	(0.49)	(0.13)	(0.23)																
8. IL1A	0.07	0.35	0.50	0.35	-0.07	0.02	0.33	-														
	(0.86)	(0.38)	(0.20)	(0.38)	(0.86)	(0.95)	(0.41)															
9. IL1B	0.35	0.59	0.80	-0.19	0.40	0.21	0.57	0.66	-													
	(0.38)	(0.11)	(0.01)	(0.65)	(0.31)	(0.61)	(0.13)	(0.07)														
10. MMP2	0.38	0.30	0.97	0.28	0.61	0.38	0.88	0.47	0.76	-												
	(0.35)	(0.45)	(0.00)	(0.49)	(0.10)	(0.35)	(0.00)	(0.23)	(0.02)													
11. MMP9	0.30	-0.28	-0.30	-0.19	-0.11	0.30	-0.26	-0.59	-0.47	-0.38	-											
	(0.45)	(0.49)	(0.45)	(0.65)	(0.77)	(0.45)	(0.53)	(0.11)	(0.23)	(0.35)												
12. NFKB1	0.35	0.14	0.33	-0.02	-0.02	0.11	0.45	0.61	0.47	0.23	-0.50	-										
	(0.38)	(0.73)	(0.41)	(0.95)	(0.95)	(0.77)	(0.26)	(0.10)	(0.23)	(0.57)	(0.20)											
13. NFKB2	0.61	0.26	0.40	0.09	0.66	-0.04	-0.04	0.64	-0.04	0.14	0.33	0.47	-									
	(0.10)	(0.53)	(0.31)	(0.82)	(0.07)	(0.91)	(0.08)	(0.91)	(0.73)	(0.41)	(0.91)	(0.23)										
14. NFKBIA	0.47	-0.66	0.16	0.19	0.14	0.71	0.33	-0.26	-0.11	0.26	0.28	-0.14	0.07	-								
	(0.23)	(0.07)	(0.69)	(0.65)	(0.73)	(0.04)	(0.41)	(0.53)	(0.77)	(0.53)	(0.49)	(0.73)	(0.86)									
15. NFKBIB	0.21	-0.80	-0.09	0.02	-0.54	0.78	0.14	-0.09	-0.26	-0.14	0.38	0.23	-0.04	0.54	-							
	(0.61)	(0.01)	(0.82)	(0.95)	(0.16)	(0.02)	(0.73)	(0.82)	(0.53)	(0.73)	(0.35)	(0.57)	(0.91)	(0.16)								
16. NFKBIE	0.16	0.76	0.50	0.21	0.78	-0.40	0.26	0.21	0.45	0.54	-0.21	-0.23	0.21	-0.19	-0.83	-						
	(0.69)	(0.02)	(0.20)	(0.61)	(0.02)	(0.31)	(0.53)	(0.61)	(0.26)	(0.16)	(0.61)	(0.57)	(0.61)	(0.65)	(0.01)							
17. REL	0.47	-0.35	-0.09	0.07	-0.19	0.38	-0.09	0.26	0.07	-0.04	0.09	0.21	-0.19	0.57	0.38	-0.21	-					
	(0.23)	(0.38)	(0.82)	(0.86)	(0.65)	(0.35)	(0.82)	(0.53)	(0.86)	(0.91)	(0.82)	(0.61)	(0.65)	(0.13)	(0.35)	(0.61)						
18. RELA	0.57	-0.54	0.09	-0.19	-0.04	0.80	0.14	-0.19	0.04	0.11	0.45	-0.07	-0.11	0.85	0.61	-0.30	0.71	-				
	(0.13)	(0.16)	(0.82)	(0.65)	(0.91)	(0.01)	(0.73)	(0.65)	(0.91)	(0.77)	(0.26)	(0.86)	(0.77)	(0.00)	(0.10)	(0.45)	(0.04)					
19. RELB	-0.42	0.02	0.50	0	-0.14	0.33	0.33	0.35	0.50	0.52	-0.45	0.02	-0.38	-0.09	0.02	0.02	-0.33	-0.09	-			
	(0.28)	(0.95)	(0.20)	(1.00)	(0.73)	(0.41)	(0.41)	(0.38)	(0.20)	(0.18)	(0.26)	(0.95)	(0.35)	(0.82)	(0.95)	(0.95)	(0.41)	(0.82)				
20. SELE	0.14	-0.30	0.23	0.71	0.38	0.09	0.52	-0.19	-0.33	0.30	0.02	-0.07	0.54	0.42	0.14	0.11	-0.19	-0.04	-0.14	-		
	(0.73)	(0.45)	(0.57)	(0.04)	(0.35)	(0.82)	(0.18)	(0.65)	(0.41)	(0.45)	(0.95)	(0.86)	(0.16)	(0.28)	(0.73)	(0.77)	(0.65)	(0.91)	(0.73)			
21. TNFα	-0.19	-0.11	0.28	0.54	0.19	0.19	0.19	-0.11	-0.11	0.38	0.23	-0.69	-0.23	0.28	-0.09	0.40	-0.23	0.04	0.33	0.47	-	
	(0.65)	(0.77)	(0.49)	(0.16)	(0.65)	(0.65)	(0.65)	(0.77)	(0.77)	(0.35)	(0.57)	(0.05)	(0.57)	(0.49)	(0.82)	(0.31)	(0.57)	(0.91)	(0.41)	(0.23)		
22. VCAM1	0.26	0.04	0.16	-0.11	-0.11	0.07	0.33	0.47	0.33	0.07	-0.47	0.97	0.45	-0.16	0.28	-0.38	0.16	-0.09	-0.02	-0.09	-0.78	-
	(0.53)	(0.91)	(0.69)	(0.77)	(0.77)	(0.86)	(0.41)	(0.23)	(0.41)	(0.86)	(0.23)	(0.00)	(0.26)	(0.69)	(0.49)	(0.35)	(0.69)	(0.82)	(0.95)	(0.82)	(0.02)	

Source: Author (2017).

Table 6.2 – Spearman correlation coefficients for the genes in GBM samples of GSE16011 (GRAVENDEEL et al., 2009). P-values are given in parenthesis.

Genes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1. BCL3	-																					
2. CCL2	0.55 (0.00)	-																				
3. CHUK	-0.24 (0.00)	-0.19 (0.00)	-																			
4. CXCL8	0.56 (0.00)	0.62 (0.00)	-0.28 (0.00)	-																		
5. ICAM1	0.65 (0.00)	0.65 (0.00)	-0.18 (0.00)	0.59 (0.00)	-																	
6. IKBKB	0.35 (0.00)	0.32 (0.00)	0.03 (0.55)	0.13 (0.02)	0.32 (0.00)	-																
7. IKBKG	0.31 (0.00)	0.00 (0.88)	-0.03 (0.55)	-0.01 (0.82)	0.25 (0.00)	0.15 (0.01)	-															
8. IL1A	0.36 (0.00)	0.36 (0.00)	-0.07 (0.19)	0.37 (0.00)	0.31 (0.00)	0.23 (0.00)	0.07 (0.21)	-														
9. IL1B	0.41 (0.00)	0.63 (0.00)	-0.12 (0.03)	0.53 (0.00)	0.48 (0.00)	0.25 (0.00)	0.03 (0.00)	0.44 (0.00)	-													
10. MMP2	0.36 (0.00)	0.11 (0.05)	-0.06 (0.26)	0.12 (0.03)	0.16 (0.00)	0.23 (0.00)	0.13 (0.02)	0.03 (0.58)	0.05 (0.39)	-												
11. MMP9	0.45 (0.00)	0.37 (0.00)	-0.27 (0.00)	0.57 (0.00)	0.39 (0.00)	0.14 (0.01)	0.15 (0.00)	0.19 (0.00)	0.25 (0.00)	0.33 (0.00)	-											
12. NFKB1	0.56 (0.00)	0.49 (0.00)	-0.15 (0.00)	0.44 (0.00)	0.46 (0.00)	0.34 (0.00)	0.08 (0.15)	0.28 (0.00)	0.36 (0.00)	0.23 (0.00)	0.38 (0.00)	-										
13. NFKB2	0.29 (0.00)	0.22 (0.00)	-0.00 (0.89)	0.15 (0.00)	0.42 (0.00)	0.27 (0.00)	0.26 (0.00)	0.16 (0.00)	0.17 (0.00)	0.05 (0.39)	0.11 (0.06)	0.18 (0.00)	-									
14. NFKBIA	0.32 (0.00)	0.30 (0.00)	-0.04 (0.45)	0.29 (0.00)	0.42 (0.00)	0.07 (0.20)	0.20 (0.00)	0.20 (0.00)	0.28 (0.00)	-0.09 (0.13)	0.14 (0.01)	0.32 (0.00)	0.30 (0.00)	-								
15. NFKBIB	0.40 (0.00)	0.17 (0.00)	-0.24 (0.00)	0.19 (0.00)	0.33 (0.00)	0.22 (0.00)	0.12 (0.04)	0.22 (0.00)	0.17 (0.00)	0.25 (0.00)	0.12 (0.03)	0.22 (0.00)	0.20 (0.00)	0.17 (0.00)	-							
16. NFKBIE	0.28 (0.00)	0.31 (0.00)	-0.11 (0.06)	0.29 (0.00)	0.26 (0.00)	0.12 (0.03)	0.11 (0.04)	0.26 (0.00)	0.28 (0.00)	0.08 (0.13)	0.29 (0.00)	0.37 (0.00)	0.13 (0.02)	0.20 (0.00)	-0.02 (0.68)	-						
17. REL	0.32 (0.00)	0.27 (0.00)	0.01 (0.82)	0.20 (0.00)	0.41 (0.00)	0.21 (0.00)	0.24 (0.00)	0.30 (0.00)	0.33 (0.00)	0.01 (0.76)	0.15 (0.00)	0.25 (0.00)	0.36 (0.00)	0.42 (0.00)	0.19 (0.00)	0.24 (0.00)	-					
18. RELA	0.20 (0.00)	0.01 (0.84)	0.03 (0.58)	0.04 (0.43)	0.19 (0.00)	0.13 (0.02)	0.30 (0.90)	-0.00 (0.78)	0.01 (0.42)	0.04 (0.09)	0.09 (0.00)	0.19 (0.00)	0.26 (0.00)	0.39 (0.71)	0.02 (0.00)	0.22 (0.03)	0.12	-				
19. RELB	0.65 (0.00)	0.53 (0.00)	-0.22 (0.00)	0.45 (0.00)	0.64 (0.00)	0.34 (0.00)	0.31 (0.00)	0.33 (0.00)	0.42 (0.00)	0.17 (0.00)	0.35 (0.00)	0.49 (0.00)	0.43 (0.00)	0.31 (0.00)	0.37 (0.00)	0.33 (0.00)	0.34 (0.00)	0.25 (0.00)	-			
20. SELE	0.27 (0.00)	0.25 (0.00)	-0.02 (0.67)	0.25 (0.00)	0.36 (0.00)	0.11 (0.05)	0.09 (0.12)	0.00 (0.96)	0.20 (0.00)	0.18 (0.00)	0.21 (0.00)	0.16 (0.00)	0.18 (0.00)	0.13 (0.02)	0.03 (0.59)	0.11 (0.05)	0.13 (0.02)	0.03 (0.51)	0.28 (0.00)	-		
21. TNF α	-0.06 (0.26)	0.03 (0.52)	0.12 (0.03)	-0.06 (0.24)	0.12 (0.04)	0.08 (0.16)	0.16 (0.00)	0.03 (0.58)	0.34 (0.00)	-0.10 (0.09)	-0.09 (0.11)	-0.05 (0.35)	0.12 (0.03)	0.16 (0.00)	0.08 (0.14)	-0.07 (0.21)	0.20 (0.00)	0.03 (0.52)	-0.00 (0.96)	0.06 (0.30)	-	
22. VCAM1	0.39 (0.00)	0.57 (0.00)	-0.17 (0.00)	0.37 (0.00)	0.42 (0.00)	0.34 (0.00)	0.00 (0.93)	0.33 (0.00)	0.39 (0.00)	0.09 (0.11)	0.17 (0.00)	0.34 (0.00)	0.21 (0.00)	0.24 (0.00)	0.29 (0.00)	0.18 (0.00)	0.22 (0.00)	0.01 (0.77)	0.41 (0.00)	0.08 (0.14)	0.02 (0.67)	-

Source: Author (2017).

Table 6.3 – Spearman correlation coefficients for the genes in GBM samples of TCGA (The Cancer Genome Atlas (TCGA) Research Network, 2008). P-values are given in parenthesis.

Genes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1. BCL3	-																					
2. CCL2	0.44	-																				
	(0.00)																					
3. CHUK	-0.04	-0.12	-																			
	(0.50)	(0.05)																				
4. CXCL8	0.46	0.48	0.04	-																		
	(0.00)	(0.00)	(0.48)																			
5. ICAM1	0.60	0.50	0.02	0.46	-																	
	(0.00)	(0.00)	(0.67)	(0.00)																		
6. IKBKB	0.03	0.07	-0.06	0.02	0.14	-																
	(0.58)	(0.27)	(0.33)	(0.68)	(0.03)																	
7. IKBKG	0.14	-0.11	0	-0.20	0.12	0.14	-															
	(0.03)	(0.08)	(0.89)	(0.00)	(0.07)	(0.03)																
8. IL1A	0.17	0.18	0.11	0.33	0.23	0.19	0.18	-														
	(0.00)	(0.00)	(0.09)	(0.00)	(0.00)	(0.00)	(0.00)															
9. IL1B	0.32	0.55	0.08	0.54	0.39	0.11	-0.02	0.48	-													
	(0.00)	(0.00)	(0.20)	(0.00)	(0.00)	(0.10)	(0.75)	(0.00)														
10. MMP2	0.19	0.02	-0.05	0.11	0.23	0.10	0.03	-0.05	0.02	-												
	(0.00)	(0.71)	(0.45)	(0.08)	(0.00)	(0.11)	(0.60)	(0.40)	(0.76)													
11. MMP9	0.27	0.25	-0.05	0.38	0.28	0.06	0	0.13	0.30	0.27	-											
	(0.00)	(0.00)	(0.44)	(0.00)	(0.00)	(0.33)	(0.92)	(0.05)	(0.00)	(0.00)												
12. NFKB1	0.39	0.26	0.04	0.23	0.49	0.13	0.08	0.10	0.30	0.08	0.24	-										
	(0.00)	(0.00)	(0.46)	(0.00)	(0.00)	(0.04)	(0.22)	(0.11)	(0.00)	(0.23)	(0.00)											
13. NFKB2	0.20	0	0.05	0.11	0.17	0.05	0.29	0.32	0.13	0.03	0.05	0.30	-									
	(0.00)	(0.97)	(0.38)	(0.09)	(0.00)	(0.37)	(0.00)	(0.00)	(0.03)	(0.62)	(0.45)	(0.00)										
14. NFKBIA	0.39	0.32	-0.02	0.28	0.34	0.11	0.03	0.12	0.21	-0.06	0.09	0.31	0.15	-								
	(0.00)	(0.00)	(0.72)	(0.00)	(0.00)	(0.09)	(0.57)	(0.06)	(0.00)	(0.32)	(0.14)	(0.00)	(0.01)									
15. NFKBIB	0.01	-0.12	-0.02	-0.04	0.08	0.01	0.14	-0.03	-0.10	0.20	-0.01	-0.03	0.11	-0.16	-							
	(0.87)	(0.06)	(0.71)	(0.55)	(0.19)	(0.80)	(0.03)	(0.58)	(0.12)	(0.00)	(0.79)	(0.63)	(0.09)	(0.01)								
16. NFKBIE	0.22	0.21	0	0.28	0.37	-0.08	0.13	0.24	0.36	0.06	0.19	0.23	0.19	0.17	0.02	-						
	(0.00)	(0.00)	(0.89)	(0.00)	(0.00)	(0.20)	(0.04)	(0.00)	(0.00)	(0.35)	(0.00)	(0.00)	(0.00)	(0.00)	(0.75)							
17. REL	0.25	0.11	0.07	0.07	0.38	0.26	0.28	0.40	0.22	0.03	0.11	0.37	0.47	0.27	0	0.15	-					
	(0.00)	(0.10)	(0.24)	(0.25)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.56)	(0.08)	(0.00)	(0.00)	(0.00)	(0.93)	(0.01)						
18. RELA	0.01	0.07	-0.22	0	0.03	0.11	0.10	-0.10	-0.02	0.01	0.03	-0.01	-0.08	0.24	-0.02	0.07	-0.16	-				
	(0.80)	(0.25)	(0.00)	(0.99)	(0.59)	(0.09)	(0.10)	(0.11)	(0.69)	(0.86)	(0.58)	(0.86)	(0.22)	(0.00)	(0.73)	(0.25)	(0.01)					
19. RELB	0.42	0.30	-0.01	0.13	0.45	0.07	0.13	0.14	0.22	0.01	0.07	0.35	0.10	0.29	0.06	0.27	0.30	0	-			
	(0.00)	(0.00)	(0.80)	(0.04)	(0.00)	(0.25)	(0.04)	(0.03)	(0.00)	(0.85)	(0.26)	(0.00)	(0.12)	(0.00)	(0.30)	(0.00)	(0.00)	(0.93)				
20. SELE	0.08	0.02	0.11	0.07	0.09	-0.08	0.07	0.08	0.11	0	0.04	0.20	0.13	0.07	-0.01	0.03	0.22	-0.11	0.14	-		
	(0.23)	(0.74)	(0.07)	(0.28)	(0.13)	(0.20)	(0.25)	(0.18)	(0.08)	(0.90)	(0.50)	(0.00)	(0.04)	(0.26)	(0.81)	(0.57)	(0.00)	(0.09)	(0.02)			
21. TNFα	-0.05	0.02	0.10	0.03	0.04	0.10	0.04	0.25	0.26	-0.12	-0.01	0.03	0.19	0.07	-0.02	0.12	0.29	-0.16	0.11	0.20	-	
	(0.41)	(0.67)	(0.10)	(0.64)	(0.51)	(0.12)	(0.48)	(0.00)	(0.00)	(0.06)	(0.87)	(0.56)	(0.00)	(0.29)	(0.68)	(0.07)	(0.00)	(0.01)	(0.09)	(0.00)		
22. VCAM1	0.21	0.36	-0.03	0.22	0.29	0.25	-0.03	0.17	0.29	0.05	0.06	0	-0.01	0.35	-0.05	0.14	0.08	0.20	0.20	-0.12	0.02	-
	(0.00)	(0.00)	(0.60)	(0.00)	(0.00)	(0.00)	(0.56)	(0.01)	(0.00)	(0.42)	(0.35)	(0.93)	(0.78)	(0.00)	(0.43)	(0.03)	(0.23)	(0.00)	(0.00)	(0.05)	(0.70)	

Source: Author (2017).

Table 6.4 – Spearman correlation coefficients for the genes in GBM samples of TCGA (BRENNAN et al., 2013). P-values are given in parenthesis.

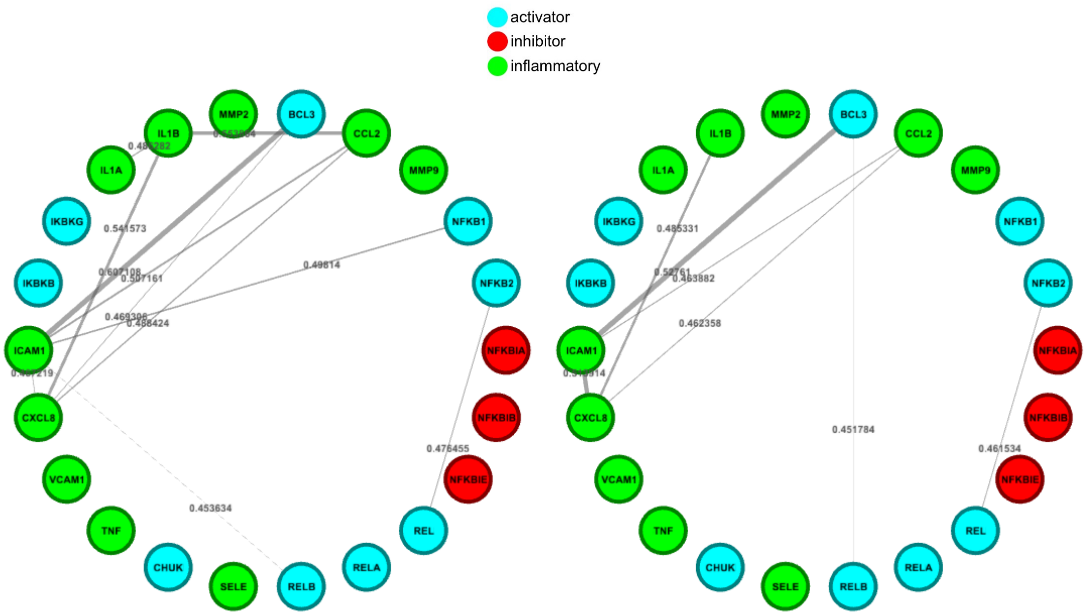
Genes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1. BCL3	-																					
2. CCL2	0.52 (0.00)	-																				
3. CHUK	-0.12 (0.00)	-0.10 (0.01)	-																			
4. CXCL8	0.49 (0.00)	0.51 (0.00)	-0.03 (0.43)	-																		
5. ICAM1	0.61 (0.00)	0.60 (0.00)	-0.04 (0.29)	0.54 (0.00)	-																	
6. IKBKB	0.05 (0.21)	0.11 (0.00)	-0.14 (0.00)	-0.07 (0.09)	0.10 (0.01)	-																
7. IKBKG	0.22 (0.00)	0.03 (0.43)	-0.10 (0.02)	-0.06 (0.11)	0.17 (0.00)	0.13 (0.00)	-															
8. IL1A	0.26 (0.00)	0.23 (0.00)	-0.04 (0.26)	0.31 (0.00)	0.30 (0.00)	0.16 (0.00)	0.19 (0.00)	-														
9. IL1B	0.41 (0.00)	0.54 (0.00)	-0.03 (0.42)	0.55 (0.00)	0.48 (0.00)	0.09 (0.03)	0.08 (0.06)	0.46 (0.00)	-													
10. MMP2	0.10 (0.02)	0.01 (0.81)	-0.03 (0.43)	0.06 (0.16)	0.10 (0.01)	0.05 (0.18)	0.00 (0.89)	-0.04 (0.33)	0.05 (0.22)	-												
11. MMP9	0.31 (0.00)	0.25 (0.00)	-0.10 (0.01)	0.37 (0.00)	0.31 (0.00)	-0.02 (0.64)	0.04 (0.29)	0.14 (0.00)	0.30 (0.00)	0.25 (0.00)	-											
12. NFKB1	0.40 (0.00)	0.37 (0.00)	0.02 (0.58)	0.24 (0.00)	0.47 (0.00)	0.23 (0.00)	0.17 (0.00)	0.17 (0.00)	0.38 (0.00)	0.09 (0.02)	0.19 (0.00)	-										
13. NFKB2	0.22 (0.00)	0.13 (0.00)	0.04 (0.29)	0.09 (0.02)	0.24 (0.00)	0.11 (0.00)	0.26 (0.00)	0.26 (0.00)	0.16 (0.00)	-0.02 (0.52)	0.03 (0.36)	0.33 (0.00)	-									
14. NFKBIA	0.40 (0.00)	0.34 (0.00)	-0.07 (0.08)	0.30 (0.00)	0.40 (0.00)	0.19 (0.00)	0.09 (0.02)	0.18 (0.00)	0.23 (0.00)	-0.02 (0.49)	0.09 (0.03)	0.35 (0.00)	0.20 (0.00)	-								
15. NFKBIB	0.01 (0.69)	-0.09 (0.02)	0 (0.95)	-0.10 (0.01)	-0.02 (0.58)	0 (0.92)	0.08 (0.04)	-0.07 (0.08)	-0.13 (0.00)	0.13 (0.00)	0 (0.86)	0.02 (0.56)	0.05 (0.22)	-0.09 (0.02)	-							
16. NFKBIE	0.23 (0.00)	0.20 (0.00)	0.030 (0.48)	0.31 (0.00)	0.25 (0.00)	-0.19 (0.00)	0.07 (0.08)	0.11 (0.00)	0.27 (0.00)	0 (0.96)	0.21 (0.00)	0.09 (0.02)	0.04 (0.31)	0.07 (0.09)	-0.05 (0.25)	-						
17. REL	0.29 (0.00)	0.22 (0.00)	0.03 (0.46)	0.08 (0.05)	0.34 (0.00)	0.25 (0.00)	0.25 (0.00)	0.35 (0.00)	0.28 (0.00)	0.09 (0.02)	0.11 (0.00)	0.34 (0.00)	0.39 (0.00)	0.23 (0.00)	-0.09 (0.03)	0.01 (0.69)	-					
18. RELA	0.11 (0.01)	0.07 (0.08)	-0.02 (0.59)	-0.02 (0.50)	0.12 (0.00)	0.15 (0.00)	0.11 (0.00)	-0.07 (0.10)	0.03 (0.36)	0.06 (0.13)	0 (0.97)	0.25 (0.00)	0.11 (0.00)	0.16 (0.00)	0.12 (0.00)	0.04 (0.31)	0 (0.96)	-				
19. RELB	0.44 (0.00)	0.37 (0.00)	-0.02 (0.54)	0.25 (0.00)	0.42 (0.00)	0.11 (0.00)	0.15 (0.00)	0.19 (0.00)	0.26 (0.00)	0 (0.97)	0.11 (0.00)	0.35 (0.00)	0.17 (0.00)	0.32 (0.00)	0.01 (0.76)	0.17 (0.00)	0.23 (0.00)	0.14 (0.00)	-			
20. SELE	0.15 (0.00)	0.08 (0.05)	0.13 (0.00)	0.14 (0.00)	0.12 (0.00)	-0.12 (0.00)	0.08 (0.04)	0.03 (0.38)	0.08 (0.04)	0.01 (0.71)	0.12 (0.00)	0.08 (0.04)	0.09 (0.03)	0.03 (0.46)	-0.04 (0.33)	0.09 (0.03)	0.17 (0.00)	-0.08 (0.05)	0.11 (0.00)	-		
21. TNFα	-0.01 (0.81)	0.11 (0.00)	0.04 (0.27)	0.04 (0.32)	0.11 (0.00)	0.14 (0.00)	0.06 (0.12)	0.15 (0.00)	0.29 (0.00)	-0.03 (0.48)	0.05 (0.17)	0.12 (0.00)	0.11 (0.01)	0.05 (0.17)	-0.02 (0.59)	0.02 (0.62)	0.23 (0.00)	-0.04 (0.26)	0.11 (0.00)	0.07 (0.08)	-	
22. VCAM1	0.17 (0.00)	0.39 (0.00)	-0.02 (0.50)	0.19 (0.00)	0.31 (0.00)	0.21 (0.00)	0 (0.82)	0.17 (0.00)	0.27 (0.00)	-0.07 (0.08)	0.01 (0.74)	0.13 (0.00)	0.09 (0.03)	0.38 (0.00)	-0.11 (0.00)	0.08 (0.04)	0.16 (0.00)	0.14 (0.00)	0.22 (0.00)	-0.02 (0.61)	0.03 (0.42)	-

Source: Author (2017).

Table 6.5 – Mean, standard deviation (SD), p-value and adjusted p-value of the gene expression values of the inflammatory genes.

Gene	Control Mean	Control SD	GBM Mean	GBM SD	p-value	adjusted p-value
CCL2	216.76	291.63	2740.85	4606.41	3.2871e-15	8.2179e-15
ICAM1	188.23	16.19	288.71	277.05	1.0968e-18	3.6561e-18
IL1A	65.32	22.13	101.71	50.26	1.7684e-03	2.2105e-03
IL1B	204.42	169.47	650.93	829.54	1.3687e-05	1.9553e-05
CXCL8	99.44	83.72	1654.08	3342.20	3.2944e-13	6.5888e-13
MMP2	93.13	15.85	616.98	504.23	5.7933e-45	5.7933e-44
MMP9	169.52	23.66	935.65	1898.20	1.2483e-10	2.0804e-10
SELE	27.21	21.78	37.01	127.74	3.7567e-01	3.7567e-01
VCAM1	156.19	94.22	1698.01	1788.33	5.6545e-32	2.8273e-31
TNFα	74.77	50.39	95.05	103.74	3.1123e-01	3.4582e-01

Source: Author (2017).



(a) GBM samples of TCGA (The Cancer Genome Atlas (TCGA) Research Network, 2008)

(b) GBM samples of TCGA (BRENNAN et al., 2013)

Figure 6.4 – NFκB and inflammatory genes graphs (threshold correlation 0.45) of TCGA experiments (The Cancer Genome Atlas (TCGA) Research Network, 2008; BRENNAN et al., 2013). Nodes are represented as circles and correspond to genes. Connections between nodes represent a weighted undirected relationship (edge), its value represents the coeficient of correlation between genes.

Source: Author (2017).

We repeated the analysis of the NFκB pathway including among the genes of interest CCL2, ICAM1, IL1A, IL1B, CXCL8, MMP2, MMP9, SELE, TNFα and VCAM1. Figure 6.3 shows the graphs created for both control and GBM samples of the experiment available in GEO (GSE16011 (GRAVENDEEL et al., 2009)) for the NFκB pathway and aforementioned inflammatory genes with correlation coefficient of at least 0.45. Spearman correlation coefficients and p-values for control and GBM samples are shown in Tables 6.1 and 6.2, respectively.

We selected 22 genes as main involved in this set: 9 activators, 3 inhibitors and 10 inflammatory gene. The graph from the control samples has 22 nodes and 17 edges, its average degree is 1.545, its diameter is 5 and its density is 0.074. The graph from the GBM samples has 22 nodes and 20 edges, its average degree is 1.818, its diameter is 3 and its density is 0.087. One can note that inhibitors genes (colored in red) are connected in the control samples graph, but none connection was observed in the GBM samples graph, evidencing the same dysregulation pattern described in Figure 6.2. This same behavior was also observed in the graphs with threshold correlation 0.45 in the TCGA experiments (The Cancer Genome Atlas (TCGA) Research Network, 2008; BRENNAN et al., 2013) shown in Figure 6.4. Spearman correlation coefficients and p-values for GBM samples of TCGA experiments (The Cancer Genome Atlas (TCGA) Research Network, 2008) and (BRENNAN et al., 2013) are shown in Tables 6.3 and 6.4, respectively.

We found 5 cycles in the graph of control samples, among which 3 positive cycles (all cycles with three nodes) and 2 negative cycles (one cycle with three nodes and one cycle with four nodes). In the graph of GBM samples, we found 385 positive cycles: (i) 20 cycles with three nodes; (ii) 47 cycles with four nodes; (iii) 93 cycles with five nodes; (iv) 117 cycles with six nodes; (v) 87 cycles with seven nodes; and (vi) 21 cycles with eight nodes.

Six cycles of GBM samples graphs appears in the analysis of the three (GSE16011 (GRAVENDEEL et al., 2009) and TCGA (The Cancer Genome Atlas (TCGA) Research Network, 2008; BRENNAN et al., 2013)) experiments with threshold correlation 0.45. Table 6.6 shown these cycles that could give tips to biochemists about possible laboratory experiments, as well as on potential drug target genes.

Table 6.6 – Cycles with potential target genes.

Genes (vertices) of the cycle	Size of the cycle
BCL3 - CXCL8 - ICAM1	3
CCL2 - CXCL8 - ICAM1	3
CCL2 - CXCL8 - IL1B	3
BCL3 - CCL2 - CXCL8 - ICAM1	4
CCL2 - ICAM1 - CXCL8 - IL1B	4
BCL3 - CXCL8 - IL1B - CCL2 - ICAM1	5

Source: Author (2017).

6.4 Discussion

In this study, we aimed to detect stable cycles with the basic genes of NFκB network and NFκB with inflammatory genes network in GBM patients, that are known to have an aberrant activation in these pathways but not completely understood mechanisms, comparing them with controls to further understand physiopathology. Our main strategy is based on the idea that a network must present stable feedback connections (negative cycles). Hence, we look for these stable connections among genes after pre-processing. Our main results show that the stoichiometric relationship between genes involved in these networks is unbalanced in GBM samples *versus* control samples. This dysregulation can be measured and explained by the identification of a higher percentage of positive feedback (positive cycles) in these networks, comprising activators genes without the presence of inhibitors genes. This conclusion helps to understand more about the biology of this tumor type.

The results still evidence the differences between the GCNs of the networks among the control samples and GBM samples. The control graphs present negative (and intrinsically stable) cycles. In these cycles, we observe the existence of an inhibitor gene (NFKBIA, NFKBIB and NFKBIE). These cycles indicate stoichiometric balance between activators and inhibitors, because there are inhibitors genes regulating activators genes. Based on the current knowledge of NFκB system functioning from in-bench assays, this equilibrium implies in the blockade of the NFκB transcription factor, an expected phenomena in normal brain tissues (RAYCHAUDHURI et al., 2007; ZANOTTO-FILHO et al., 2011).

In GBM samples, the stoichiometric relationship between activators and inhibitors drastically changes. The cycles are positive, consisting of activators. There is no inhibitor gene in these cycles; consequently the balance activator/inhibitor seems to be broken. Thus, from these results, one could conclude there is activation of the NFκB transcription factor, agreeing with findings from in-bench studies in GBM. Particularly in the NFκB regulatory system, an inhibitor gene such as IκB proteins (NFKBIA, NFKBIB, NFKBIE genes), for example, do not decrease the amount of the activator gene; it reduces the operation of these genes products, by direct binding and cytoplasmic retention of activators thus impeding its nuclear activity (BARKETT; GILMORE, 1999; AGGARWAL, 2004). Hence, with the absence of an inhibitor gene in the cycle, NFκB activators genes can migrate into nucleus and regulate the activation of its target genes ultimately affecting cell function.

Therefore, one could conclude that the inhibitory genes can participate in the cycles of control samples, but are not involved in the cycles of GBM samples. In this way, it suggests that the NFκB genes and the inflammatory genes in the control sample remain with balanced operation, since there is the negative feedback provided by NFκB inhibitors. Moreover, the GBM samples are still showing an unbalanced behavior, without the inhibitors genes there is no negative feedback and activators and inflammatory genes are unregulated.

Our challenge in the analysis of the NFκB pathway is to understand the NFκB network. We explore the structure through the cycles of the network to be able to understand more of the relationships established by genes of the pathway and identify and quantify differences between GBM and control networks. Our analysis is focused on the genes from the NFκB pathway - which is in fact a really intricate and multi-regulated transcription factor - so the goal was not to identify modules, pathways or motifs, but rather to better understand the relationships of the genes of the NFκB pathway and its variations between samples of GBM and control to get insights on how alterations in the levels of upstream controllers (inhibitors) may affect the activation of the NFκB based on target genes evaluation. Our findings show that stoichiometric relation of genes within the NFκB network seems to be dysregulated in tumors *versus* normal tissues, and it seems to be caused by absence of inhibitors in the system, which lead to increased expression of NFκB targets.

As mentioned earlier, analysis results could provide tips to biochemists about possible laboratory experiments. The results of this analysis drew the attention of the group of biochemists who worked with us for the BCL3 gene. This gene does not have relations in the control samples graphs. However, on the other hand, presents a set of connections in the GBM samples graphs. In addition, BCL3 gene is present in the only positive cycle found in the GBM samples graph (threshold correlation 0.45) built for the NFκB pathway genes, as well as is present in the cycles found in the GBM samples graph (threshold correlation 0.45) construct for the NFκB pathway and inflammatory genes. So the biochemists decided to perform laboratory experiments for testing this specific gene.

7 CONCLUSIONS

The amount of biological data available grows a lot. However, these data alone do not help the experts in biology and biochemistry, computational models, techniques, tools and methodologies are needed to analyze and extract relevant information from such data sets. Systems biology (IDEKER; GALITSKI; HOOD, 2001) studies an organism modeled as a *network* of biological components (genes, proteins, and chemical reactions). Its focus are all components and the interaction between them, all as part of a single system, as indeed happens in reality. In computer science, a *network* is often called *graph*, an abstract data type that is meant to implement the graph concept from mathematics (DIESTEL, 2012). Thus, the network language is particularly intuitive for both areas and may be used as a basis to model and analyze biological data, extracting relevant information. In this context, a widely used approach is the construction and analysis of gene networks (SERIN et al., 2016) based on gene expression similarity. In this thesis, we present a new way of analyzing co-expression networks constructed using quantized data obtained from the microarrays, based on the different kinds of cycles found among genes of these networks.

The input of the proposed method is the raw data of a set of samples (a microarray), a set of interest genes (for example, genes from a known pathway) and a function (activator or inhibitor) of these genes. The output is a set of cycles. Thanks to the new way of finding relations among genes, a more robust interpretation of gene correlations is possible, because cycles are associated with feedback mechanisms, that are very common in biological networks.

Our hypothesis is that negative feedbacks allow finding relations among genes that may help explaining the stability of the regulatory process within the cell. Positive feedback cycles, on the other hand, may show the amount of imbalance of a certain cell at a given time. The cycle-based analysis allows identifying the stoichiometric relationships among the genes of the network. This methodology provides a better understanding of the tumor biology. As a consequence, it may enable the development of more effective treatment therapies. Furthermore, cycles help differentiate, measure and explain the phenomena identified in healthy and diseased tissues. Cycles may also be used as a new method for classification of samples of a microarray (cancer diagnosis). Compared to other classification methods, cycle-based classification provides a richer explanation of the proposed classification, that can give hints on the possible therapies.

Therefore, the main original contributions of this thesis are:

(i) A new cycle-based analysis method. The new way of analyzing relations among genes allows a more robust interpretation of gene correlations. The cycle-based analysis enables the identification of the stoichiometric relationships between the genes of a network. Furthermore, it provides a better understanding of the tumor biology. Thus, it may aid the development of more appropriate treatment therapies that act effectively on problematic components. Cycles help differentiate, measure and explain the phenomena identified in healthy and diseased tissues.

(ii) A new microarray samples classification method. The use of cycles as the basis of a new method for classification of samples of a microarray allows, for example, cancer diagnosis. The main advantage of its use for these purposes is the embedded biological significance. In the most classic algorithms, classification occurs in a black box through complex mathematical formulas. With the proposed method, it is easy to understand the classification: if the sample has cycles of a particular class, then the sample belongs to that class.

(iii) Application and achievement of practical results. We use the proposed methodology to analyze the genes of four networks closely related to cancer – apoptosis, glucolysis, cell cycle and NFκB – in tissues of the most aggressive type of brain tumor (Gliobastoma multiforme – GBM) and in healthy tissues.

Our main results show that the stoichiometric relationship between genes involved in apoptosis, glucolysis, cell cycle and NFκB pathways is unbalanced in GBM samples *versus* control samples. This dysregulation can be measured and explained by the identification of a higher percentage of positive cycles in these networks. This conclusion helps to understand more about the biology of this tumor type. These results also confirmed our hypothesis.

The results concerning the analysis of the first three networks validate our method, since they can be confirmed in results of previous studies. The results of the analysis of the NFκB pathway, in turn, allowed us to obtain a new significant result: the highlight of the gene BCL3, because its behavior is different in control and GBM networks.

The proposed cycle-based classification method achieved the same performance metrics as a neural network, a classical classification method. However, our method has a significant advantage with respect to neural networks. The proposed classification method not only classifies samples, providing diagnosis, but also explains why samples were classified in a certain way in terms of the feedback mechanisms that are present/absent. This way, the method provides hints to biochemists about possible laboratory experiments, as well as on potential drug target genes.

GBM is the most aggressive type of brain tumor, with generally extremely poor prognosis. Thus, these tumors have drawn significant attention. Hence the discovery of novel molecular therapeutic targets can be important to optimize treatment strategies (MRUGALA, 2013). Therefore, it becomes evident the relevance of the results obtained in this thesis, as they help to understand the tumor biology, they may assist biochemical formulate new and more efficient protocols, propose new approaches for therapies, and even review the current concepts of traditional pharmacological treatments for the disease.

The research presented in this thesis can be extended in several ways, some of which are listed below:

(i) Multiple Data Integration. Our method of analysis could also integrate multiple types of data (for example, miRNA data and proteomic data), thus extending the analysis to post-transcription and translation levels. These possibilities are promising in particular because transcript levels does not correlate with protein levels due to variety of reasons like differences in translation efficiency, protein stability and miRNA regulation (SOMASUNDARAM; NIJAGUNA; KUMAR, 2011-08-23).

(ii) Single Cell Through single cell correlations complex interactions between genes might be identified, so better suggestions could be provided to biochemists about possible laboratory experiments, as well as better potential drug target genes. To use single cell data in our method, some changes considering the characteristic of the new format, including an abundance of zeros, increased heterogeneity, and complex expression distributions are required (BACHER; KENDZIORSKI, 2016). Although the construction of the networks and the identification of the cycles will basically remain the same with single cell data, the pre-processing of the new format will be different. Especially, it would be interesting to compare the data obtained in our control samples with other experiments data (for example, using the data published in (DARMANIS et al., 2015)).

(iii) Temporal Analysis. - The analysis using cycles could also be introduced into a dynamic model (temporal). In the dynamic model (with temporal information), we would identify the direction of the edges of the graphs that would then represent gene regulatory networks, thus providing more information about the relationships between genes. In addition, the analysis using cycles allows one to observe the stoichiometric relationships over time, and identify and define rules regarding the changes over time associated with cycles. This information can complement the static model information regarding the interactions between genes and, moreover, could allow the prediction of future stoichiometric changes associated with cycles. Thus, the predictions will have biological explanation.

The development of such a new way of analyzing networks can help biochemists to answer, for example, the following questions:

- what are the genes that remain in the cycles?
- what are the genes that appear and disappear in the cycles?
- what are the cycles that remain in the network?
- what are the cycles that disappear in the network?
- what are the new cycles in the network?
- how is the process of creating and destroying cycles?

The last question is the most important, since it is related to the evolution of the cell treatment (i.e., the cell change process). This process will give us information about the behavior of the cell in response to treatment. If we can identify patterns of behavior over time, we can better understand the behavior of tumor progression/regression and we will have more chances to interfere in it. Our methodology identifies the network and the cycles, so it does not only describe the genes involved in a given moment, but it also identifies the processes that are occurring (cycles) at the moment. In a temporal analysis, we would analyze how these cycles are modified over time. Thus, we might explain how the processes (cycles) evolve over time, which is quite different from other temporal analysis approaches.

The extension of the methodology would allow a more complete analysis of microarray data. Thus, the biological processes will be better understood, and biochemical operations will be performed safely in the cell, allowing, for example, as already mentioned, treatment of diseases.

(iv) Cycles Visualization. - The visualization of the networks cycles in a graphical perspective is very important further work, since ir could allow researchers to intuitively explore the cycles and develop hypotheses. One possibility of visualization could be a graph that could show the interactions between cycles and bring interesting insights. In this graph, the nodes could represent the cycles and the edges could indicate the common genes, the weight of a edge could indicate how many genes are shared by the connected nodes (cycles). Another possibility of visualization would be a correlogram, which could allow a visual perspective of how correlated the cycles are (a possible measure of correlation could be the amount of shared genes).

REFERENCES

AGGARWAL, B. B. Nuclear factor-kappab: the enemy within. **Cancer Cell**, v. 6, n. 3, p. 203–208, 2004.

ALON, U. Network motifs: theory and experimental approaches. **Nature Reviews Genetics**, Nature Publishing Group, v. 8, n. 6, p. 450–461, 2007.

ALTAY, G.; EMMERT-STREIB, F. Inferring the conservative causal core of gene regulatory networks. **BMC Systems Biology**, v. 4, p. 132, 2010.

AOKI, K.; OGATA, Y.; SHIBATA, D. Approaches for extracting practical information from gene co-expression networks in plant biology. **Plant and Cell Physiology**, v. 48, n. 3, p. 381, 2007. Available from Internet: <http://dx.doi.org/10.1093/pcp/pcm013>. Accessed: October 19, 2017.

ASYALI, M. H. et al. Gene expression profile classification: A review. **Current Bioinformatics**, v. 1, p. 55–73, 2006.

BACHER, R.; KENDZIORSKI, C. Design and computational analysis of single-cell rna-sequencing experiments. **Genome Biology**, v. 17, n. 1, p. 63, 2016. ISSN 1474-760X. Available from Internet: <http://dx.doi.org/10.1186/s13059-016-0927-y>. Accessed: October 19, 2017.

BARILLOT, E. et al. **Computational Systems Biology of Cancer**. [S.l.]: CRC Press, 2012. (Chapman & Hall/CRC Mathematical and Computational Biology). Available from Internet: <https://books.google.com.br/books?id=njfNBQAAQBAJ>. Accessed: October 19, 2017. ISBN 9781439831458.

BARKETT, M.; GILMORE, T. D. Control of apoptosis by rel/nfkb transcription factors. **Oncogene**, v. 18, p. 6910–6924, 1999.

BASAK, S.; SHIH, V. F.; HOFFMANN, A. Generation and activation of multiple dimeric transcription factors within the nf-kappab signaling system. **Mol. Cell. Biol.**, v. 28, n. 10, p. 3139–3150, 2008.

BAUD, V.; KARIN, M. Is nfκb a good target for cancer therapy? hopes and pitfalls. **Nat. Rev. Drug. Discov.**, v. 8, p. 33–44, 2009.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 57, n. 1, p. 289–300, 1995. ISSN 00359246. Available from Internet: <http://www.jstor.org/stable/2346101>. Accessed: October 19, 2017.

BERG, J.; TYMOCZKO, J.; STRYER, L. The glycolytic pathway is tightly controlled. In: **Biochemistry**. 5. ed. New York: W H Freeman, 2002. chp. 16. Available from Internet: <https://www.ncbi.nlm.nih.gov/books/NBK22395/>. Accessed: October 19, 2017.

BOCCALETTI, S.; LATORA, V.; MORENO, Y. **Handbook on Biological Networks**. [S.l.]: World Scientific, 2010. (World Scientific lecture notes in complex systems). Available from Internet: <https://books.google.com.br/books?id=jeiSmAEACAAJ>. Accessed: October 19, 2017. ISBN 9789812838797.

BORNHOLDT, S.; SCHUSTER, H. G. (Ed.). **Handbook of Graphs and Networks: From the Genome to the Internet**. New York, NY, USA: John Wiley & Sons, Inc., 2003. ISBN 3527403361.

BRENNAN, C. W. et al. The somatic genomic landscape of glioblastoma. **Cell**, v. 155, n. 2, p. 462–477, 10 2013. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3910500/>. Accessed: October 19, 2017.

BUTTE, A. J. et al. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. **Proceedings of the National Academy of Sciences of the United States of America**, United States, v. 97, n. 22, p. 12182–12186, oct. 2000. ISSN 0027-8424. Available from Internet: <http://dx.doi.org/10.1073/pnas.220392197>. Accessed: October 19, 2017.

CARTER, S. L. et al. Gene co-expression network topology provides a framework for molecular characterization of cellular state. **Bioinformatics**, v. 20, n. 14, p. 2242–2250, 2004. Available from Internet: <http://dx.doi.org/10.1093/bioinformatics/bth234>. Accessed: October 19, 2017.

CASTRO, M. A. A. et al. Evolutionary origins of human apoptosis and genome-stability gene networks. **Nucleic Acids Research**, Oxford University Press, v. 36, n. 19, p. 6269–6283, 11 2008. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2577361/>. Accessed: October 19, 2017.

CHOOBDAR, S.; RIBEIRO, P.; SILVA, F. Discovering weighted motifs in gene co-expression networks. In: **Proceedings of the 30th Annual ACM Symposium on Applied Computing**. New York, NY, USA: ACM, 2015. (SAC ’15), p. 10–17. ISBN 978-1-4503-3196-8. Available from Internet: <http://doi.acm.org/10.1145/2695664.2695773>. Accessed: October 19, 2017.

CORMEN, T. H. et al. **Introduction to Algorithms**. 2nd. ed. [S.l.]: McGraw-Hill Higher Education, 2001. ISBN 0070131511.

CRICK, F. Central dogma of molecular biology. **Nature**, v. 227, n. 5258, p. 561–563, 08 1970. Available from Internet: <http://dx.doi.org/10.1038/227561a0>. Accessed: October 19, 2017.

DAM, S. van et al. Gene co-expression analysis for functional classification and gene–disease predictions. **Briefings in Bioinformatics**, p. bbw139, 2017. Available from Internet: <http://dx.doi.org/10.1093/bib/bbw139>. Accessed: October 19, 2017.

DARMANIS, S. et al. A survey of human brain transcriptome diversity at the single cell level. **Proceedings of the National Academy of Sciences of the United States of America**, National Academy of Sciences, v. 112, n. 23, p. 7285–7290, 06 2015. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466750/>. Accessed: October 19, 2017.

DAS, S. et al. **Handbook of Research on Computational Methodologies in Gene Regulatory Networks**. 1st. ed. Hershey, PA: Information Science Reference - Imprint of: IGI Publishing, 2009. ISBN 1605666858, 9781605666853.

DIESTEL, R. **Graph Theory**. 4. ed. [S.l.]: Springer, 2012. (Graduate texts in mathematics, v. 173). ISBN 978-3-642-14278-9.

ELMORE, S. Apoptosis: A review of programmed cell death. **Toxicologic pathology**, v. 35, n. 4, p. 495–516, 2007. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2117903/>. Accessed: October 19, 2017.

FAITH, J. J. et al. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. **PLoS Biol**, v. 5, n. 1, p. 8, 2007.

FUENTE, A. de la. From ‘differential expression’ to ‘differential networking’ - identification of dysfunctional regulatory networks in diseases. **Trends in Genetics**, Elsevier, v. 26, n. 7, p. 326–333, July 2010.

FULLER, T. F. et al. Weighted gene coexpression network analysis strategies applied to mouse weight. **Mammalian Genome**, v. 18, n. 6, p. 463–472, 2007. ISSN 1432-1777. Available from Internet: <http://dx.doi.org/10.1007/s00335-007-9043-3>. Accessed: October 19, 2017.

GAITERI, C. et al. Beyond modules & hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. **Genes, brain, and behavior**, v. 13, n. 1, p. 13–24, 01 2014. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3896950/>. Accessed: October 19, 2017.

GILLIES, R. J.; ROBEY, I.; GATENBY, R. A. Causes and consequences of increased glucose metabolism of cancers. **Journal of Nuclear Medicine**, v. 49, n. Suppl 2, p. 24S–42S, 2008. Available from Internet: <http://jnm.snmjournals.org/content/49/Suppl_2/24S.abstract>. Accessed: October 19, 2017.

GRAVENDEEL, L. A. et al. Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. **Cancer Research**, American Association for Cancer Research, v. 69, n. 23, p. 9065–9072, 2009. ISSN 0008-5472. Available from Internet: <http://cancerres.aacrjournals.org/content/69/23/9065>. Accessed: October 19, 2017.

GRIZZLE, W. E.; BELL, W. C.; SEXTON, K. C. Issues in collecting, processing and storing human tissues and associated information to support biomedical research. **Cancer biomarkers: section A of Disease markers**, v. 9, n. 1-6, p. 531–549, 2010. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3445033/>. Accessed: October 19, 2017.

HACHE, H.; LEHRACH, H.; HERWIG, R. Reverse engineering of gene regulatory networks: A comparative study. **Journal on Bioinformatics and Systems Biology**, Hindawi Publishing Corp., New York, NY, United States, v. 2009, p. 8:1–8:12, jan. 2009. ISSN 1687-4145.

HANAHAN, D.; WEINBERG, R. A. Hallmarks of cancer: The next generation. **Cell**, Elsevier, v. 144, n. 5, p. 646–674, 2011.

HANSEN, B. O. et al. Elucidating gene function and function evolution through comparison of co-expression networks of plants. **Frontiers in Plant Science**, Frontiers Media S.A., v. 5, p. 394, 2014. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4137175/>. Accessed: October 19, 2017.

HEIDEN, M. G. V.; CANTLEY, L. C.; THOMPSON, C. B. Understanding the warburg effect: The metabolic requirements of cell proliferation. **Science**, American Association for the Advancement of Science, v. 324, n. 5930, p. 1029–1033, 2009. ISSN 0036-8075. Available from Internet: <http://science.sciencemag.org/content/324/5930/1029>. Accessed: October 19, 2017.

HOLLAND, E. C. Glioblastoma multiforme: The terminator. **Proceedings of the National Academy of Sciences**, v. 97, n. 12, p. 6242–6244, 2000. Available from Internet: <http://www.pnas.org/content/97/12/6242.short>. Accessed: October 19, 2017.

HORVATH, S. **Weighted Network Analysis: Applications in Genomics and Systems Biology**. [S.l.]: Springer New York, 2011. (SpringerLink : Bücher). Available from Internet: <https://books.google.com.br/books?id=ZCh06NgMFesC>. Accessed: October 19, 2017. ISBN 9781441988195.

HSU, C.-L.; JUAN, H.-F.; HUANG, H.-C. Functional analysis and characterization of differential coexpression networks. **Scientific Reports**, The Author(s) SN -, v. 5, p. 13295 EP –, 08 2015. Available from Internet: <http://dx.doi.org/10.1038/srep13295>. Accessed: October 19, 2017.

IANCU, O. D. et al. Differential network analysis reveals genetic effects on catalepsy modules. **PLoS ONE**, Public Library of Science, San Francisco, USA, v. 8, n. 3, p. e58951, 2013. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3605410/>. Accessed: October 19, 2017.

IDEKER, T.; GALITSKI, T.; HOOD, L. A new approach to decoding life: Systems biology. **Annual Review of Genomics and Human Genetics**, v. 2, n. 1, p. 343–372, 2001. Available from Internet: <http://dx.doi.org/10.1146/annurev.genom.2.1.343>. Accessed: October 19, 2017.

JEANMOUGIN, M. et al. Should we abandon the t-test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. **PLOS ONE**, Public Library of Science, v. 5, n. 9, p. 1–9, 09 2010.

JIANG, L. et al. Microrna-30e* promotes human gbm cell invasiveness in an orthotopic xenotransplantation model by disrupting the nf-κb/iκbα negative feedback loop. **J. Clin. Invest.**, v. 122, n. 1, p. 33–47, 2012.

JIN, X.; HAN, J. K-medoids clustering. In: ____. **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2010. p. 564–565. ISBN 978-0-387-30164-8. Available from Internet: <http://dx.doi.org/10.1007/978-0-387-30164-8_426>. Accessed: October 19, 2017.

JOHNSON, D. B. Finding all the elementary circuits of a directed graph. **SIAM Journal of Computing**, v. 4, n. 1, p. 77–84, 1975.

KANEHISA, M. et al. From genomics to chemical genomics: new developments in kegg. **Nucleic Acids Research**, Oxford University Press, v. 34, n. Database issue, p. D354–D357, 01 2006. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1347464/>. Accessed: October 19, 2017.

KOSCHÜTZKI, D.; SCHREIBER, F. Centrality analysis methods for biological networks and their application to gene regulatory networks. **Gene Regulation and Systems Biology**, Libertas Academica, v. 2, p. 193–201, 2008. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2733090/>. Accessed: October 19, 2017.

KREMPELS, D. **The Genetics of Cancer: Failure of the Regulation of Cell Division and Programmed Cell Death**. 2016. Available from Internet: <http://www.bio.miami.edu/dana/250/250SS15_15.html>. Accessed: October 19, 2016.

LANGFELDER, P.; HORVATH, S. WGCNA: an R package for weighted correlation network analysis. **BMC Bioinformatics**, v. 9, n. 1, p. 1–13, 2008. ISSN 1471-2105. Available from Internet: <http://dx.doi.org/10.1186/1471-2105-9-559>. Accessed: October 19, 2017.

LEE, H. K. et al. Coexpression analysis of human genes across many microarray data sets. **Genome Research**, Cold Spring Harbor Laboratory Press, v. 14, n. 6, p. 1085–1094, 06 2004. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC419787/>. Accessed: October 19, 2017.

LEE, W.-P.; TZOU, W.-S. Computational methods for discovering gene networks from expression data. **Briefings in Bioinformatics**, Oxford University Press, v. 10, n. 4, p. 408–423, 2009.

LEMAIGRE, F. P.; ROUSSEAU, G. G. Transcriptional control of genes that regulate glycolysis and gluconeogenesis in adult liver. **Biochemical Journal**, Portland Press Limited, v. 303, n. 1, p. 1–14, 1994. ISSN 0264-6021. Available from Internet: <http://www.biochemj.org/content/303/1/1>. Accessed: October 19, 2017.

LEMÉE, J.-M.; CLAVREUL, A.; MENEI, P. Intratumoral heterogeneity in glioblastoma: don't forget the peritumoral brain zone. **Neuro-Oncology**, Oxford University Press, v. 17, n. 10, p. 1322–1332, 10 2015. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4578587/>. Accessed: October 19, 2017.

LEMéE, J.-M. et al. Proteomic analysis of glioblastomas: What is the best brain control sample? **Journal of Proteomics**, v. 85, p. 165 – 173, 2013. ISSN 1874-3919. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1874391913002285>. Accessed: October 19, 2017.

LIM, W. K. et al. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. **Bioinformatics**, v. 23, n. 13, p. i282–i288, 2007. Available from Internet: <http://bioinformatics.oxfordjournals.org/content/23/13/i282.abstract>. Accessed: October 19, 2017.

LIU, C.-C. et al. Topological-based classification using artificial gene networks. In: **IEEE ICCI**. [S.l.]: IEEE, 2005. p. 49–56.

MA, S. et al. Incorporating motif analysis into gene co-expression networks reveals novel modular expression pattern and new signaling pathways. **PLOS Genetics**, Public Library of Science, v. 9, n. 10, p. 1–20, 10 2013. Available from Internet: <http://dx.doi.org/10.1371%2Fjournal.pgen.1003840>. Accessed: October 19, 2017.

MA, S. et al. Incorporating gene co-expression network in identification of cancer prognosis markers. **BMC Bioinformatics**, v. 11, n. 271, 2010.

MANTIONE, K. J. et al. Comparing bioinformatic gene expression profiling methods: Microarray and rna-seq. **Medical Science Monitor Basic Research**, International Scientific Literature, Inc., v. 20, p. 138–141, 2014. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4152252/>. Accessed: October 19, 2017.

MARGOLIN, A. A. et al. Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. **BMC Bioinformatics**, BioMed Central, v. 7, n. S-1, 2006.

MENDOZA, M. R. **Exploring ensemble learning techniques to optimize the reverse engineering of gene regulatory networks**. Thesis (PhD) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2014.

MEYER, P. E. et al. Information-theoretic inference of large transcriptional regulatory networks. **EURASIP Journal Bioinformatics and Systems Biology**, v. 2007, 2007.

MILO, R. et al. Network motifs: Simple building blocks of complex networks. **Science**, American Association for the Advancement of Science, v. 298, n. 5594, p. 824–827, 2002. ISSN 0036-8075. Available from Internet: <http://science.sciencemag.org/content/298/5594/824>. Accessed: October 19, 2017.

MOVAHEDI, S. et al. Comparative co-expression analysis in plant biology. **Plant, Cell & Environment**, Blackwell Publishing Ltd, v. 35, n. 10, p. 1787–1798, 2012. ISSN 1365-3040.

MRUGALA, M. M. Advances and challenges in the treatment of gbm: a clinician's perspective. **Discov. Med.**, v. 15, n. 83, p. 221–230, 2013.

NAKANISHI, C.; TOI, M. Nuclear factor-kappab inhibitors as sensitizers to anticancer drugs. **Nat. Rev. Cancer**, v. 5, n. 4, p. 297–309, 2005.

National Center for Biotechnology Information - NCBI. **Gene Expression**. 2014. Available from Internet: <http://www.ncbi.nlm.nih.gov/genome/probe/doc/ApplExpression.shtm>. Accessed: March 15, 2014.

National Center for Biotechnology Information - NCBI. **LDHA – lactate dehydrogenase A [Homo sapiens (human)]**. 2016. Available from Internet: <https://www.ncbi.nlm.nih.gov/gene/3939>. Accessed: October 20, 2016.

ORLOWSKI, R. Z.; JR, A. S. B. Nf-kappa b as a therapeutic target in cancer. **Trends Mol. Med.**, v. 8, p. 385–389, 2002.

PARMIGIANI, G. et al. The analysis of gene expression data: An overview of methods and software. In: PARMIGIANI, G. et al. (Ed.). **The Analysis Of Gene Expression Data: Methods and Software**. New York: Spring Verlag, 2003.

PAVLOPOULOS, G. A. et al. Using graph theory to analyze biological networks. **BioData Mining**, BioMed Central, v. 4, p. 10–10, 2011. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3101653/>. Accessed: October 19, 2017.

PENG, X. et al. Detecting network motifs in gene co-expression networks through integration of protein domain information. In: ____. **Methods of Microarray Data Analysis V**. Boston, MA: Springer US, 2007. p. 89–102. ISBN 978-0-387-34569-7. Available from Internet: <http://dx.doi.org/10.1007/978-0-387-34569-7_7>. Accessed: October 19, 2017.

PRIDDY, K. L.; KELLER, P. E. **Artificial Neural Networks: An Introduction**. Bellingham: SPIE (The International Society for Optical Engineering) Press, 2005.

RAPAPORT, F. et al. Classification of microarray data using gene networks. **BMC Bioinformatics**, v. 8, 2007.

RAYCHAUDHURI, B. et al. Aberrant constitutive activation of nuclear factor kappab in gbmmultiforme drives invasive phenotype. **J. Neurooncol.**, v. 85, n. 1, p. 39–47, 2007.

RITCHIE, M. E. et al. limma powers differential expression analyses for rna-sequencing and microarray studies. **Nucleic Acids Research**, Oxford University Press, v. 43, n. 7, p. e47–e47, 04 2015. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402510/>. Accessed: October 19, 2017.

RUAN, J.; DEAN, A. K.; ZHANG, W. A general co-expression network-based approach to gene expression analysis: comparison and applications. **BMC Systems Biology**, v. 4, n. 1, p. 1–21, 2010. ISSN 1752-0509. Available from Internet: <http://dx.doi.org/10.1186/1752-0509-4-8>. Accessed: October 19, 2017.

SARHAN, A. M. Cancer classification based on microarray gene expression data using dct and ann. **Journal of Theoretical and Applied Information Technology**, v. 6, n. 2, p. 208–216, 2009.

SCHLITT, T.; BRAZMA, A. Current approaches to gene regulatory network modelling. **BMC Bioinformatics**, BioMed Central, London, v. 8, n. Suppl 6, p. S9–S9, 2007. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1995542/>. Accessed: October 19, 2017.

SCHWARTZ, G. K.; SHAH, M. A. Targeting the cell cycle: A new approach to cancer therapy. **Journal of Clinical Oncology**, v. 23, n. 36, p. 9408–9421, 2005. Available from Internet: <http://jco.ascopubs.org/content/23/36/9408.abstract>. Accessed: October 19, 2017.

SERIN, E. A. R. et al. Learning from co-expression networks: possibilities and challenges. **Frontiers in Plant Science**, v. 7, n. 444, 2016. ISSN 1664-462X. Available from Internet: <http://www.frontiersin.org/plant_systems_and_synthetic_biology/10.3389/fpls.2016.00444/abstract>. Accessed: October 19, 2017.

SHEN-ORR, S. S. et al. Network motifs in the transcriptional regulation network of escherichia coli. **Nat Genet**, v. 31, n. 1, p. 64–68, 05 2002. Available from Internet: <http://dx.doi.org/10.1038/ng881>. Accessed: October 19, 2017.

SOMASUNDARAM, K.; NIJAGUNA, M. B.; KUMAR, D. M. Glioma proteomics: Methods and current perspective. In: ____. **Brain Tumors - Current and Emerging Therapeutic Strategies**. Rijeka: InTech, 2011–08–23. p. Ch. 0. Available from Internet: <http://dx.doi.org/10.5772/24710>. Accessed: October 19, 2017.

STATNIKOV, A. et al. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. **Bioinformatics**, Oxford University Press, Oxford, UK, v. 21, n. 5, p. 631–643, mar. 2005. ISSN 1367-4803. Available from Internet: <http://dx.doi.org/10.1093/bioinformatics/bti033>. Accessed: October 19, 2017.

STEKEL, D. **Microarray Bioinformatics**. Cambridge: Cambridge University Press, 2003.

STURROCK, M. **Spatio-temporal modelling of gene regulatory networks containing negative feedback loops**. Thesis (PhD) — University of Dundee, Dundee, 2013.

STåHLBERG, A.; KUBISTA, M.; ÅMAN, P. Single-cell gene-expression profiling and its potential diagnostic applications. **Expert Review of Molecular Diagnostics**, v. 11, n. 7, p. 735–740, 2011. Available from Internet: <http://dx.doi.org/10.1586/erm.11.60>. Accessed: October 19, 2017.

TCHOGHANDJIAN, A. et al. Identification of non-canonical nf-κb signaling as a critical mediator of smac mimetic-stimulated migration and invasion of gbm cells. **Cell Death Dis.**, v. 4, n. e564, 2013.

TEJERA, E.; BERNARDES, J.; REBELO, I. Co-expression network analysis and genetic algorithms for gene prioritization in preeclampsia. **BMC Medical Genomics**, v. 6, n. 1, p. 1–10, 2013. ISSN 1755-8794. Available from Internet: <http://dx.doi.org/10.1186/1755-8794-6-51>. Accessed: October 19, 2017.

The Cancer Genome Atlas (TCGA) Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. **Nature**, v. 455, n. 7216, p. 1061–1068, 10 2008. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2671642/>. Accessed: October 19, 2017.

THIEFFRY, D.; THOMAS, R. Qualitative analysis of gene networks. **Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing**, p. 77—88, 1998. ISSN 2335-6936. Available from Internet: <http://europepmc.org/abstract/MED/9697173>. Accessed: October 19, 2017.

TRAPNELL, C. Defining cell types and states with single-cell genomics. **Genome Research**, Cold Spring Harbor Laboratory Press, v. 25, n. 10, p. 1491–1498, 10 2015. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4579334/>. Accessed: October 19, 2017.

TUSHER, V. G.; TIBSHIRANI, R.; CHU, G. Significance analysis of microarrays applied to the ionizing radiation response. **Proceedings of the National Academy of Sciences of the United States of America**, The National Academy of Sciences, v. 98, n. 9, p. 5116–5121, 04 2001.

VAULONT, S.; VASSEUR-COGNET, M.; KAHN, A. Glucose regulation of gene transcription. **Journal of Biological Chemistry**, v. 275, n. 41, p. 31555–31558, 2000. Available from Internet: <http://www.jbc.org/content/275/41/31555.short>. Accessed: October 19, 2017.

VIALANEIX, N. V. et al. The structure of a gene co-expression network reveals biological functions underlying eqtls. **PLoS ONE**, Public Library of Science, v. 8, n. 4, p. 1–13, 04 2013. Available from Internet: <http://dx.doi.org/10.1371%2Fjournal.pone.0060045>. Accessed: October 19, 2017.

XULVI-BRUNET, R.; LI, H. Co-expression networks: graph properties and topological comparisons. **Bioinformatics**, Oxford University Press, v. 26, n. 2, p. 205–214, 01 2010. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2804297/>. Accessed: October 19, 2017.

YANG, Y. et al. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. **Nature communications**, v. 5, p. 3231–3231, 2014. Available from Internet: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3951205/>. Accessed: October 19, 2017.

ZANOTTO-FILHO, A. et al. Proteasome inhibitor mg132 induces selective apoptosis in glioblastoma cells through inhibition of pi3k/akt and nfkappab pathways, mitochondrial dysfunction, and activation of p38-jnk1/2 signaling. **Invest New Drugs**, v. 30, n. 6, p. 2252–2262, 2012.

ZANOTTO-FILHO, A. et al. The curry spice curcumin selectively inhibits cancer cells growth in vitro and in preclinical model of glioblastoma. **J Nutr Biochem.**, v. 23, n. 6, p. 591–601, 2012.

ZANOTTO-FILHO, A. et al. Nfκb inhibitors induce cell death in glioblastomas. **Biochem.Pharmacol.**, v. 81, n. 3, p. 412–424, 2011.

ZANOTTO-FILHO, A. et al. The pharmacological nfkappab inhibitors bay117082 and mg132 induce cell arrest and apoptosis in leukemia cells through ros-mitochondria pathway activation. **Cancer Lett.**, v. 288, n. 2, p. 192–203, 2010.

ZANOTTO-FILHO, A. et al. The nf kappa b-mediated control of rs and jnk signaling in vitamin a-treated cells: duration of jnk-ap-1 pathway activation may determine cell death or proliferation. **Biochem Pharmacol**, v. 77, n. 7, p. 1291–1301, 2009.

ZHANG, A. **Advanced Analysis of Gene Expression Microarray Data**. [S.l.]: World Scientific, 2006. Available from Internet: <https://books.google.com.br/books?id=6QLCJQ25VUwC>. Accessed: October 19, 2017. ISBN 9789812566454.

ZHANG, B.; HORVATH, S. A general framework for weighted gene coexpression network analysis. **Statistical Applications in Genetics and Molecular Biology**, v. 4, n. 17, 2005.

APPENDIX A — APOPTOSIS – LIST OF GENES

A.1 Activators Genes

APAF1, BAD, BAX, BID, CAPN1, CAPN2, CAPNS1, MCH4, CASP3, CASP6, CASP7, CASP8, CASP9, CAST, CFLAR, CYCS, DFFA, DFFB, FADD, FAS1, FASLG, RIPK1, TNF, TNFRSF10A, TNFRSF10B, TNFRSF10C, TNFRSF10D, TNFRSF1A, TNFSF10, TP53, TRADD, TRAP3.

A.2 Inhibitors Genes

AKT1, AKT2, AKT3, ATM, BCL2, BCL2L1, BIRC2, BIRC3, BIRC4, CHUK, CSF2RB, IKBKB, IKBKG, IL1A, IL1B, IL1R1, IL1RAP, IL3, IL3RA, IRAK1, IRAK2, IRAK4, MAP3K14, MYD88, NFKB1, NFKB2, NFKBIA, NGFB, NTRK1, PDCD8, PIK3C2A, PIK3C2B, PIK3C2G, PIK3C3, PIK3CA, PIK3CB, PIK3CD, PIK3CG, PIK3R1, PIK3R2, PIK3R3, PIK3R4, PIK3R5, PPP3CA, PPP3CB, PPP3CC, PPP3R1, PRKCA, PRKACB, PRKACG, PRKAR1B, PRKAR2A, PRKAR2B, RELA.

APPENDIX B — GLUCOLYSIS – LIST OF GENES

B.1 Activators Genes

ACSS1, ACSS2, ADH1A, ADH1B, ADH1C, ADH4, ADH5, ADH6, ADH7, ADPGK, AKR1A1, ALDH1A3, ALDH1B1, ALDH2, ALDH3A1, ALDH3A2, ALDH3B1, ALDH3B2, ALDH7A1, ALDH9A1, ALDOA, ALDOB, ALDOC, BPGM, DLAT, DLD, ENO1, ENO2, ENO3, FBP1, FBP2, G6PC, G6PC2, G6PC3, GALM, GAPDH, GAPDHS, GCK, GPI, HK1, HK2, HK3, HKDC1, LDHA, LDHAL6A, LDHAL6B, LDHB, LDHC, MINPP1, PCK1, PCK2, PDHA1, PDHA2, PDHB, PFKL, PFKM, PFKP, PGAM1, PGAM2, PGAM4, PGK1, PGK2, PGM1, PGM2, PKLR, PKM, TPI1.

B.2 Inhibitor Gene

C12orf5.

APPENDIX C — CELL CYCLE – LIST OF GENES

C.1 Activators Genes

ABL1, ANAPC1, ANAPC10, ANAPC11, ANAPC13, ANAPC2, ANAPC4, ANAPC5, ANAPC7, BUB1, BUB1B, BUB3, CCNA1, CCNA2, CCNB1, CCNB2, CCNB3, CCND1, CCND2, CCND3, CCNE1, CCNE2, CCNH, CDC14A, CDC14B, CDC16, CDC20, CDC23, CDC25A, CDC25B, CDC25C, CDC26, CDC27, CDC45, CDC6, CDC7, CDK1, CDK2, CDK4, CDK6, CDK7, CREBBP, CUL1, DBF4, E2F1, E2F2, E2F3, E2F4, E2F5, EP300, ESPL1, FZR1, GSK3B, HDAC1, HDAC2, MCM2, MCM3, MCM4, MCM5, MCM6, MCM7, MDM2, MYC, ORC1, ORC2, ORC3, ORC4, ORC5, ORC6, PCNA, PLK1, PRKDC, RAD21, SKP1, SKP2, SMAD2, SMAD3, SMAD4, SMC1A, SMC1B, SMC3, STAG1, STAG2, TFDP1, TFDP2, TTK, ZBTB17.

C.2 Inhibitors Genes

ATM, ATR, CHEK1, CHEK2, RB1, RBL1, RBL2, RBX1, SFN, TP53, CDKN1A, CDKN1B, CDKN1C, CDKN2A, CDKN2B, CDKN2C, CDKN2D, GADD45A, GADD45B, GADD45G, MAD1L1, MAD2L1, MAD2L2, PKMYT1, PTTG1, PTTG2, WEE1, WEE2.

C.3 Dual Genes

TGFB1, TGFB2, TGFB3, YWHAB, YWHAE, YWHAG, YWHAH, YWHAQ, YWHAZ.

APPENDIX D — NFκB – LIST OF GENES

D.1 Activators Genes

BCL3, CHUK, IKBKB, IKBKG, NFKB1, NFKB2, REL, RELA, RELB.

D.2 Inhibitors Genes

NFKBIA, NFKBIB, NFKBIE.

D.3 Target Genes

CCL2, ICAM1, IL1A, IL1B, CXCL8 (IL8), MMP2, MMP9, SELE, TNFα and VCAM1.

APPENDIX E — RESUMO ESTENDIDO

Este capítulo apresenta um resumo das principais contribuições desta tese. Primeiramente, o contexto do trabalho de pesquisa é discutido. Em seguida, são descritos os métodos de análise e classificação propostos. Por fim, os principais resultados são relatados e discutidos.

E.1 Contextualização

A expressão gênica fornece informações para a construção de modelos de sistemas biológicos. A análise de expressão genética comparando tecidos normais e neoplásicos tem sido utilizada para identificar genes associados com a gênese tumoral e potenciais alvos terapêuticos (PARMIGIANI et al., 2003). Tecnologias genômicas de alta performance, como *microarrays*, podem facilitar consideravelmente o perfil molecular de tumores. Milhares de genes podem ser analisados usando um único chip de hibridização de um *microarray* (STEKEL, 2003). O perfil de expressão de um tumor reflete o estado de eventos de um tecido tumoral num determinado momento. Para generalizar os achados e fornecer evidências conclusivas sobre uma alteração molecular, muitas vezes é necessário analisar centenas de tumores. Usando a metodologia molecular tradicional, tal verificação pode levar vários meses, ou mesmo anos. Para facilitar a pesquisa em grande escala, novas técnicas são necessárias.

Uma das principais áreas de pesquisa em biologia de sistemas refere-se à análise de caminhos biológicos ou descoberta de redes a partir de um conjunto de dados de *microarrays*. Uma rede de co-expressão de genes (GCN) pode ser inferida a partir de experimentos de *microarrays* (DAS et al., 2009). Uma GCN é um grafo não direcionado, em que os nodos representam genes e as arestas, por sua vez, representam relações significativas de co-expressão. Em uma GCN, dois genes são conectados por uma aresta se suas atividades tiverem associação significativa (geralmente quantificada por correlação ou informação mútua) considerando uma série de medidas de expressão gênica em diferentes condições ou em diferentes momentos. Comparada com uma rede regulatória de genes, uma GCN não tenta representar relações causais diretas entre os genes participantes na forma de arestas direcionadas. Embora existam outras técnicas estatísticas para analisar a correlação, a rede provê uma linguagem particularmente intuitiva para os biólogos e permite analogias com redes sociais (LANGFELDER; HORVATH, 2008).

Esta tese apresenta uma nova maneira de analisar os dados de *microarrays*, com base nos diferentes tipos de ciclos encontrados entre os genes da GCN construída utilizando dados quantizados obtidos a partir dos *microarrays*. A entrada do método de análise é formada pelos dados brutos, um conjunto de genes de interesse (por exemplo, genes de uma via conhecida) e a função (ativador ou inibidor) destes genes. A saída do método é um conjunto de ciclos. Um ciclo é um caminho fechado com todos os vértices distintos (exceto o primeiro e o último) (DIESTEL, 2012). Graças à nova forma de encontrar relações entre genes, é possível uma interpretação mais robusta das correlações de genes, porque os ciclos são associados a mecanismos de *feedback*, muito comuns em redes biológicas. A hipótese é que o *feedback* negativo permitiria encontrar relações entre os genes que poderiam ajudar a explicar a estabilidade do processo regulatório dentro da célula. Ciclos de *feedback* positivo, por outro lado, poderiam mostrar o desequilíbrio que uma determinada célula está sofrendo em um dado estado. A análise baseada em ciclos permite identificar a relação estequiométrica entre os genes da rede. Assim, esta metodologia permite compreender melhor a biologia do tumor e pode permitir o desenvolvimento de terapias de tratamento mais adequadas que atuem de forma eficaz sobre componentes problemáticos. Além disso, os ciclos ajudam a diferenciar, mensurar e explicar os fenômenos identificados em tecido saudável e tecido doente. Adicionalmente, os ciclos podem também ser utilizados como um novo método para a classificação de amostras de um *microarray* (diagnóstico de câncer). A principal vantagem de seu uso para este propósito é o significado biológico embutido.

Quatro redes fortemente relacionadas com o câncer - apoptose, glicólise, ciclo celular e NFκB - foram utilizadas como estudos de caso das metodologias propostas. Estas redes foram comparadas em tecidos de tumor cerebral Gliobastoma multiforme (GBM), conhecido por ter ativação alterada nessas vias e cujos mecanismos não são completamente entendidos, e em tecidos saudáveis . GBM é o tipo mais agressivo de tumor cerebral. A maioria dos pacientes com GBM morrem em menos de um ano e, essencialmente, nenhum paciente tem sobrevivência a longo prazo, por isso estes tumores atraem atenção significativa. O prognóstico do GBM ainda é extremamente pobre e a descoberta de novos alvos terapêuticos moleculares pode ser importante para otimizar as estratégias de tratamento (MRUGALA, 2013).

Os principais resultados mostram que a relação estequiométrica entre os genes envolvidos nas quatro vias é desequilibrada nas amostras de GBM em relação as amostras de controle. Este desequilíbrio pode ser mensurado e explicado pela identificação de um percentual maior de ciclos positivos nestas redes. Esta conclusão ajuda a entender mais sobre a biologia deste tipo de tumor. O método de classificação proposto obteve as mesmas métricas de desempenho que uma rede neural. Entretanto, enquanto a rede neural é uma caixa preta, o método proposto tem ganho de informação em relação à análise de dados. Se a via analisada está relacionada à doença, o método pode explicar a ação da doença através de ciclos, uma vez que estes estão associados com mecanismos de *feedback*. Portanto, o novo método permite a classificação e também ajuda a explicar a biologia do tumor. Assim, poderia prover dicas para bioquímicos sobre possíveis experiências de laboratório, bem como sobre potenciais genes alvo de medicações.

Portanto, as principais contribuições da tese são: (i) um novo método de análise baseado em ciclos; (ii) um novo método de classificação; (iii) e, finalmente, a aplicação dos métodos e a obtenção de resultados práticos.

E.2 Método de Análise

O método de análise proposto é explicado em duas fases: (i) a construção de GCNs usando *microarrays* e a obtenção dos ciclos; e (ii) a análise baseada em ciclos.

E.2.1 Fase I – Grafos e Ciclos

O método de análise começa com as seguintes etapas, que são mostradas na Figura E.1, para a construção de GCNs e obtenção de ciclos:

1) Pré-processamento de dados de *microarrays* Affymetrix. Essa etapa consiste em importar os dados brutos e definir os valores de expressão para cada conjuntos de sondas. Essas operações são suportadas no pacote *affy* do Bioconductor. A função *ReadAffy* foi usada importar os dados e a função *mas5* para normalizar os valores de expressão. Lim et al. (2007) sugerem que MAS5 permite a construção de uma rede mais fiel. Vale ressaltar que os dados de medições de expressão gênica de alta performance são afetados por um nível relativamente alto de ruído (BOCCALETTI; LATORA; MORENO, 2010).

2) Anotação de dados. A finalidade da anotação é fornecer informações detalhadas sobre os dados. Estas operações são suportadas no Bioconductor pelos pacotes *annotate* e *hgu133plus2.db*. Os nomes dos genes foram extraídos com a função *featureNames* e os símbolos dos genes com a função *getSYMBOL*. Para cade gene envolvido na via de interesse, foi criado um registro com o nome, o símbolo, a função (ativador ou inibidor) do genes e o valor de expressão calculado na etapa 1. Os registros gerados foram classificados em amostras de GBM e amostras de controle.

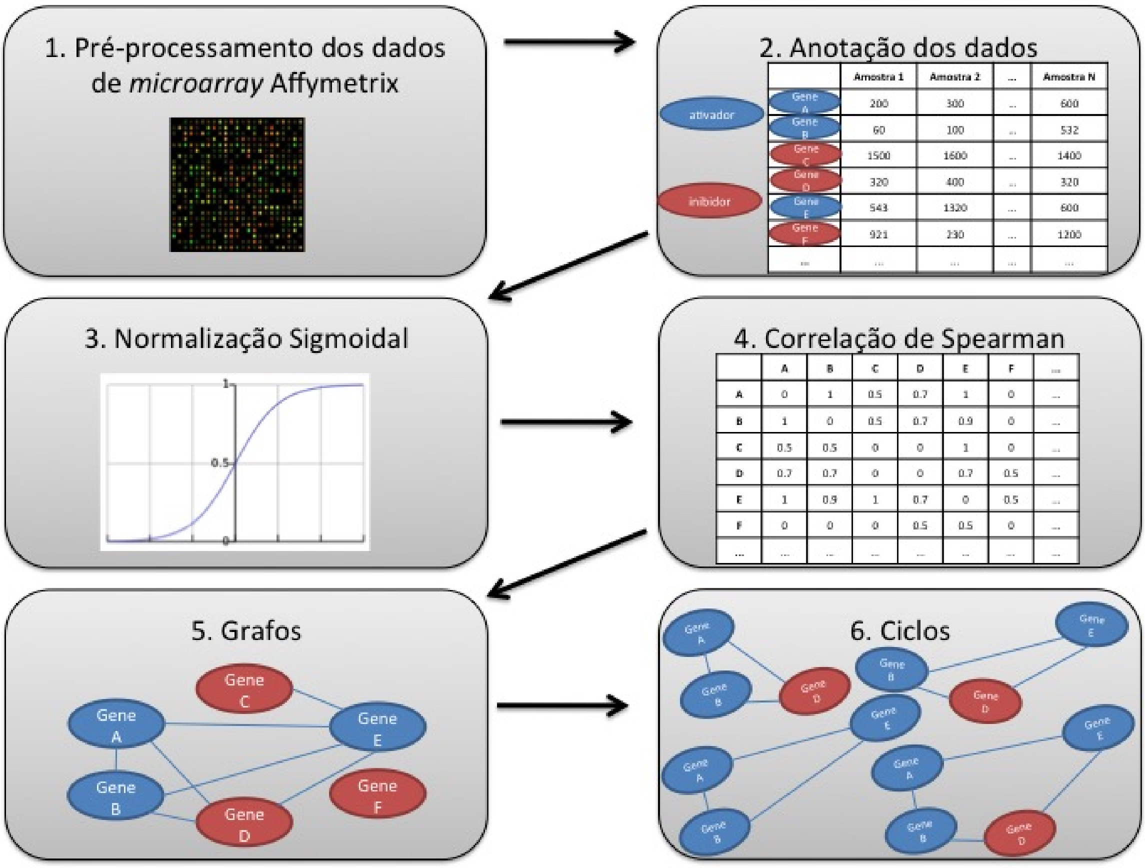


Figure E.1 – Visão geral da análise de redes de co-expressão usando ciclos.
Fonte: Autora (2017).

3) Normalização Sigmoidal. Esta etapa reduz a influência de valores extremos ou *outliers* nos dados sem removê-los do conjunto de dados. Os dados são transformados de forma não linear usando uma função sigmoidal (PRIDDY; KELLER, 2005) e os valores normalizados variam de 0 a 1.

4) Correlação de Spearman. A correlação é usada para descobrir conjuntos de genes com perfis de expressão semelhantes. Métodos de correlação tem sido amplamente utilizados para analisar dados de expressão gênica. Os genes com perfis de expressão correlacionados podem corresponder a complexos protéicos, vias ou participar em circuitos reguladores e de sinalização (HORVATH, 2011). O coeficiente de correlação de Spearman não é paramétrico e permite identificar se duas variáveis (genes) se relacionam em uma função monotônica. A função *cor* foi utilizada (especificando o parâmetro para o tipo de correlação como *method* = *spearman*) para a criação da matriz de correlação de valores de expressão entre os genes selecionados.

5) Grafos. Os grafos não direcionados (que representam a GCN) são construídos calculando um coeficiente de correlação para cada par de genes. Se o coeficiente estiver acima de um certo limiar e for estatisticamente significativo ($p < 0,05$), o par de genes é conectado no grafo; se não, ele permanece desconectado. Foram usados grafos não direcionados ponderados representados por uma matriz de adjacência simétrica $A = [a_{ij}]$, onde i e j são nodos e representam genes e a adjacência a_{ij} tem valores $0 < a_{ij}\ leq 1$ e representa o coeficiente de correlação entre i e j; se i e j estão conectados, é maior que zero; e 0 caso contrário. Os elementos da diagonal foram definidos como 0 para facilitar a notação. O pacote R *igraph* foi usado para obter estatísticas dos grafos e a função *graph.adjacency* foi usada para criar os grafos.

6) Ciclos. A fim de procurar a explicação biológica da associação de genes observados, buscaram-se os ciclos na rede de genes. Uma implementação C ++ do algoritmo de Johnson (JOHNSON, 1975) foi usada para encontrar os ciclos nos grafos. Após a conclusão do algoritmo de busca, os ciclos repetidos são excluídos da análise. Os mecanismos de *feedback* são muito comuns em redes biológicas. A hipótese é que o *feedback* negativo permitiria encontrar relações entre os genes que poderiam ajudar a explicar a estabilidade do processo regulatório dentro da célula. Ciclos de *feedback* positivo, por outro lado, poderiam mostrar o desequilíbrio que uma determinada célula está sofrendo em um dado estado. Os genes de interesse podem ser de dois tipos: ativadores e inibidores. Assumiu-se que um ciclo é positivo quando o número de inibidores no mesmo é zero. Similarmente, é dito negativo quando o número de inibidores no ciclo é maior ou igual a um. A identificação de ciclos é um processo computacionalmente caro. Desta forma, pode ser necessário limitar o tamanho dos ciclos para tornar o método computacionalmente viável. O tamanho máximo a ser considerado depende do tamanho da rede que está sendo analisada e da via sob investigação. Nesta tese, ciclos de tamanho pequeno (até o tamanho dez) foram suficientes para mostrar diferenças significativas entre as amostras de controle e GBM.

E.2.2 Fase II – Análise

Uma vez que os grafos e os ciclos foram obtidos, começa a análise.

1) Gerando estatísticas de grafos. Foram usadas métricas estatísticas que fornecem uma visão geral da estrutura da rede: (i) grau médio - que define o número médio de arestas conectadas a um nodo (uma medida de quantas arestas estão no grafo em comparação com o número de nodos); (ii) densidade - a medida do percentual de arestas conectadas na rede em relação ao total de conexões possíveis; a densidade é definida por um valor decimal entre zero e um; grafos com valores próximos de um são tipicamente densos, enquanto que perto de zero são chamados de esparsos; (iii) diâmetro - este conceito refere-se à medição da distância entre os dois nodos mais distantes da rede; pode ser um indicativo de quão complexa é a rede; por exemplo, um grafo com diâmetro três é geralmente menos complexo do que um com diâmetro sete. Estas medidas podem fornecer números tangíveis que suportam uma avaliação visual da rede, bem como podem ajudar a explorar facetas da rede que são menos óbvias de uma perspectiva visual.

2) Analisando ciclos. Em seguida, os ciclos são considerados. Realiza-se uma análise de subgrupos, que se refere à pesquisa de padrões em subconjuntos das amostras. Este tipo de análise visa comparar os resultados de cada subgrupo particular de amostras. Nesta análise foram considerados dois subconjuntos de amostras: GBM e controle. Portanto, a execução das etapas da seção anterior resulta em um conjunto de ciclos para cada subconjunto de amostras. Cada ciclo é representado por um registro de dados, contendo o símbolo dos genes que compõem o ciclo (ou seja, os nodos), o tamanho do ciclo (determinado pelo número de nodos), a correlação absoluta (produto do peso das arestas) e o tipo de ciclo (positivo ou negativo). Os registros de dados gerados são sumarizados por tipo de ciclo, bem como por tamanho de ciclo. A hipótese considerada é de que o número de ciclos positivos e negativos (isto é, mecanismos de *feedback* positiva e negativa) estão relacionados com a funcionalidade e complexidade de uma determinada via. Para aumentar a confiança na relevância dos ciclos encontrados, é possível considerar bases de dados de amostras de diferentes experimentos. Foram consideradas bases de dados para as amostras de GBM em três experimentos diferentes: um disponível no GEO e dois no TCGA.

3) Encontrando genes alvo. Finalmente, verificou-se quais genes aparecem mais freqüentemente em ciclos. Foram contados os ciclos positivos e negativos em que cada gene aparece em cada subconjunto de amostras e os valores foram comparamos. Na seqüência, as médias do nível de expressão desses genes entre os subgrupos foram comparadas usando o teste t para verificar se eles têm uma diferença estatisticamente significativa. A identificação destes genes proporciona uma forma alternativa de destacar nodos importantes para a rede, bem como permite um ganho de informação em relação à análise do nível de expressão genética nos casos em que a diferença média não é estatisticamente significativa.

Os dados brutos das amostras de 276 GBM de todas as histologias e oito amostras normais de cérebro adulto utilizadas neste estudo estão disponíveis como número de experimento GSE16011 (GRAVENDEEL et al., 2009) no Gene Expression Omnibus (GEO)[1]. Além disso, foram analisadas amostras de GBM de dois estudos independentes disponíveis no on TCGA[2]: 220 amostras caracterizadas em (The Cancer Genome Atlas (TCGA) Research Network, 2008) e 520 amostras caracterizadas em (BRENNAN et al., 2013). Os dados experimentais utilizados na análise estão disponíveis no formato AffymetrixGeneChip Human Genome U133 Plus 2.0 Array. As análises de dados dos microarrays foram realizadas usando R[3] and Bioconductor[4].

E.2.3 Resultados e Discussão

O objetivo do estudo foi detectar ciclos estáveis formados pelos genes básicos de quatro redes fortemente relacionadas com o câncer - apoptose, glicólise, ciclo celular e NFκB - em pacientes com GBM - que são conhecidos por apresentar ativação aberrante nessas vias e mecanismos não completamente compreendidos - comparando-os com controles para compreender melhor a fisiopatologia. A estratégia é baseada na idéia de que uma rede deve apresentar conexões de *feedback* estáveis. Assim, procuram-se estas conexões entre os genes após o pré-processamento. Os principais resultados mostram que a relação estequiométrica entre

[1]http://www.ncbi.nlm.nih.gov/geo/
[2]http://cancergenome.nih.gov/
[3]http://www.r-project.org/
[4]http://www.bioconductor.org/

genes envolvidos na apoptose, glicólise, ciclo celular e NFκB é desequilibrada em amostras GBM. Esta desregulação pode ser medida e explicada pela identificação de uma maior porcentagem de ciclos positivos nestas redes, compreendendo genes de ativadores sem a presença de genes inibidores. Esta conclusão ajuda a entender melhor a biologia deste tipo de tumor.

Os resultados evidenciam as diferenças entre as GCNs das amostras de controle e GBM. Sobre as primeiras três vias, pode-se dizer que os ciclos dos grafos de controle usam todos os genes de cada rede; enquanto os ciclos dos grafos GBM usam um pequeno grupo de genes de cada rede. Além disso, analisando os genes mais comuns encontrados nos ciclos, observou-se que o teste t com nível de significância de 0,05 indicou que não há diferença significativa entre a média do nível de expressão gênica das amostras de controle e GBM de alguns genes das três vias. Assim, há um ganho de informação com a análise usando ciclos em relação à análise do nível de expressão gênica, uma vez que os ciclos destacam a diferença entre as amastras de controle e as amostras GBM. Ainda sobre estas vias:

Apoptose. Apenas alguns ciclos foram encontrados no grafo GBM, o que indicaria que a célula não pode morrer (HANAHAN; WEINBERG, 2011).

Glicólise. LDHA é o gene que codifica uma enzima essencial no processo de metabolismo da glicólise para a formação de ácido láctico. No grafo de controle, este gene aparece em ciclos negativos (mais regulados, já que estes ciclos têm inibidor). No grafo de GBM, ele aparece em ciclos positivos (desregulados, sem inibidor), o que pode estar associado ao metabolismo acelerado da glicólise no tumor e, além disso, gerar acidificação no tumor, o que é bem conhecido por acelerar o crescimento tumoral (HANAHAN; WEINBERG, 2011).

Ciclo celular. O maior número de ciclos foi encontrado no grafo GBM, o que pode indicar que o tumor tem mecanismos mais ativos do ciclo celular, uma vez que é mais proliferado (HANAHAN; WEINBERG, 2011).

Na via do NFκB, os grafos de controle apresentam ciclos negativos (e intrinsecamente estáveis). Nestes ciclos, observa-se a existência de um gene inibidor (NFKBIA, NFKBIB e NFKBIE). Estes ciclos indicam equilíbrio estequiométrico entre ativadores e inibidores, porque existem genes inibidores que regulam os genes ativadores. Com base no conhecimento atual do sistema NFκB a partir de ensaios in-bench, esse equilíbrio implica no bloqueio do fator de transcrição NFκB, fenômeno esperado em tecidos cerebrais normais (RAYCHAUDHURI et al., 2007; ZANOTTO-FILHO et al., 2011). Nas amostras de GBM, a relação estequiométrica entre ativadores e inibidores muda drasticamente. Os ciclos são positivos, consistindo em ativadores.

Não existe um gene inibidor nestes ciclos; consequentemente, o equilíbrio ativador/inibidor parece estar quebrado. Assim, a partir desses resultados, pode-se concluir que há ativação do fator de transcrição NFκB, concordando com os achados dos estudos in-bench de GBM. Particularmente no sistema regulador NFκB, um gene inibidor (NFKBIA, NFKBIB, NFKBIE), por exemplo, não diminui a quantidade do gene ativador; reduz a operação destes produtos de genes, por ligação direta e retenção citoplasmática de ativadores, impedindo assim a sua atividade nuclear (BARKETT; GILMORE, 1999; AGGARWAL, 2004). Assim, com a ausência de um gene inibidor no ciclo, os genes ativadores de NFκB podem migrar para o núcleo e regular a ativação dos seus genes alvo, afetando em última instância a função celular. Portanto, pode-se concluir que os genes inibidores podem participar nos ciclos de amostras de controle, mas não estão envolvidos nos ciclos de amostras de GBM. Desta forma, sugere que os genes NFκB e os genes inflamatórios na amostra de controle permanecem com operação equilibrada, uma vez que existe o *feedback* negativo fornecido pelos inibidores NFκB. Além disso, as amostras GBM ainda mostram um comportamento desequilibrado, sem os genes inibidores não há *feedback* negativo e ativadores e genes inflamatórios não são regulados.

Infelizmente, tem-se um tamanho pequeno de amostra de controle, portanto os resultados podem ser insuficientes e, como conseqüência, algumas relações potencialmente importantes podem ser ignoradas. Estudos futuros devem incluir um maior número de amostras para confirmar esses resultados. A constituição deste grupo de amostras esta relacionada a questão ética da amostragem do tecido cerebral normal. O acesso ao tecido cerebral vivo normal é muito limitado, porque normalmente não é removido durante uma cirurgia de rotina (GRIZZLE; BELL; SEXTON, 2010). Consequentemente, amostras de controle comummente utilizadas em estudos GBM incluem tecido cerebral obtido durante a cirurgia para epilepsia intratável ou das margens da cavidade de ressecção durante a cirurgia de GBM, com a consentimento informado do paciente (LEMéE et al., 2013). No entanto, o uso de tecido adjacente ao tumor como controle deve ser evitado, uma vez que existe a possibilidade de infiltração de células tumorais. Por outro lado, os espécimes GBM são facilmente obtidos, uma vez que são obtidos a partir de tecidos em excesso coletados de pacientes submetidos à ressecação cirúrgica de seus tumores (GRAVENDEEL et al., 2009).

Vários estudos descreveram a heterogeneidade do GBM no nível celular e molecular (LEMÉE; CLAVREUL; MENEI, 2015). Sua heterogeneidade é um tema atual em neuro-oncologia. O glioblastoma é multiforme como o nome implica (HOLLAND, 2000). É multiforme em muitos níveis: grosseiramente, mostrando regiões de necrose e hemorragia; microscopicamente, com regiões de necrose pseudopalisadora, núcleos pleomórficos e células, e proliferação microvascular; e geneticamente, com várias deleções, amplificações e mutações pontuais que levam à ativação de vias de transdução de sinal dos receptores (HOLLAND, 2000). Estes tumores também mostram heterogeneidade genética intratumoral (LEMÉE; CLAVREUL; MENEI, 2015). A heterogeneidade do GBM é uma das razões para sua resistência à intervenção terapêutica (HOLLAND, 2000). Esta heterogeneidade também explica os menores coeficientes de correlação observados entre os genes nas amostras de GBM. Portanto, uma possibilidade interessante de trabalho futuro é analisar os ciclos em amostras GBM subdivididas em subgrupos histológicos e moleculares definidos em (GRAVENDEEL et al., 2009) e (BRENNAN et al., 2013). Espera-se que os grupos mais agressivos tenham coeficientes de correlação mais baixos do que os menos agressivos, possivelmente também apresentam menos ciclos.

O uso de dados de célula única com método proposto é uma possibilidade de trabalho futuro muito interessante. Através de correlações de células únicas interações complexas entre os genes podem ser identificados, por isso melhores sugestões poderiam ser fornecidas aos bioquímicos sobre possíveis experiências de laboratório, bem como melhores genes potenciais alvo de drogas. Para usar dados de célula única com o método proposto, algumas mudanças considerando a característica do novo formato, incluindo uma abundância de zeros, maior heterogeneidade, e distribuições de expressão complexas são necessárias (BACHER; KENDZIORSKI, 2016). Embora a construção das redes e a identificação dos ciclos sejam basicamente as mesmas com os dados de célula única, o pré-processamento do novo formato será diferente. Especialmente, interessante seria comparar os dados obtidos com as amostras de controle com dados de outros experimentos (por exemplo, usando os dados publicados em (DARMANIS et al., 2015)).

Esta tese não descreve um novo software ou método para construir redes. O desafio não é encontrar a rede. Muitos estudos têm sido desenvolvidos nessa direção e hoje a rede pode ser encontrada facilmente (LANGFELDER; HORVATH, 2008; LEE; TZOU, 2009; HACHE; LEHRACH; HERWIG, 2009; ALTAY; EMMERT-STREIB, 2010; BUTTE et al., 2000; MARGOLIN et al., 2006; FAITH et al., 2007; MEYER et al., 2007). Em vez disso, o desafio é entender a rede: explorou-se a estrutura da rede para compreender melhor as relações estabelecidas entre os genes e identificar e quantificar as diferenças entre redes de controle e GBM. Essa maior compreensão da estrutura e das diferenças pode ajudar a prever modelos que resolvam ou impedem o problema (o tumor).

O passo mais importante do método de análise é a identificação de ciclos e sua classificação em ciclos positivos e negativos. Existe uma relação intuitiva entre um ciclo (estrutura no grafo) e um conceito biologicamente importante (*feedback*). O ciclo é responsável pela maioria das propriedades dinâmicas de toda a rede. As dependências cíclicas estão associadas a muitas propriedades fundamentais dos sistemas vivos, como a homeostase, a robustez, a excitabilidade, a multiestationaridade e os ritmos biológicos. Embora o método analise um modelo estático, a análise dos ciclos relacionados à função (ativação e inibição) dos genes - *feedback* positivo e negativo - sugere a operação dinâmica da rede.

Uma enumeração completa de ciclos é computacionalmente caro. O algoritmo de Johnson (JOHNSON, 1975) é a variante mais eficiente e tem uma complexidade temporal que é proporcional ao número de ciclos no grafo onde a constante de proporcionalidade é o número de nodos e arestas. O algoritmo é linear no tamanho da saída, mas exponencial no tamanho da entrada porque o número de ciclos pode aumentar exponencialmente com o tamanho da rede (especialmente com o número de arestas). Embora a identificação dos ciclos possa ser onerosa, sua associação com mecanismos de *feedback* pode permitir uma maior compreensão da estabilidade ou desequilíbrio do processo na célula em um dado estado, justificando assim o uso do método. Dependendo das vias sob investigação, é possível obter diferenças significativas entre as amostras de doença e controle, mesmo considerando apenas ciclos de tamanho pequeno (até o tamanho dez), o que pode ser calculado em tempo viável (dependendo do tamanho da rede).

O método proposto permite contrastar uma rede com outra (por exemplo, controle *versus* GBM). Esta análise de rede diferencial pode ser usada para identificar mudanças nos padrões de ciclos em diferentes condições. Além disso, pode-se encontrar ciclos compartilhados entre duas ou mais redes (análise de ciclos de consenso). Os ciclos de consenso podem representar propriedades estruturais fundamentais da rede. O método proposto de análise e classificação não depende do método de construção da rede, então outro método poderia ser usado para a construção da rede. Assim, nossa metodologia poderia ser usada como um complemento à análise de modularidade. Depois de identificar módulos de GCNs, cada módulo pode ser tomado como uma nova rede, de modo que pode-se procurar ciclos nessas redes para identificar a relação entre os genes e entender melhor o papel de um determinado gene em um módulo de interesse.

Na literatura, alguns artigos focam em como detectar módulos biologicamente significativos (ZHANG; HORVATH, 2005) e padrões recorrentes chamados *motifs* (ALON, 2007) em redes. Langfelder and Horvath (2008), por exemplo, não assumem informações prévias sobre a via e constroem módulos de forma não supervisionada. Eles relacionam módulos com traços clínicos para encontrar módulos clinicamente interessantes, usando conectividade intramodular juntamente com a significância do gene para a detecção de genes hub significativos. A análise proposta está focada nos genes de uma via, de modo que o objetivo não é identificar módulos, vias ou *motifs*, mas sim entender melhor as relações entre os genes da via de interesse e suas variações em amostras de tecidos doentes e de controle. O objetivo geral é obter informações sobre como as alterações nos níveis de expressão podem afetar a ativação da via com base na avaliação dos genes alvo. A abordagem proposta inova usando os ciclos existentes na rede para análise, em vez de usar a conectividade de toda a rede ou a conectividade intramodular como a medida da importância do nodo como outras abordagens (MA et al., 2010), fornecendo assim uma estratégia diferente e potencialmente frutífera para analisar interações complexas em vias.

Vale ressaltar que a metodologia proposta é genérica e pode ser usada para estudar outras redes. Concentrações de proteínas intracelulares chave que são prevalentes em uma gama diversa de importantes processos celulares são controlados por *loops* de *feedback* negativo (STURROCK, 2013). A inflamação e a meiose são exemplos disso. Assim, nossa abordagem para analisar dados de *microarrays* usando ciclos de redes pode ser usada para estudar, por exemplo, esses processos.

E.3 Método de Classificação

O método de classificação de amostras de *microarray* usa os diferentes tipos de ciclos encontrados entre os genes de uma GCN construída utilizando dados quantificados obtidos a partir de *microarrays*. Inicialmente, é preciso aplicar a metodologia descrita na seção anterior. Essas etapas são necessárias para construir a rede (grafo) e para identificar os ciclos que podem ser usados para classificar as amostras do *microarray* de entrada. O método de classificação continua com as seguintes etapas, que são descritas na Figura **??**:

1) **Seleção de ciclos.** Esta etapa seleciona os ciclos com a maior correlação absoluta no grafo de amostras de controle e no grafo de amostras GBM (usamos no máximo dez ciclos, porque não se observou que mais ciclos diminuem o erro absoluto; como trabalho futuro, o número ideal de ciclos pode ser ajustados de acordo com o erro obtido).

2) **Caracterização do ciclo por um sistema de equações.** Cada ciclo selecionado é associado a um sistema de equações. Cada vértice (gene) do ciclo é representado por uma equação, que é definida como uma função de regressão, porque cada aresta indica a existência de uma correlação significativa entre dois vértices (genes). A correlação determina que há uma relação entre os comportamentos dos dois genes; a regressão determina quão forte é a relação.

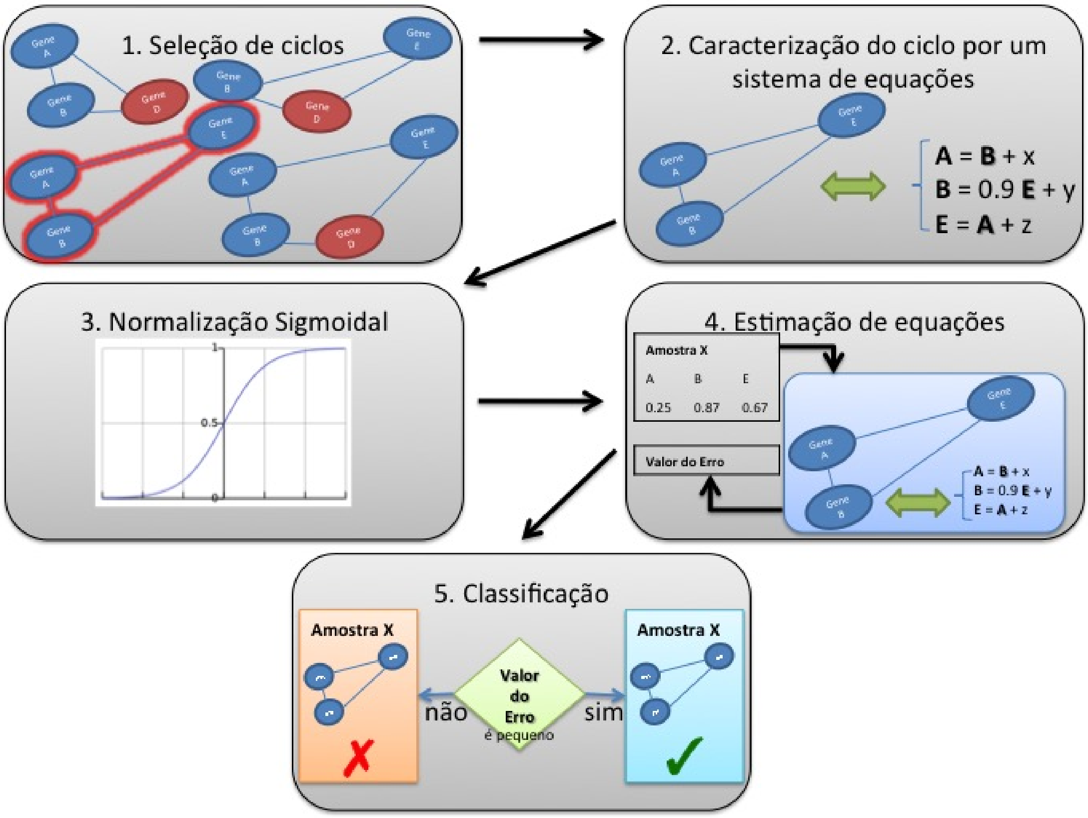


Figure E.2 – Visão geral da classificação de amostras de *microarrays* usando ciclos.
Fonte: Autora (2017).

3) Normalização Sigmoidal. Para a classificação, os dados também precisam ser normalizados, pois os dados do *microarray* foram normalizados para a construção das redes. Os dados são normalizados de duas maneiras: (i) considerando a média e o desvio padrão das amostras de controle e (ii) considerando a média e o desvio padrão das amostras de GBM. Isto é necessário porque não se sabe *a priori* se a amostra que está sendo classificada é de um indivíduo saudável ou paciente, o objetivo deste método é fazer essa diferenciação.

4) Estimando equações. A estimativa do sistema de equações de todos os ciclos seleccionados é realizada utilizando os dados normalizados gerados no passo anterior. Como o sistema de equações é cíclico, algumas iterações são feitas para alcançar a convergência do sistema (definiu-se o número de iterações como 7, porque foi observada uma diminuição no erro absoluto com este número; como trabalho futuro, o número de iterações pode ser definido de acordo com o erro obtido). Em cada iteração, o erro percentual é calculado. Após a última iteração, calcula-se a média do erro absoluto das equações que compõem o sistema.

5) Classificação. Este último passo é baseado no erro calculado no passo anterior. Se o erro for pequeno (diferentes limites devem ser testados para obter resultados mais robustos, o usuário pode definir esse limite), considera-se que a amostra tem o ciclo. Mas para avaliar o erro, deve-se também considerar que os dados foram normalizados de duas maneiras, então tem-se as possibilidades mostradas na Tabela E.1.

Table E.1 – Possibilidades de classificação.

Amostra normalizada como	Erros pequenos em	Erros grandes em	Classificação da amostra
controle	ciclos de controle	ciclos de GBM	controle
controle	ciclos de GBM	ciclos de controle	GBM
GBM	ciclos de controle	ciclos de GBM	controle
GBM	ciclos de GBM	ciclos de controle	GBM

Source: Autora (2017).

Como descrito anteriormente, ss dados brutos das amostras de 276 GBM de todas as histologias e oito amostras normais de cérebro adulto utilizadas neste estudo estão disponíveis como número de experimento GSE16011 (GRAVENDEEL et al., 2009) no Gene Expression Omnibus (GEO)[5]. Além disso, foram analisadas amostras de GBM de dois estudos independentes disponíveis no on TCGA[6]: 220 amostras caracterizadas em (The Cancer Genome Atlas

[5]http://www.ncbi.nlm.nih.gov/geo/
[6]http://cancergenome.nih.gov/

(TCGA) Research Network, 2008) e 520 amostras caracterizadas em (BRENNAN et al., 2013). Os dados experimentais utilizados na análise estão disponíveis no formato AffymetrixGeneChip Human Genome U133 Plus 2.0 Array. As análises de dados dos microarrays foram realizadas usandoR[7] and Bioconductor[8].

E.3.1 Resultados e Discussão

Sobre o resultado do método de classificação, conclui-se que as métricas do método proposto e de uma rede neural são as mesmas na classificação usando os ciclos das redes de apótese, glicólise e ciclo celular para classificar as amostras. Como uma grande vantagem do método proposto sobre a rede neural, pode-se apontar as informações associadas com a classificação. No método proposto, quando um indivíduo é colocado em um determinado grupo, sabemos que ele tem os ciclos utilizados para a classificação deste grupo, por isso há informações biológicas incluídas na classificação. Em redes neurais, por outro lado, a classificação é realizada usando fórmulas que são como caixa preta. Além disso, apesar da paridade no diagnóstico da doença, nosso método é superior a uma rede neural, pois fornece ganho de informação com a análise de dados. Como mencionado anteriormente, o novo método permite a classificação e ajuda a explicar a biologia do tumor: se o conjunto de genes analisado está relacionado com a doença, os ciclos encontrados podem ajudar a explicar a ação da doença, uma vez que estão associados a mecanismos de *feedback*. Assim, os ciclos podem fornecer informações para desenvolver experiências de laboratório, bem como para identificar potenciais genes-alvo para novos medicamentos.

O método proposto também pode ser utilizado para analisar e classificar amostras de GBM com diferentes tipos histológicos. O método convencional de diagnóstico de câncer utiliza uma combinação de exame clínico e exame do aspecto morfológico de espécimes de tecido sob microscopia óptica. No entanto, este método é subjetivo e depende de patologistas altamente treinados. Os conjuntos de dados de *microarrays* podem permitir uma classificação de câncer mais objetiva e altamente precisa. O uso de conjuntos de dados de *microarrays* para determinar o diagnóstico de uma doença que um indivíduo está sofrendo ou prever o prognóstico de um curso de uma terapia ou o resultado de um experimento é uma das áreas mais desafiadoras da pesquisa de *microarrays* (STEKEL, 2003).

[7] http://www.r-project.org/
[8] http://www.bioconductor.org/