OPTIMIZING THERAPEUTIC TARGETS FOR BREAST CANCER USING BOOLEAN NETWORK MODELS

Domenico Sgariglia

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientadores: Carlos Eduardo Pedreira

Fabricio Alves Barbosa da Silva

Rio de Janeiro

Maio de 2024

OPTIMIZING THERAPEUTIC TARGETS FOR BREAST CANCER USING BOOLEAN NETWORK MODELS

Domenico Sgariglia

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientadores: Carlos Eduardo Pedreira
Fabricio Alves Barbosa da Silva

Aprovada por: Prof. Carlos Eduardo Pedreira
Prof. Fabricio Alves Barbosa da Silva
Prof. Geraldo Bonorino Xexeo
Prof. Geraldo Zimbrão Da Silva
Prof. Nicolas Carels
Prof. Carlos Eduardo Ribeiro de Mello

RIO DE JANEIRO,RJ-BRASIL

MAIO DE 2024

Sgariglia, Domenico

Optimizing Therapeutic Targets for Breast Cancer Using Boolean Networks Models / Domenico Sgariglia. – Rio de Janeiro: UFRJ/COPPE, 2024.

XII, 112 p.: il.; 29.7cm.

Orientadores: Carlos Eduardo Pedreira

Fabrício Alves Barbosa da Silva

Tese (doutorado) – UFRJ / COPPE / Programa de Engenharia de Sistemas e Computação, 2024.

Referências Bibliográficas: p. 96-111.

1. Boolean Networks. 2. Systems Biology of Cancer. 3. Gene Regulatory Network Analysis. 4. Epigenetic Landscape Attractors. 5. Apoptosis. I. Pedreira, Carlos Eduardo *et al*. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Agradecimentos

Acontece na vida encontrar uma pessoa que pode inspirá-lo, dar-Ihe confiança, esperança e ajudá-lo concretamente a seguir esse novo caminho, sem nunca nos deixar sozinhos. Para mim, essa pessoa è o Prof. Luis Alfredo Vidal de Carvalho, meu mentor, ispirador e irmão. Obrigado Luis Alfredo, de todo o meu coração.

Um agradecimento especial ao meu orientador, Professor Fabricio Alves Barbosa da Silva, pela confiança e pelo apoio constante, fundamentais para o trabalho realizado e inestimáveis para o meu crescimento pessoal.

Agradeço meu orientador Carlos Eduardo Pedreira por sua disponibilidade durante esse período.

Agradeço ao professor Nicolas Carels, à professora Flavia Raquel Gonçalves Carneiro e à Dra. Alessandra Jordano Conforte por sua valiosa contribuição durante esse período.

Gostaria de agradecer ao Sr. Gutierrez da Costa por sua disponibilidade gentil e constante durante todo esse período.

Ao meu pai Giovanni, que está no céu, e à minha mãe Francesca, por seu amor sempre presente em minha vida.

À minha sogra Anna e ao meu sogro Umberto, que está no céu, que tanto contribuíram para que esse objetivo fosse alcançado.

Às minhas irmãs, Cinzia e Ester, e aos meus cunhados Teresa, Bárbara e Silvestro, que compartilharam comigo as alegrias e os momentos difíceis desse período.

Ao Sr. Alfredo, Sra. Manoelina, Rachel, Israel, Ivy, Ana Luisa e Andres, minha família brasileira, que me acolheu como filho e irmão com amor e carinho.

À Telma e ao André, meus queridos amigos cariocas, pelo apoio e carinho, e pelos momentos agradáveis que passamos juntos.

Ao Renato, por sua ajuda no início dessa aventura empreendida.

Per te Domenica, mia amata sposa

Resumo da tesi apresentada à COPPE/UFRJ como parte dos requisitos necessários para obtenção do grau de Doutor em Ciências (D.Sc.)

OTIMIZAÇÃO DE ALVOS TERAPÊUTICOS PARA CÂNCER DE MAMA USANDO MODELOS DE REDE BOOLEANA

Domenico Sgariglia

Maio/2024

Orientadores: Carlos Eduardo Pedreira
Fabricio Alves Barbosa da Silva
Programa: Engenharia de Sistemas e Computação

As redes reguladoras de genes booleanos permitem a análise dinâmica de sistemas biológicos característicos do funcionamento celular. Sua abstração de alto nível do sistema biológico em estudo é compensada por sua capacidade de fornecer informações úteis sobre sistemas dinâmicos de tamanho considerável.

Esta tese propõe o uso de redes booleanas para modelar uma rede reguladora de genes relacionada ao câncer de mama. Por meio da modelagem dinâmica da rede analisada, foi possível identificar os elementos mais críticos do sistema para a definição de um determinado fenótipo celular relacionado ao câncer.

Além disso, esta dissertação apresenta uma metodologia capaz de otimizar, em nível computacional, o número de alvos identificados na experimentação celular in vitro. As simulações computacionais indicam que ela poderia induzir a morte celular em uma célula cancerosa inibindo um conjunto reduzido de genes-alvo.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

OPTIMIZING THERAPEUTIC TARGETS FOR BREAST CANCER
USING BOOLEAN NETWORK MODELS

Domenico Sgariglia

May/2024

Advisors: Carlos Eduardo Pedreira
Fabricio Alves Barbosa da Silva

Program: Systems Engineering and Computer Science

Boolean gene regulatory networks allow for the dynamic analysis of biological systems characteristic of cellular functioning. Their high-level abstraction of the biological system under study is compensated by their ability to provide helpful information on dynamic systems of considerable size.

This thesis proposes using Boolean networks to model a gene regulatory network related to breast cancer. Through the dynamic modeling of the analyzed network, it was possible to identify the most critical elements of the system for the definition of a particular cellular phenotype related to cancer.

In addition, this dissertation presents a methodology capable of optimizing, at a computational level, the number of targets identified in cell experimentation in vitro. Computational simulations indicate that it could induce cell death in a cancer cell by inhibiting a reduced set of target genes.

Contents

List of Figures

Supplementary materials

For supplementary materials listed in Chapter 3 refer to:

https://github.com/Domenico321/attractors-search

For supplementary materials listed in Chapter 4 refer to:

https://github.com/Domenico321/therapeutic-optimization

CHAPTER 1

INTRODUCTION

The objective of this study is to model a specific computational system related to breast cancer and, consequently, its dynamics. It will also provide useful insights that can be used in a potential therapeutic approach. To this end, this introductory chapter will provide all the conceptual elements used in this research that will allow for easy reading and interpretation of the choices and techniques used throughout the path followed in realizing this thesis.

1.1 PART I – Biological Concepts

1.1.1 - Pathology

The pathology on which the modeling of this research is applied is cancer. It is a group of diseases characterized by unregulated cell growth and the invasion and spread of cells from the site of origin to other sites in the body. Cancer is a genetic disease [Volgestein and Kinzler, 2004]. It is caused by gene changes that control how cells grow and multiply.

The type of cancer investigated in this study is breast cancer, the most commonly occurring cancer in women worldwide [Sung Hyuna et al., 2021]. It is a group of biologically and molecularly heterogeneous diseases originating from the breast. It comprises several biological subtypes with distinct behaviors and responses to therapy. These molecular subtypes are usually divided into five categories [Feng et al., 2018]:

1- Luminal A breast cancer: estrogen receptor (ER) and progesterone-receptor (PR). It accounts for about 40% of all breast cancer.
2- Luminal B breast cancer: Accounting for < 20% of all breast cancer. Luminal B cancer grows slightly faster than luminal A.
3- HER-enriched breast cancer: Accounting for 10%-15% of breast cancer and is characterized by the absence of ER and PR expression.
4- Triple-negative/basal-like breast cancer (TNBC): Accounting for approximately 20% of all breast cancer and is characterized as ER-negative, PR-negative, and

HER2-negative. TNBC usually behaves more aggressively than other types of breast cancer, making it a high-grade breast cancer.

5- Normal-like breast cancer: It is similar to luminal A disease. It is ER and/or PR positive and HER2 negative.

1.1.2 – Hallmarks of cancer

The large number of genes involved in cancer can be organized into a limited number of biological functions, termed Hallmarks of cancer [Hanahan., 2022].

These hallmarks have been proposed as capabilities acquired by human cells in their transition phase from normal to neoplastic growth states, with the aim of providing knowledge capable of rationalizing the complex phenotypes of different human tumor types into a common set of cellular parameters. The hallmarks comprise the capability for sustaining proliferative signaling, evading growth suppression, enabling re plicative immortality, tumor-promoting inflammation, activating invasion and metastasis, inducing or accessing vasculature, genome instability, and mutation, resisting cell death, deregulating cellular metabolism, and avoiding immune destruction. With a view to constant progress in understanding the mechanisms underlying the nature of cancer, new emerging hallmarks have been proposed: unlocking phenotypic plasticity, which is a capability that enables disruptions of cellular differentiation; non mutational epigenetic reprogramming, which involves purely epigenetically regulated changes in gene expression that like DNA mutations, can contribute to the acquisition of hallmarks capabilities during tumor development; polymorphic microbiomes, for which the polymorphic variability in the microbiomes can have a profound impact on cancer phenotypes [Dzutsev et al., 2017] ; senescent cells, seen until now as a protective mechanism against neoplasia, but for which a growing body of evidence instead reveals its ability to stimulate tumor development in certain contexts [Koward et al., 2020]

These common features for each type of cancer are crucial capabilities of a cell in the formation of a malignant tumor. Cancer is daunting in the breadth and scope of its diversity. The concept embodied in these hallmarks is helping to tackle this complexity with the perspective to understand mechanisms of cancer. They constitute an organizing principle for rationalizing the complexities of neoplastic disease.

1.1.3 – Cellular apoptosis

Apoptosis is a highly regulated process of programmed cell death that plays in developmental cells but also controls cell numbers and gets rid of damaged cells [Pecorino, 2012]. It is a type of cell suicide that is intrinsic to the cell, and these characteristics make it an important factor in tumor suppression. In fact, if the apoptotic capacity of a cell is damaged, for example, due to a mutation, this cell will continue to divide without limit, turning into a cancerous cell. Cells can be induced into the process of apoptosis by extracellular signals, also called "death factors," which trigger a series of chain reactions in the cell referred to as “extrinsic pathways”, or by internal physical-chemical causes such as DNA damage or oxidative stress. In this case, the resulting reaction triggered within the cell is commonly called the "Intrinsic pathway".
A group of proteins called caspases plays a central role in both apoptotic pathways in a cascade activation mode, where one caspase activates another in a chain reaction. Another group of proteins, Bcl2 family, turns out to be crucial in the induction of the intrinsic pathway. Some genes in this group promotes apoptosis and others instead inhibits it :[Pecorino, 2012] . The correct balance of these different functions enables the functioning of this essential cellular process.

1.1.4 – Cellular reprogramming

Cellular reprogramming aims to artificially induce changes in a cell phenotype through perturbation of specific genes.

For a long time, biological processes such as differentiation, tumorigenesis, and cellular aging have been thought irreversible. This means that the transition of a cell from one state to another based on genetic or epigenetic mutations has always been seen as a unidirectional phenomenon. Through recent studies [Yamanaka and Blau., 2010], however, it has been shown how this process can become bidirectional, that is, how it is possible to allow the cell to move out of a given phenotype to acquire different functional characteristics. Now, let us contextualize this concept in a carcinogenic context.

Cancer is generally caused by genetic and epigenetic alterations considered irreversible. Experimental evidence [Cho et al., 2017] supports a strategy of reverting cancer cells into normal cells by inducing permanent differentiation.
Cancer reversion involves a cellular reprogramming methodology by which cancer cells lose their malignant properties and acquire the phenotypic characteristic of normal cells, suppressing malignancy [Shin and Cho., 2023].

1.2 PART II – MODELING CONCEPTS

1.2.1 – Network construction

The starting point of the modeling was to find the constituent elements of the adopted network representing the interactions between certain genes. Gene regulatory networks regulate the expression of genes in any given developmental process [Davidson and Levin., 2005]. In this control system, each node in the network receives and integrates multiple inputs in the form of regulatory proteins that may be activating or repressing gene expression, providing as output the transcript value associated with the gene target. This result was achieved through the use of the public repository MSigDB [Liberzon et al., 2015], through which four lists of genes linked with two hallmarks of cancer, "Evasion of cell death" and "Unlimited replicative potential" were taken. This group of genes was then compared with the differentially expressed genes of the MB231 triple-negative breast cancer cell line. This difference was obtained by comparing the expression level of the MB231 cell line with the MCF10 type cell line, representing a noncancer cell line. The choice of the triple negative cancer type was due to its therapeutic difficulty in treating this disease. At the end of this operation, only those genes obtained through the MSigDB repository that were also differentially expressed in the MB231 line were retained. Using the human interactome from the intact-micluster.txt file, the existing interactions between the component genes of the above-selected group were found. An interactome is the total set of molecular interactions in a particular cell. It specifically refers to physical interactions between molecules but also describes sets of indirect interactions between genes [Caldera et al., 2017]. From the interactions found, genes with a number of connections greater than 50 were chosen, considering the greater connectivity of a node within a network as an indicator of its greater influence on the dynamics of the system [Albert, 2005]. Having

performed this filtering operation, transcription factors were associated with these remaining vertices through the use of the online tool TRRUST [Han et al., 2015].

Having defined the relationships existing between the vertices of the gene regulatory network, it was necessary to determine the type of relationship existing between neighboring genes. This means knowing the type of influence exerted by a gene, which can be activation or inhibition of gene expression on its neighbor in which a direct linkage is present. This type of information can be obtained through considerable consultation of existing literature, when it exists, specifically related to verified interactions between two or more genes. The tool through which this curation process was carried out is the Metacore database [Ekins et al., 2007]. Metacore made it possible to identify the type of interaction between the various genes in the network. Each identified interaction is correlated with an indication of the specific existing literature justifying the type of interaction assigned to the analyzed gene pair.

The procedure described above led to the creation of the model related to the investigated pathology, a model represented by a gene regulatory network, which shows the causal nature of the interactions present between the vertices of the network.
This search for causality guided the entire process of building the model.

1.2.2 – Boolean network model construction

The study of the dynamics of a gene regulatory network system can be done from a quantitative or qualitative point of view. In the former case, the tool used is differential equations that offer considerable detail in the description of the phenomenon under investigation but require knowledge of a large number of parameters, making it prohibitive in networks with a large number of nodes.

In the research described in this thesis, we opted for a qualitative investigation of the system dynamics using Boolean networks [Thomas, 1973]. It is an approach based an abstract representation of the system, where every node can take two possible values: zero for inactive and one for active. Inactive or active is indicated as an approximation of the gene expression level of the genes that make up the network at a given time [Schwab et al., 2020].

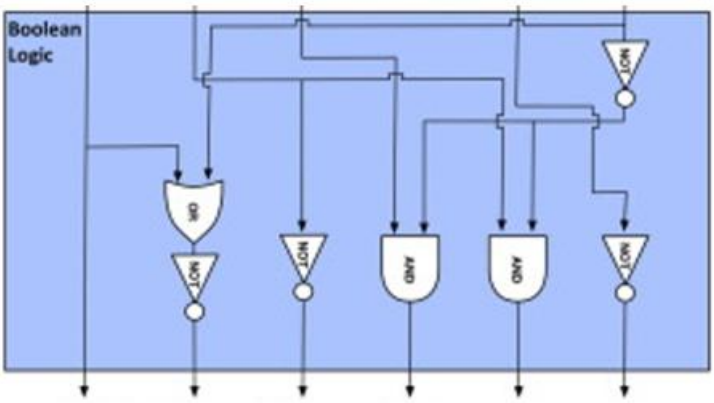

Fig. 1.1 Representation of the network nodes through logical gates of type AND, OR and NOT. Figure adapted from [Schwab et al., 2020].

Figure 1.1 shows how network nodes in the formalism of Boolean networks are equivalent to logical gates of type AND, OR, and NOT. The only output of each port, which is equal to a Boolean value, is the result of the processing implemented by the port on several Boolean input values.

The choice of the type of Boolean function to apply to each node in the network represents a complex task crucial to the result produced by the system. The question sought to be answered is: how can we choose the set of appropriate Boolean functions such that the ensuing network dynamics mimic cell fate behavior? In this work, we choose to use functions of type Nested Canalizing Function. The characteristic feature of this type of function is that a single input or a group of inputs determines the corresponding output [Schwab et al., 2020].

An important aspect of complex dynamical systems, like our Boolean network, is the existence of two dynamical regimes, order and chaotic, and a critical phase transition boundary between the two [Nykter et al., 2008]. Ordered regimes are intrinsically robust with simple dynamics. Contrary to this, networks in chaotic regimes are very sensitive in front of small perturbations, which can propagate on the whole system, preventing the necessary relative robustness for cellular homeostasis. Between these two regimes, there is a third regime named phase transition, which represents a trade-off between the need for stability and the need to have a range of dynamic behavior to respond to a variable environment. Kauffman [Zhou et al., 2013] showed that the Boolean functions that belong to the canalizing functions shift the dynamics of the network from the chaotic to the phase transition. This is because every genes have

less effective upstream regulators than it appears on the network diagram, which improves the robustness of the networks.

Having defined the type of Boolean function to be applied to each node in the network, it remains to determine the method of evolution of the system as time passes. The choice was to allow the system to employ a synchronous evolution mode, in which all nodes in the network are updated simultaneously with each temporal evolution of the system.

1.2.3 – Attractor definition

The model simulated with the Boolean network can reach a stable dynamic behaviour, called attractor, that is interpreted as a physiological endpoint [Wang et al., 2012].

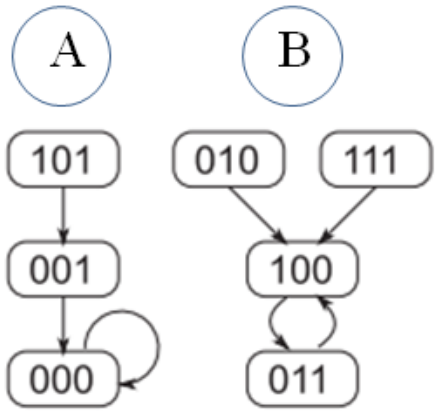

Fig. 1.2 A: Steady-state attractor B: Simple-cycle attractor. Figure adapted from [Mori and Akutsu., 2022].

As shown in Figure 1.2, an attractor is a state of a Boolean network with no outgoing edges in the state transition graph. Steady-state attractors comprise only one state, while cycle attractor is formed by a sequence of states that are periodically repeated. They represent the long -term behavior of the Boolean network, and once they are reached, they cannot be left unless an external perturbation occurs.

The basin of attraction comprises all states which lead to a corresponding attractor.

Now, we can use the Boolean network's attractor concept to represent a specific cellular state. Considering the space state of the constructed gene regulatory network as the space that contains all theoretically possible gene expression patterns of this network, each point in the state space represents the combination of gene expression of all component genes in the system. The attractor state is a set of points in the state space with a particular property: a stable equilibrium. Based on this statement, Kauffman [Kauffman, 1969] proposed that attractor states correspond to the gene expression profiles associated with each cell type. Consequently, if cell types are attractors and cancer cells are viewed as abnormal cells , then cancer cells should also be represented by attractors [Huang et al., 2009].

Based on these considerations, it is possible to define common characteristics between a cellular state and a Boolean-type attractor. Both are:

- Discrete
- Dynamically stable
- Mutually exclusive

These statements provide a basis for using Boolean networks in dynamic modeling of a cellular state geared toward the goal of cellular reprogramming.

1.3 PART III – STRUCTURE OF THE THESIS

1.3.1 – General overview

The strategy adopted in the work presented in this thesis is schematically depicted in the following figure 1.3

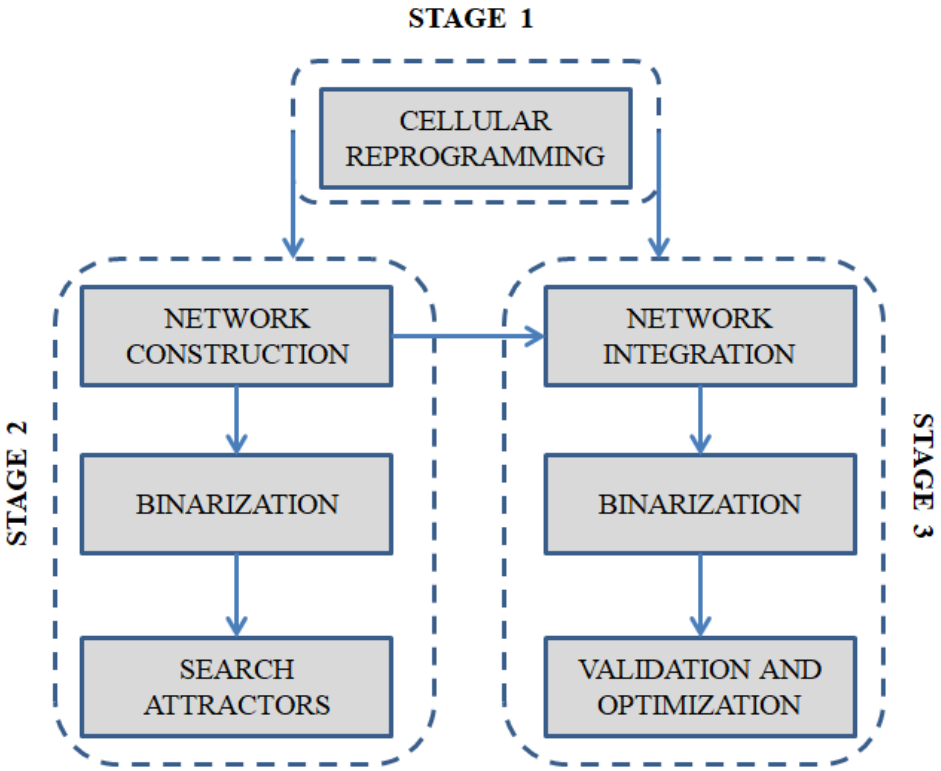

Fig. 1.3 The three implementation phases of this research.

In Stage 1 [Sgariglia et al., 2018], contained in Chapter 2 of the thesis, we provide an overview of cellular reprogramming. The possibility of being able to produce a guided mutation of a cell's evolutionary fate conceptually realized on the epigenetic landscape [Baedke, 2013] that describes its dynamic evolution is the conceptual premise on which the following two stages constituting the research described here are based. Boolean networks allowed the representation of this epigenetic landscape and the possibility of modeling the action on it to guide, in silico, the evolution of a cell's state within this landscape.

In stage 2, the breast cancer gene regulatory network was constructed, RNA-seq data were binarized, and specific attractors related to particular patients were found, identifying the peculiar elements in the system that characterize these attractors.

Finally, stage 3 [Sgariglia et al., 2024] tested the validity of the implemented model against a n actual biological process carried out in vitro and optimized the results obtained in this research [Tilli et al., 2016].

Note the close correlation of the three sections shown in Figure 1.3. Stage 1 represents the theoretical premise on which the following two steps rest, and the starting point of stage 3 is represented by a well-defined step in stage 2.

1.3.2 – Binarization process

RNA, or ribonucleic acid, is a biological macromolecule that plays a central role in protein generation from DNA. Since DNA cannot leave the cell nucleus, it cannot generate a protein. This occurs via the transcription of RNA molecules that code for protein. This quantitative information can be obtained through RNA-seq, which is a technique that uses next-generation sequencing to detect the presence and amount of RNA molecules in a biological sample, providing a snapshot of gene expression in the sample, also called transcriptome [Chaffey et al., 2003].

The breast cancer-related RNA-seq data used in this thesis are Single cell RNA-seq (scRNA-seq) from tumor cells (Stage 2 of figure 1.3) and bulk RNA-seq from in vitro cell culture (Stage 3). The scRNA-seq examines the gene expression level of individual cells in a given population by simultaneously measuring the RNA concentration of hundreds to thousands of genes. It can reveal complex and rare cell populations, uncover regulatory relationships between genes, and track the trajectories of distinct cell lineages [Hwang et al., 2018]. The in vitro cell culture is a biological process reproduced in the laboratory outside the organism. In this case, transcript values are obtained by Bulk RNA sequencing, which is a method of choice for transcriptomic analysis of pooled cell populations. It measures the average expression level of individual genes across hundreds to millions of input cells.

Translating a given gene expression value into its corresponding dichotomous value represents one of the most critical steps in this type of modeling, in which the interpretive component of an observed quantitative phenomenon has a significant bearing on the quality of the final result. In stage 2 we used the BASC algorithm [Hopfensitz et al., 2012], which detects the discriminating thresholds on the data for the attribution of the corresponding dichotomous value.

In this algorithm, there are two approaches, named BASC A and BASC B, which have the following steps in common:

- definition of a step function in which the analyzed data is reorganized in ascending order.
- find the highest discontinuity in the step function.
- estimate the location and variation of the strongest discontinuities.

In BASC A, the step function is calculated to minimize the Euclidean distance from the initial step function. BASC B obtains step functions from smoothened versions of the input function.

In this work, we used the BASC B, due to the best performance shown on the analyzed data.

Having assigned the RNA-seq value to the corresponding gene in the network and transformed it into a Boolean value, the remaining action was to find the attractors related to the specific individual cells of a given patient suffering from the disease under study.

In stage 3, Bulk RNA-seq values assigned to the corresponding genes in the network are derived from cancer cell lines cultured in vitro. The following figure shows schematically how these data were organized and binarized for subsequent dynamic system analysis. Since the data we processed in stage 3 is different from stage 2, we used a different binarization process.

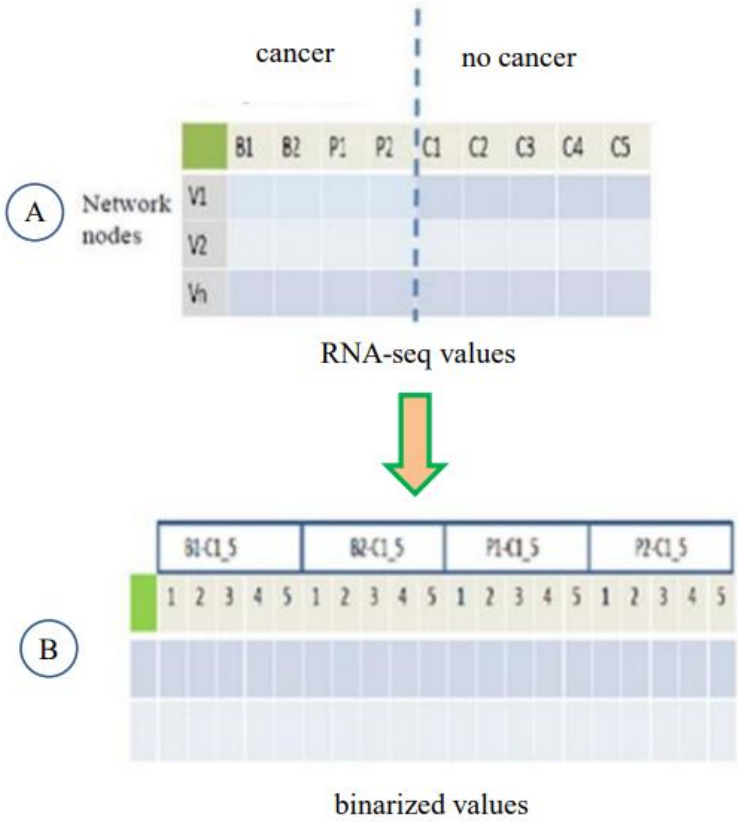

Fig. 1.4 Organization and binarization of gene expression values for stage 3. Part A of the figure represents the original data format, and part B shows the final format after appropriate interventions.

As shown in Section A of Figure 1.5, the gene expression data from in vitro cell cultivation consists of two cancer cell phenotypes, B and P, and one noncancer cell phenotype, C. Subtracting the RNA-seq value of each line C from each line B and P yielded 20 columns (section B) representing the gene expression difference values between a tumor and a nontumor phenotype for each gene in the implemented network. Through appropriate statistical calculation techniques, a discriminating threshold value was assigned to each column. Comparison of the value of the difference in gene expression present in each column position with the corresponding threshold value determines the Boolean value attributed to the specific position relative to the gene and column.

This process resulted in twenty columns representing twenty initial network configurations on which to perform dynamic analysis in search of attractors.

1.3.3 - Search for attractors

Having assigned the RNA-seq value to the corresponding gene in the network and transformed it into a Boolean value, the remaining action was to find the attractors related to the specific individual cells of a given patient suffering from the disease under study.

Finding the attractors of a Boolean network for all possible initial combinations of node values can be a computationally prohibitive task, especially in networks of similar size to the one used in this research. For the search of attractors in this work, an initial configuration of Boolean values representing the analyzed cell state was assigned to the network, which allowed the system to be placed at a point in a basin of attraction of a given attractor. Under these conditions, it was sufficient to calculate the trajectory followed by the system from this initial point to the specific attractor to which it belongs. This methodology allows the search for attractors without excessive computational resources.

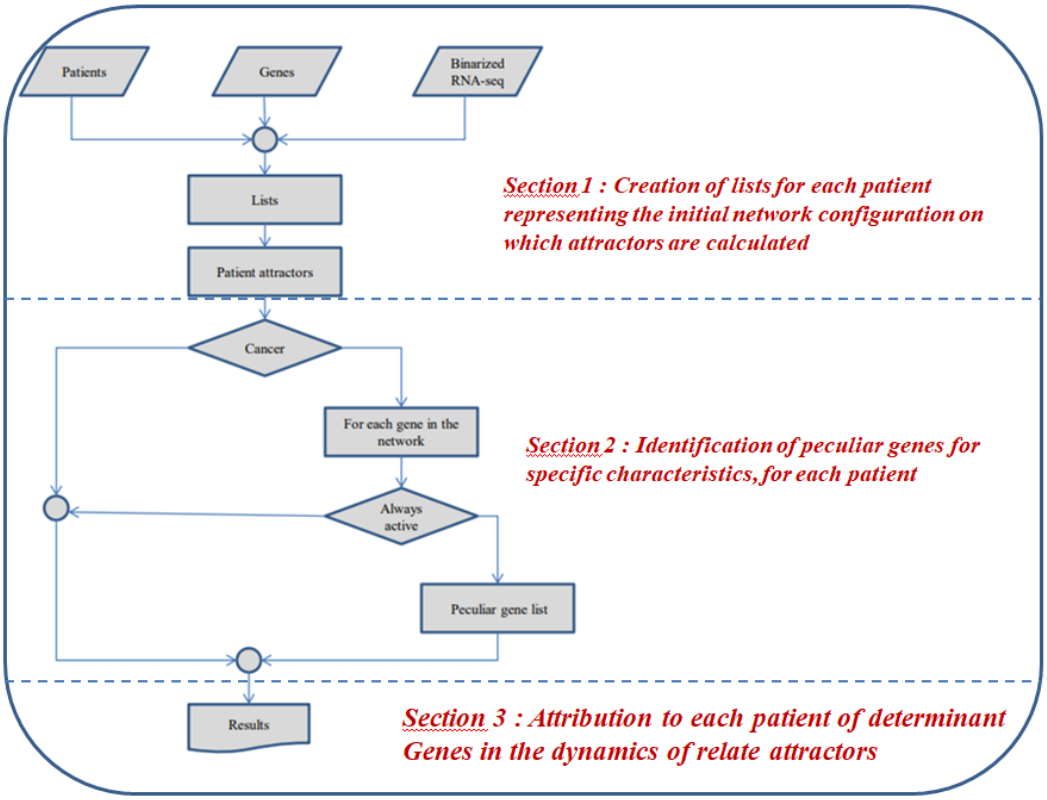

Fig. 1.5 Procedure adopted for the identification of the attractors attributable to each patient, and the specific genes related to this specific patient that characterize the dynamics of his attractors.

Figure 1.4 schematically shows the procedure adopted in stage 2 for identifying attractors and peculiar elements related to them in three separate procedural sections.

In Section 1, lists equal to the number of single cells of the specific patient are created for each patient. In these lists are the Boolean scRNA-seq values assigned to each gene in the network. The attractors of each initial configuration of the network represented by each list are calculated.

In Section 2, genes that show consistent behavior in their expression level in all attractors attributed to a given patient are identified.

In Section 3, the results consist of assigning genes in the network to each patient with a peculiar behavior in each attractor.

This procedure made it possible to study data from different patients with the same disease by focusing the search on the specific elements that could potentially characterize the evolution of the disease differently.

1.3.4 – Validation and optimization

The inspiring principle of stage 3 [Sgariglia et al., 2024] is the result obtained in an in vitro experiment [Tilli et al., 2016] by which a decrease in the proliferation of a breast cancer cell was achieved by inhibiting specific genes. The objective of the research at this stage was to emulate the experiment performed in vitro through a computational model by obtaining the same biologically significant results and trying to improve its performance at the quantitative and qualitative levels. The idea is to guide our biological model in silicon to assume a configuration in attractors and attain a state of cell death.

To reproduce in the model a process of cell apoptosis, a group of genes with a specifically biologically relevant role in the process of cell apoptosis were added to the gene regulatory network developed in stage 2, highlighting how this step represents a point of continuity, as highlighted in Figure 1.3, between the work carried out in stage 2 and the research pursued in stage 3. The functional specificity related to apoptosis of the

group of genes integrated into the model results from a thorough literature search on the topic.

This integration action aims to use these new elements added to the system as indicators of the modeled cellular phenotype once the system dynamics reach a stable state. They represent the role that direct visual verification on the ongoing process can have in the laboratory, in a biological process in silicon, allowing localization within a state space representing the epigenetic landscape.

The configuration in terms of Boolean values of the group of apoptosis-specific genes in the attractors detected as a result of appropriate interventions on the computational model designed to reproduce the methodology adopted in the in vitro experiment [Tilli et al., 2016] allowed verification of the biological compatibility of our model.

Having reproduced the in vitro experiment in terms of biologically relevant results, optimization of these results was sought in terms of fewer genes inhibited in vitro essential for a reduction of cancer cell proliferation and possibly a better configuration of the cancer gene expression group of genes related to apoptosis, expressed with Boolean values in the attractors found.

The strategy adopted to pursue this result is summarized in these two steps:

step 1: Network modularity check
step 2: Apply a specific analysis on the network structure.

The first step was to verify the implemented network's modularity by whether the apoptosis-related genes inserted into the system are concentrated in a single or two modules of the network rather than scattered throughout it. Modules, also known n as communities, essentially are groups of nodes more closely connected than the rest of the network [Raman., 2021]. This property of the network was investigated by the Clauset-Newman-More greedy modularity maximization algorithm [Clauset et al., 2004]. The steps performed by the algorithm are:

1: Assign each node to a community of its own. Therefore, we start with a number of communities equal to the number of nodes in the network.
2: Inspect each pair of communities connected by at least one link and compute the modularity variation obtained if we merge these two communities.

3: Identify the community pairs for which *M*, the value the algorithm gives to each detected community, is the largest, and merge them.

4: Repeat step 2 until all nodes are merged into a single community

5 : Record for each step and select the partition for which the modularity is maximal.

The result obtained through the network modularization procedure, i.e., verification that the genes constituting the apoptosis-related group are concentrated in specific clusters, is the prerequisite for implementing the method to improve the results obtained in *in vitro* experimentation, above defined as step 2.

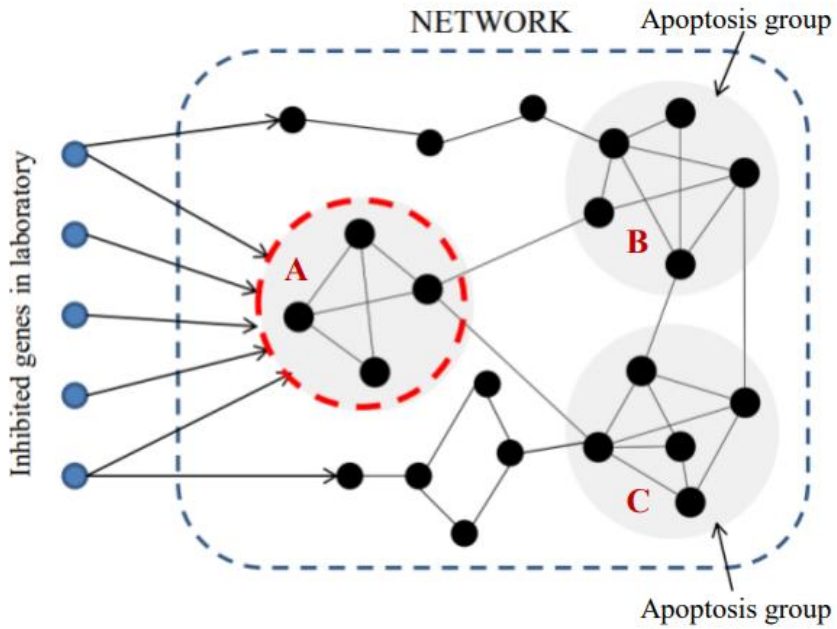

Fig 1.6 Schematization of the objective pursued. Cluster A represents the group of network nodes researched. Clusters B and C are the communities in which apoptosis-related genes are concentrated.

- Detect all shortest paths between each inhibited gene in the laboratory (blue nodes on the left of the figure) and all nodes related to apoptosis, shown in groups B and C.
- Create a list for each inhibited gene in the laboratory (blue nodes in the figure) in which the network nodes participating in at least one shortest path of this inhibited gene are included.

- Assign a score to each node of the network based on the number of lists linked to the inhibited gene (blue node in the figure) in which it is present.

At the end of this procedure, the nodes in the network with the highest score are the genes sought to be inhibited in the computational model instead of the group inhibited in the laboratory.

1.4 PART IV – OBJECTIVE OF THE THESIS

1.4.1 Primary objective

Identification of gene regulatory network vertices as potential therapeutic targets. These potential targets may enable the transition from a given pathological cell phenotype to a nonpathological one.

1.4.2 Complementary objectives

- Identify specific data to build a gene regulatory network related to breast cancer
- Finding specific system attractors through Boolean network modeling and detecting these stable states' peculiar elements.
- Verify the biological compatibility of the constructed model.

CHAPTER 2 [Sgariglia et al.,2018]

Cellular Reprogramming

With cellular reprogramming, it is possible to convert a cell from one phenotype to another without necessarily passing through a pluripotent state. This perspective is opening many interesting fields in the world of research and biomedical applications. This essay provides a concise description of the purpose of this technique, its evolution, mathematical models used, and applied methodologies. As examples, four areas in the biomedical field where cellular reprogramming can be applied with interesting perspectives are illustrated: diseases modeling, drug discovery, precision medicine, and regenerative medicine. Furthermore, the use of ordinary differential equations, Bayesian network, and Boolean network is described in these contexts. These strategies of mathematical modeling are the three main types that are applied in gene regulatory networks to analyze the dynamic interactions between their nodes. Ultimately, their application in disease research is discussed considering their benefits and limitations.

The concept of cellular reprogramming began in the 1960s with the idea of reversing the direction of cell differentiation, which was so far conceived only as occurring in a single irreversible direction. The differentiation of cellular state was schematically described through theWaddington epigenetic landscape [Waddington, 1957; Waddington, 1940], where the metaphorical valleys represent states of cellular stability, and the hills around them represent the epigenetics barriers that prevent the transition from one state to another. The goal of cellular reprogramming is to induce cells to overcome these barriers and move from one stable state (attractor) to another according to the simulations described in this chapter. Among the various scientific advances in this field, one may quote the work done by Takahashi and Yamanaka [Takahashi and Yamanaka, 2006], concerning the generation of induced *pluripotent stem cells* (PSC), as an important reference in the progress of cellular reprogramming. The ability of a cell to reprogram itself from one attractor to another in the epigenetic landscape according to external and internal perturbations, or the overexpression of some key genes, has opened a huge field of investigation in the world of scientific research. Different strategies were followed with the aim of inducing phenotypic cell changes using the different mathematical and biological modeling techniques available.

Technological integration in different scientific areas such as biology, mathematics, statistics, and computational sciences is essential for the success in the simulation of cellular reprogramming. For this reason, the contribution of systems biology is determinant for the success of this emerging field. This chapter first defines cellular reprogramming and its objective. Next, it provides a review of the methods used to achieve cellular reprogramming and the approaches to build the network models analyzed. Lastly, we discuss the applications of cellular reprogramming to diseases, highlighting the benefits and limitations of this technique and its potential application in different areas.

2.1 What is cellular reprogramming?

2.1.1 Premise

We define cellular reprogramming as the conversion of one specific cell type to another one. Eukaryote cells transit from one state to another through changes in geneexpression and, consequently, protein levels in response to signals coming from the extracellular environment. The goal of cellular reprogramming is to artificially induce changes in a cell phenotype through perturbation of specific genes.

Until few years ago, cellular differentiation has long been thought of as “one-way traffic,” without any possibility of returning to a previous cellular state. The idea that a cell could be induced to reverse its differentiated state toward a less specialized one was not even imagined.

The demonstration in 1963 [Siminovitch et al., 1963] of cell dedifferentiation in culture of adult fibroblast through interaction with stem cells of a mouse teratocarcinoma [Martin GR, 1981] was a great step toward the concept that cellular differentiation is, indeed, reversible.

In 2006, Takahashi and Yamanaka induced PSCs from adult fibroblast cultures of mouse under the incubation with the transcriptional factors POU5F1, SOX2, KLF4, and MYC [Takahashi and Yamanaka , 2006].

This remarkable discovery was a milestone for further advances and developments in the cellular reprogramming field. For the first time, it was shown to the scientific community that reversibility in the cell differentiation process was possible.

Mature cells could be reverted to a previous pluripotent state, and it was possible to control the gene expression pattern with few transcription factors.

2.1.2 Meaning of cellular reprogramming

We begin with the mechanism of cell reprogramming by the definition of epigenetic given by Conrad Waddington (Fig. 1): “Epigenetic is the branch of biology that studies the causal interactions between genes and their products, which bring the phenotype into being” [Waddington , 1968]. He conceived the epigenetic landscape as an inclined surface with a cascade of branches ridges, and valleys [Waddington ,1939; Waddington ,1940; Waddington ,1957].

The goal of cellular reprogramming is to bring a cell (the ball of Fig. 1) from a valley of differentiation back to a state of pluripotency or to another differentiated state into a different valley passing a ridge. Following the same logic, it becomes clear that inducing a cell to move from one specialized cell state to another without necessarily passing through the pluripotent state is also possible. Indeed, the transition from a differentiated state toward a progenitor state is referred to as *dedifferentiation*, while the transition between two differentiated states is called *transdifferentiation*.

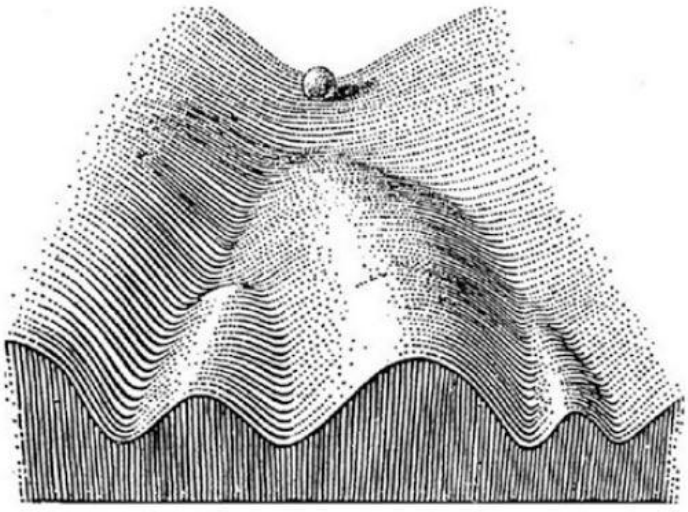

Fig. 2.1 Waddington landscape representation of epigenetic space where the ball that can roll down from an undifferentiated cell state into a specialized state. The branches are the different potential states, and the ridges are the epigenetic barriers that prevent a

cell from taking a different differentiation trajectory than the one in which it is already engaged.

Keeping in mind the Waddington landscape representation described above, we might answer the following two questions:

(a) What are the barriers we must overcome to move from one cellular state to another?
(b) How can we induce cellular state transitions?

Answering the first question, we know that a stable cell state can be seen as a high-dimensional attractor of the gene regulatory network [Huang et al., 2005]. Attractors correspond to stable states associated with specific cell types [Huang et al., 2009].

In this context, cell fates are determined by gene expression and epigenetic patterns controlled by multiple factors [Lang et al., 2014], such as DNA methylation and histone modifications [Seah et al., 2015]. Both modifications can affect gene expression without inducing changes in DNA. DNA methylation involves the addition of methyl groups to the DNA molecule that usually results in the inhibition of eukaryotic gene transcription.

Histone modifications are posttranslational processes that occur in the histone tails, which inhibit or induce local gene expression depending on the modification type [Goldberg et al., 2007].

After illustrating the role of the epigenetic activity that controls cellular states, the second question can be answered: How can we induce state transitions? as outlined above, there are attractors corresponding to different cell fates and different epigenetic barriers that prevent transitions from one cell state to another. A stable cellular state is characterized by a given gene expression pattern. The perturbation of this pattern can induce cells to overcome these barriers by changing their steady state from one attractor to another in the epigenetic space [Ding and Wang, 2011]. This transition has the consequence of changing the cell phenotype. As an example, we can cite the positive regulation of transcription factors responsible for the regulation of a gene expression pattern. The scheme of Fig. 2 may represent both dedifferentiation and transdifferentiation processes. In general, we can think at epigenetic landscape as an

energy configuration, where the cellular state is defined by the underlying transcriptional and epigenetic regulation [del Sol and Buckley , 2014].

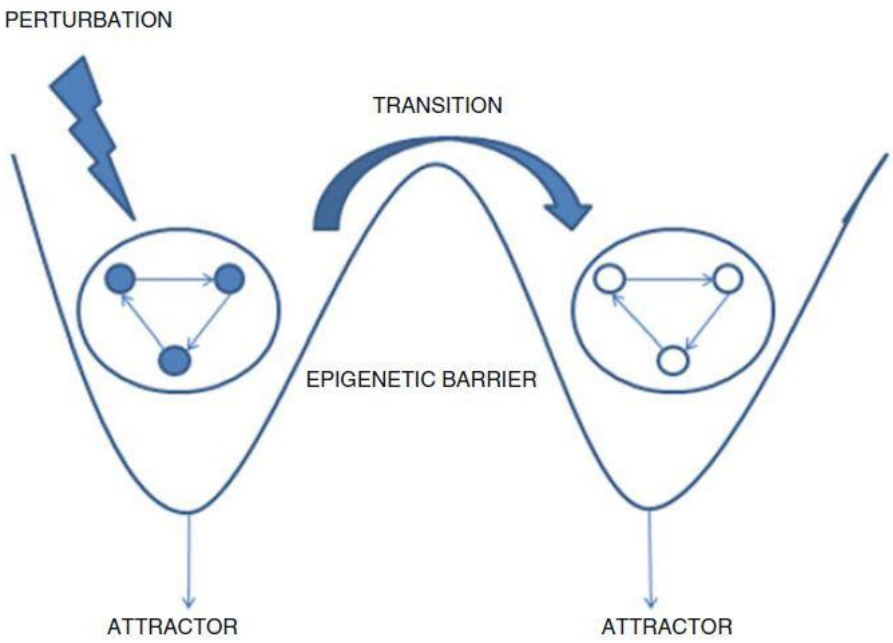

Fig. 2.2 Schematic representation of the cellular transition from one attractor to another by overcoming an epigenetic barrier between two cell states as result of a specific perturbation.

2.1.3 Applications

Basically there are four main areas where cellular reprogramming are or could be applied in the biomedical research [Mall and Wernig , 2017]:

(a) Disease modeling
(b) Drug discovery
(c) Precision medicine
(d) Regenerative medicine

With disease modeling (a), we may think about transforming a cell pathology into another desired cell condition, such as healthy, less aggressive phenotypes or even cell death.

The benefit of this approach is to work with a human-specific representation that may not be available through cells coming from animal models. As an example, astrocytes dysfunction is related to several neurological and degenerative diseases, and their cellular reprogramming provides potential for the investigation of developmental and evolutionary features of the human brain. Exploring such potentialities, Dezonne et al. [Dezonne et al., 2017] successfully generated astrocytes from human cerebral organoids.

Concerning drug discovery (b), new drug targets can be inferred from a model representation and tested for cell reprogramming in vitro and in vivo before they reach clinical trials. For example, induced PSCs can be reprogramed into insulinsecreting pancreatic β cells, and their determinant genes could serve as targets for drug development. Also, induced PSCs from diabetes patients are being used to perform drug screening for new therapies against diabetes mellitus (DM) [Kawser Hossain et al., 2016].

Precision medicine (c) aims to provide an individual treatment to patients and diseases. A key factor in this context is the pharmacogenomics that studies the influence of an individual's genetic characteristics in relation to its body's response to a drug. Succeeding in reprogramming a cell to a pluripotent state gives a chance to better understand the gentype-phenotype relationship at the individual level, which should allow the improvement of therapeutic efficacy [Hamazaki et al., 2017].

Regenerative medicine (d) is the process of replacing, engineering, or regenerating human cells, tissues, or organs to restore or establish normal function [Mason and Dunnill , 2008]. In therapies of cell replacement, the use of reprogrammed autologous cellscan theoretically be a solution against the risk of graft rejection, due to cellular mismatch between the host and donor. In order to implement this idea in humans, nonhuman primates were studied regarding their potential to generate PSC cells through different cellular reprogramming techniques [Hemmi JJ et al., 2017].

2.2 Reprogramming methods

By *cell state*, one means the phenotype features of a cell as determined by the expression pattern of some of its key genes. Based on this definition, it is necessary to act on the expression of key genes to change a cell's phenotype features, which is the main purpose of cellular reprogramming. Consequently, one way to achieve such purpose is to modulate the regulation of the transcription factors that are responsible for the expression of those key genes. This method will be discussed below, together with other cellular reprogramming techniques that were also used [Halley-Stott RP et al., 2013].

2.2.1 Cellular reprogramming through the overexpression of transcription factors

The discovery that it is possible to change cellular fate by overexpressing just four transcription factors [Takahashi and Yamanaka , 2006] boosted the field of cellular reprogramming. After transfection, the cell was induced to a pluripotent state very much similar to that of embryonic stem cells; this similarity concerned morphology, phenotype, and epigenetics.

The switch from a somatic cell phenotype to induced PSCs through the modulation of transcription factor expression has an efficiency lower than 1% [Takahashi K, 2014]. Once the genomic sequences of the original and reprogrammed cells are mostly identical, the reason for the low performance of cell reprogramming may be related to cell epigenetic factors, which indicates that induced PSCs have an epigenetic memory inherited from the previous cellular state [D'urso and Brickner , 2014].

Lineage reprogramming can also be obtained by cell reprogramming. As an example, Takahashi and Yamanaka [Takahashi and Yamanaka, 2006] performed random gene integration at multiple DNA sites to obtain the overexpression of Oct4, Sox2, Klf4, and c-Myc transcription factors in adult fibroblasts, which caused their return to a pluripotent state. This transformation with retroviral vectors was performed for experimental purposes since the DNA integrates randomly at multiple sites and might promote the knockdown of essential genes and entail oncogenicity. To avoid such

noxious risk, alternative transformation techniques were used, such as the combination of seven drug-like compounds that were able to generate iPSCs without the insertion of exogenous genes [Hou et al., 2013]. In addition to drug-like treatment, the repeated transfection of plasmids for transcriptional factor expression into mouse embryogenic fibroblast was also performed, but without any evidence of their genomic integration [Okita et al., 2008].

2.2.2 Somatic cell nuclear transfer

Somatic cell nuclear transfer (SCNT) is a technique in which the nucleus of a donor somatic cell is transferred to another enucleated one called *egg cell*. After insertion, the somatic cell nucleus is reprogrammed by the egg cell. With this method it is possible to obtain embryonic stem cell (ESCs) [Byrne et al., 2007] as well as to induce the differentiation of a cell phenotype into a different one [Wakayama et al., 2001].

2.2.3 Cell fusion

It is possible to combine two nuclei within a same cell by the fusion of two cells. The dominant nucleus, the larger and more active one, imposes its pattern and consequently reprograms the somatic hybrid cell according to its dominant characteristics [Yamanaka and Blau, 2010]. It is worth noting here that the cell fusing technique is not always efficient in achieving the desired result and the reprogramming is often incomplete.

2.3 Modeling cellular reprogramming

Reprogramming is obtained by resetting the regulation of gene expression in somatic cells, which depends on the knowledge of the key genes and proteins that may serve as target to induce this process, and the interactions between them.

The intracellular environment is continually subjected to stimuli from extracellular environment, such as nutrient availability, mechanical injury, cell competition, cooperation, etc. This type of stimulation affects the intracellular environment by changing the gene expression pattern in response to each stimulus. In this context, transcription factors are activated by external signals through transduction and promote the expression of specific genes and their respective pathways to set up a cellular response. This regulation process can be extended and include the induction of specific cell phenotypes.

Therefore, modeling the interaction between proteins in a living system and the transcription factors that regulate their expression is essential to carry out cellular reprogramming. As an approach to model such cellular systems, we may consider genes as variables and their activation state as “on” or “off.”With these observations in mind, we may address some mathematical methodologies to represent the relationship between these state variables.

2.3.1 A Data-Oriented approach

The development of new high-throughput technologies along with the growing amount of available data did promote computational frameworks based on protein interaction networks [Rackham et al., 2016] integrated to different databases, such as (i) FANTOM consortium [Forrest et al., 2014], which contains data on promoter characterization; (ii) STRING[Franceschini et al., 2012], which provides protein-protein interactions (PPI); and (iii) MARA (Motif Activity Response Analysis) [Suzuki et al., 2009], which provides interactions between proteins and DNA, to predict the reprogramming factors necessary to induce cell conversion.

In this context, Mogrify [Rackham et al., 2016] is a predictive system that integrates gene expression data and regulatory network information. It searches for differentially expressed transcriptional factors that regulate most of the differentially espresse genes between two cell types. This methodology has been validated in vitro by inducing the transdifferentiations of dermal fibroblasts into keratinocytes and of keratinocytes into microvascular endothelial cells.

Basically, one may model a biological system through three different strategies (Fig. 3).

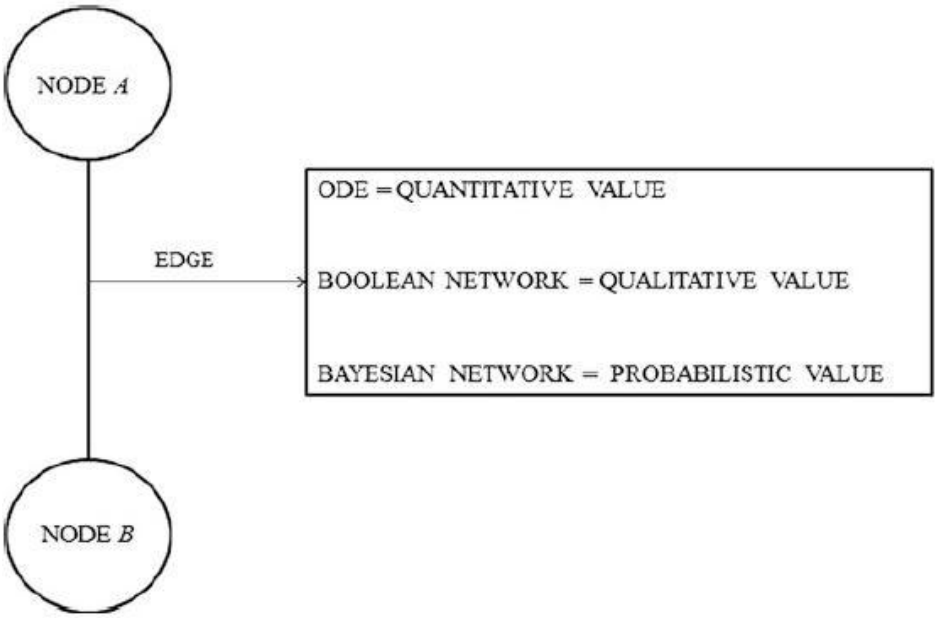

Fig. 2.3 Schematic representation showing the interpretation of an edge between two nodes by three different modeling methods: (i) ODE gives a quantitative description of the state of the connection by differential equation modeling, (ii) Boolean network gives a qualitative interpretation in terms of a connection being activated or not, and (iii) Bayesian network gives a probabilistic assessment of the connection state.

2.3.2 Ordinary differential equation

In the context of a gene regulatory network, ordinary differential equations (ODEs) are used to describe the existing quantitative relationship between variables, i.e., nodes [Cao et al., 2012]. Theoretically, the use of ODE can provide a very accurate description of the existing interactions between system elements. In practice, the use of this technique, especially in complex networks, is difficult due to the high number of data and parameters involved in the process. The differential equation (formula 2.1) for each variable in the network is

$$\frac{dx_i}{dt} = f_i(x_{i1}, x_{i2}, \dots, x_{il}) \qquad (2.1)$$

where the right side of the equation represent all variable function linked to the gene $\boldsymbol{x_i}$, and the left side is the variation in the gene $\boldsymbol{x_i}$ expression. ODE can be used to model cellular reprogramming by determining the rate of change of a given substance concentration within the cell that determines a precise cellular state in response to some kind of cellular perturbation. For example, Mitra [Mitra et al., 2014] used ordinary differential equations to prove that time delays from chemical reactions are of crucial importance to understand cell differentiation and that it allows the introduction of a new system regime between two admissible steady states with sustained oscillations due to feedback loops in gene regulation circuits.

2.3.3 Bayesian network

Bayesian network is an example of network analysis that takes into consideration the random behavior inherent to biological networks. Bayesian networks are acyclic graphs $\boldsymbol{G} = (\boldsymbol{X}, \boldsymbol{E})$, where $\boldsymbol{X}$ represents the network nodes and $\boldsymbol{E}$ the directed edges that represent the probabilistic relationship dependence between nodes. The relationship between the network's nodes is regulated by a conditional probability distribution (formula 2.2):

$$\boldsymbol{P}(\boldsymbol{x_i}|\boldsymbol{Pa}(\boldsymbol{x_i})) \tag{2.2}$$

where $\boldsymbol{Pa}(\boldsymbol{x_i})$ represent the antecessor nodes of the node $\boldsymbol{x_i}$. A Bayesian network is a representation of a join probability distribution (formula 2.3):

$$\boldsymbol{P}(\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}) = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{P}(\boldsymbol{x_i}|\boldsymbol{Pa}(\boldsymbol{x_i})) \tag{2.3}$$

It allows an intuitive visualization of the network conditional structural dependences between variables [Friedman et al., 2000].

Bayesian networks that model sequences of variables varying over time are called *dinamica Bayesian networks* (DBNs). As proposed above, one may consider each protein in the network as being active or inactive. In this context, DBN allows the inference of the likelihood of each network node state, which is necessary to calculate

the probability of each cell state [Chang et al., 2011] (an essential feature of cellular reprogramming). As an example, Chang et al. [Chang et al., 2011] established a cellstate landscape that allowed the search for optimal reprogramming combinations in human embryogenic stem cell (hESC) through the use of DBN.

2.3.4 Boolean network

An alternative to differential equations and Bayesian network to describe variables' relationships in a gene regulatory network is the use of Boolean network. It is a qualitative dynamical model, describing a system change over time, which each network node being either "on" or "off." Its representation of the system is easier to derive that the one based on ordinary differential equations, since it does not require the inference of kinetic parameters and, consequently, it can process gene networks with a higher number of nodes.

A Boolean network is a directed graph $\boldsymbol{G(X,E)}$ where $\boldsymbol{X}$ represent the nodes ofthe network and $\boldsymbol{E}$ are the edges between them. The vector of formula (2.4)

$$\boldsymbol{S(t) = (x_1(t), x_2(t), \dots, x_n(t))} \tag{2.4}$$

describes the state of the network at any given time. The Boolean value, 1 or 0, of a node represents the state "on" or "off" of the gene considered, i.e., active or inactive, respectively.

The Boolean model is suitable to represent the evolution of biological systems over time and is relatively simple to implement and interpret. The greatest limitation of this type of network is that the state, 0 or 1, of a node is just an approximation of the reality. The state updating of all nodes across the entire system can be synchronous, asynchronous, or probabilistic depending on the modeling purpose and parameter's availability [Xiao, 2009].

2.4 Cellular reprogramming using a Boolean network

To address the problem of cellular reprogramming using the Boolean network in practice, one may use a modeling strategy of *gene regulatory network* (GRN) that warrants a relative simplicity in finding attractors. It should be noticed, however, that detailed information on the interactions within the elements of the network is not taken into account by this approach, since kinetic parameters or affinity terms may take different values according to different components of the network.

As seen above, gene interactions can be modeled based on the knowledge of the relationships between the genes of a set that should be modulated, activated, or inactivated, to achieve cellular reprogramming. Therefore, it is crucial to identify specific transcription factors that regulate these genes in order to enable a cell to perform a transition between its actual state and the wanted state.

Different cell types are defined as stable states, and a stable steady state is called an attractor. An attractor is characterized by a gene expression pattern that is specific of that attractor and whose perturbation can induce a transition from a stable cellular state to another [Crespo et al., 2013]. It was shown that the number of genes to be modulated to reach attractor reprograming is relatively low, compared to the high number of genes differently expressed between two different cellular states [Lukk et al., 2010].

Considering that the complexity of a gene regulatory network increases together with its number of nodes and that a phenotypic transition requires a low number of genes to be perturbed [Crespo and del Sol, 2013],different strategies are being used to reduce the number of network nodes to be analyzed. An iterative network pruning can be used to contextualize the network to the biological condition under which the expression data were obtained [Crespo et al., 2013]. Pruning algorithms compare lists of genes and interactions from literature-based network with lists of genes differentially expressed from a bench experiment in two cellular phenotypes and then search for compatibilità between both data sets. This comparison produces a score for each sample of pruned network in order to identify the genes to be perturbed according to the data pair that best matches the cell steady state regarded as a phenotype.

The topological relationship between the elements of a specific attractor in a network can be used to construct a protocol of cell reprogramming [Crespo and del Sol, 2013]. Based on data of topological configuration, it is possible to establish a

hierarchical organization of *strongly connected components* (SCC), identify their respective *differentially expressed positive circuits* (DEPCs), and identify determinant genes able of promoting the transition from one stable cellular state to another.

The choice of genes to be perturbed can also be done based on dynamic simulation [Crespo et al., 2013] through the combination of transcriptomic profiling and analyses of network stability in order to find the minimum number of DEPCs that needs to be perturbed to complete cellular transition.

2.5 Application of cellular reprogramming to disease control

All human diseases are intrinsic multifactorial and characterized by dysregulated processes in gene regulatory networks. The knowledge of GRN is important to understand how a molecular network robustness may lead malignant cells to overcome the inactivation of single protein targets by therapeutic treatment through alternative pathways or network propagation until a system accustoms to a new equilibrium [Cornelius et al., 2013]. Thus, network pharmacology and cellular reprogramming are promising methods for the identification of protein combinations with potential to disarticulate a key subnetwork that correlates with a disease and achieve an efficient therapeutic result [Crespo and del Sol, 2013; Zickenrott et al., 2016].

A very common problem is the bias in the modeling representation induced by reference to well-known pathways already described for the disease and the use of generic models that do not consider the specific features of the cell or tissue under consideration. The methods described in the previous section overcome this problem through the integration of gene expression data and regulatory networks, which allows the reconstruction of a network specific to the case under consideration. This specific network is more accurate, indicates specific aspects of the diseased cell or tissue, and may indicate genes related to dysregulated pathways responsible for the disease development [Rackham et al., 2016; Crespo et al., 2013; Crespo and del Sol, 2013].

The use of gene expression data from both ill and healthy cells is also important to identify the differentially expressed genes and target the ones preferentially expressed in ill cells in order to minimize the negative side effects of target inactivation to healthy cells.

The Mogrify methodology [Rackham et al., 2016] considers all these features. However, it potentially may cause two types of negative effects if applied to patients in the context of a therapeutic treatment. First, with this methodology, one searches for differentially expressed transcription factors responsible for the regulation of genes related to the establishment of the disease phenotype. The problem is that transcription factors might be responsible for the regulation of hundreds of genes, and probably they

are not all significantly more expressed in ill than in healthy cells. The perturbation of hundreds of genes, even if they are mostly differentially expressed in disease cells, may affect genes that are essential to cell maintenance and cause serious side effects. Second, this methodology requires the induction of gene expression through cell transfection. As already discussed above, the insertion of a plasmid into DNA occurs randomly and might knockdown some key genes, which increases the risk of oncogenicity. The most common approach applied in patients is the inhibition of a protein target with drugs. Even new innovative alternative patient therapies based on biopharmaceuticals as RNA interference, aptamer, peptides, or antibodies also target proteins with the aim to inactivate their function [Tabernero et al., 2013].

These limitations need to be considered when applying cellular reprogramming strategies in a disease context because they may exclude a number of possible alternative solutions. Once attractors for cell reprogramming have been considered, it is important to emphasize that focusing on the full reprogramming of a cell in order to reach a given steady state is not necessary. All stable attractors have a basin of attraction, in which trajectories spontaneously converge to the steady-state attractor [Zickenrott et al., 2016]. The concept of basin of attraction should simplify the application of cellular reprogramming in diseases, since it reduces the number of requie perturbations needed to achieve the desired stable state.

The perturbation capable of overcoming an epigenetic barrier and bringing a cell from a disease attractor to another desired one considered to match a healthy, or at least a less aggressive, condition for the patient needs to be carried out in a subspace where therapeutic options overlap with the basin of attraction.

As examples, we now propose putative applications of cellular reprogramming in two different diseases, cancer (cell disease) and malaria (infection disease).

Cancer cells accumulate malignant mutations during their development and, as result, present a different network topology if compared to healthy cells [Jonsson and Bates, 2006].

Due to mutations accumulation and its consequences on genome dysregulation, it would be impossible to control a cell in order to bring it back from its malignant attractor toward its healthy one. However, the key genes involved in the malignant attractor can be analyzed at the light of malignant features, such as continuous proliferation and escape from apoptosis or cell death. In addition, both malignant and healthy conditions can be analyzed in terms of differences according to their attractor phenotypes. This would allow the identification of key genes able to reprogram dysregulated cellular processes and achieve proliferation control and/or the induction of malignant cells to apoptosis.

The vaccines used against malaria uses live attenuated *salivary gland sporozoites* (SPZ) [Phillips et al., 2017], and cannot be produced in large scale due to hurdles associated with SPZ obtainment. It is known that SPZ development occurs following three main stages according to the mosquito organs that are infected: midgut, hemolymph, and/or salivary gland. Therefore, if considering the salivary gland tissue, the cellular reprogramming analysis should allow the identification of key genes related to this tissue by comparison to the others two stages. The understanding of salivary gland SPZ genesis and maturation is crucial to develop a culture system in laboratory and produce SPZs in vitro for large-scale vaccine production. Many advances were already made toward cell reprogramming, and it is effective for a number of purposes. However, much still need to be done in regard to diseases and patient treatment. A clear example is that, unfortunately, an efficient general method for identifying basins of attraction is still lacking [Cornelius et al., 2013].

2.6 Chapter conclusion

The concept of cell reprogramming has evolved a lot during the last decade. The development of high-throughput technologies has also promoted more accurate applications of cell reprogramming through its integration with gene expression data. Currently, there is a great perspective of its application in multiple biomedical areas, such as drug screening and regenerative medicine. Nevertheless, there is still much to do in order to understand and predict the behavior of complex systems such as the biological ones.

CHAPTER 3 [Sgariglia et al., 2021]

Data-Driven Modeling of Breast Cancer Using Boolean Network

Cancer is a genomic disease involving various intertwined pathways with complex cross-communication links. Conceptually, this complex interconnected system forms a network, which allows one to model the dynamic behavior of the elements that characterize it to describe the entire system's development in its various evolutionary stages of carcinogenesis. Knowing the activation or inhibition status of the genes that make up the network during its temporal evolution is necessary for the rational intervention on the critical factors for controlling the system's dynamic evolution. In this report, we proposed a methodology for building data-driven boolean networks that model breast cancer tumors. We defined the network components and topology based on gene expression data from RNA-seq of breast cancer cell lines. We used a Boolean logic formalism to describe the network dynamics. The combination of single-cell RNA-seq and interactome data enabled us to study the dynamics of malignant subnetworks of up-regulated genes. First, we used the same Boolean function construction scheme for each network node, based on canalyzing functions. Using single-cell breast cancer datasets from The Cancer Genome Atlas, we applied a binarization algorithm. The binarized version of scRNA-seq data allowed identifying attractors specific to patients and critical genes related to each breast cancer subtype. The model proposed in this report may serve as a basis for a methodology to detect critical genes involved in malignant attractor stability, whose inhibition could have potential applications in cancer theranostics.

Cancer is a multifactorial disease resulting in uncontrolled cell growth and the spread of cancer cells from the original site to other body areas. The modification of cellular homeostasis through various processes identifies and characterizes the Hallmarks of cancer[Hanahan and Weinberg, 2011], typical to all types of tumors. Cell survival, proliferation, and metastatic dissemination are driven by different cellular pathways, with many genes involved. These highly complex interconnections modify the linearity of the pathways allowing the conceptualization of a reticular structure made

up of genes, proteins and other molecules, characterizing cancer as a network disease. This structure defines a robust state of endogenous networks [Yuan et al., 2017; Su et al., 2017], which dynamically describes the cellular network as composed of oncogenic factors, tumor suppressors, and other acting agents, which modulate the main molecular functions.

Breast cancer, which is the type of cancer addressed in this report, is the leading cause of death due to cancer of the world's female population. It accounts for 23% of all cancer deaths of postmenopausal women[Akram et al., 2017]. Current therapies used to combat this disease frequently produce harmful side effects. In patients undergoing chemotherapy, 38 symptoms were identified, classified into 5 clusters characterizing the symptomatology [Chan et al., 2017].Therefore new therapeutic strategies aiming to decrease the undesirable effects produced by current treatment approaches, together with improved therapeutic efficacy, are needed. Personalized medicine seems to increasingly gain importance in patient care. The purpose of this therapeutic approach is to adapt the treatment to the unique characteristics of the individual patient's disease [Sabatier et al., 2014], which are based not only on the site of the tumor but also on genetic characteristics such as mutations and gene expression profiles. There are different methodologies to model gene regulatory networks. The ordinary differential equations (ODE) and stochastic differential equations (SDE) are quantitative approaches that allow an instrumental and detailed description of the system's dynamic functioning when the exact mechanisms and kinetic parameters are well known. Given the noise level of cellular processes, the precise determination of ODE and SDE parameters is challenging [(Nasti, 2020]. A qualitative approach would help avoid ODE and SDE limitations while providing useful information on the system under study. Boolean network Modeling is an example of this methodology [Somogyi and Sniegoski, 1996]. It is composed of Boolean variables representing the nodes (which corresponds to vertices in a graph) making up the network, whose values are periodically updated synchronously (i.e., all nodes are updated simultaneously) or asynchronously. These updated values represent the activation/inhibition status of the genes that make up the studied system [Barbuti et al., 2020]. The dynamic simulation of the network, guided by the Boolean functions that regulate the relations between the various vertices, reaches a set of final stable states, which can be cyclic or not. These repetitive states compose network attractors. The formulation of the concept of "Epigenetic Landscape" by Waddington [Waddington, 1957] offers the opportunity for modeling cellular

functioning through attractor theory [Huang et al., 2009]. The Boolean paradigm allows the processing and analysis of vast gene regulatory networks, resulting in an improved capacity to model the complexity of cancer since no parameters are required. This report analyzed a gene regulatory network specifically adapted to breast cancer through a qualitative dynamic analysis using Boolean network modeling. From the choice of the network vertices (genes), the network topology, and the definition of the functional relationships at each vertex, one may found the attractors within the system through the assignment of binary gene expression values. We adopted a step-by-step network pruning approach to identify the genes being key determinants of specific basins of attraction with therapeutic relevance. Generally, when looking for attractors in a Boolean network, one considers every possible vertex configuration [Barbuti et al., 2020]. On the other hand, in our approach, we identify biologically relevant attractors through trajectories. The initial point of these trajectories is the binarization of the cellular data of specific gene expression of a given tumor belonging to a given individual, enabling different and specific outcomes for different patients.

The network's basins of attraction that emerged from the single-cell RNA-seq (scRNA-seq) data [Saliba et al., 2014] represent this research's culmination. The essential genes that contribute to the stability of a basin of attraction can be considered potential therapeutic targets since they may modify the epigenetic landscape in which they are involved. The results described in this work show a difference between the various basins of attraction related to cancer and control cells, therefore confirming the relevance of the data-driven customization procedure based on patients' transcriptional data. This work also describes methods for identifying potential therapeutic targets specific to each patient using boolean network modeling.

3.1 Materials and methods

3.1.1 Overhead description of the method

The main steps of the method adopted in this report are as follows:

1) Selection of breast cancer-related genes and subsequent gene regulatory network construction based on this gene set.

2) Adoption of the Boolean formalism for the dynamic modeling of the system and assignment of a specific type of Boolean function (i.e. canalyzing functions) to all nodes of the network.

3) Selection of single-cell RNA-seq (scRNA-seq) data relating to breast cancer, assigning expression values to the gene regulatory network's corresponding nodes

4) Binarization, for the set of cells in the dataset of step 3, of the expression values assigned to each gene.

5) Search for attractors in each cell provided in the dataset. We use the binarized values assigned to the network genes for each cell (step 4) as the initial value for a trajectory simulation. The set of states that compose the final cycle of the trajectory corresponds to the cell's attractor.

This procedure allowed us to highlight attractors and related genes constantly expressed in the dataset of different patients.

In subsection 3.1.2, we describe the procedure by which we selected the constituent elements of the gene regulatory network used in this report . The description of Boolean formalism used to model the network dynamic is in subsection 3.1.3.In subsection 3.1.4 we describe the scRNA-seq data used to quantify the network genes. In subsection 3.1.5, we illustrate the method by which the scRNA-seq values assigned to the constituent elements of the gene network have been binarized, and descrive the tool used in this report to obtain this result. The last subsection 3.1.6 describes the essential characteristics of the network's attractors and the procedure, through an appropriate software tool, of its identification by simulating trajectory dynamics.

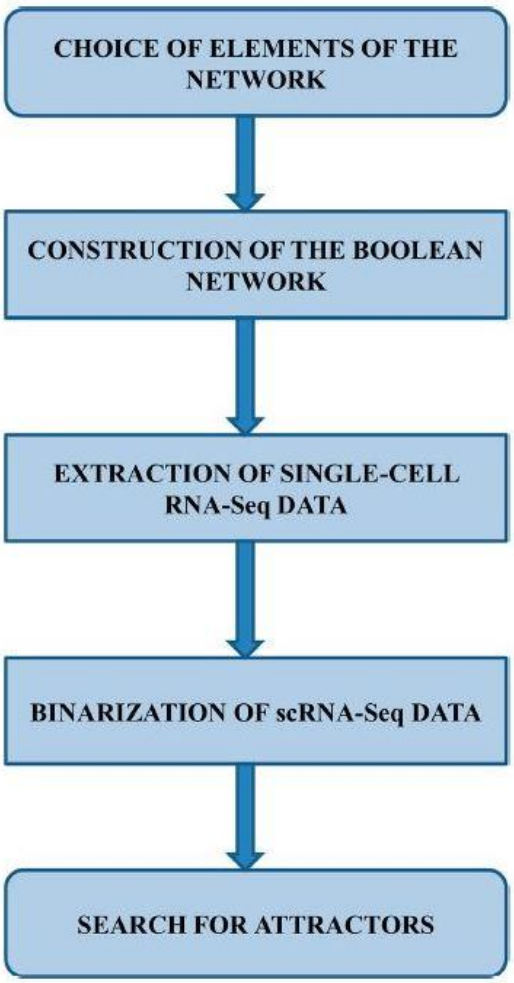

Fig. 3.1: Workflow illustrating the various stages of the method used in this work

The following subsections detail the steps shown in Figure 3.1.

3.1.2 Choice of the elements of the gene regulatory network

Hallmarks of cancer represent groups of acquired biological features that are critical for its development [Hanahan and Weinberg, 2011]. We considered two of these hallmarks, UNLIMITED REPLICATIVE POTENTIAL, and EVASION OF CELL DEATH, as starting points for constructing a representative gene regulatory network of cancer. This modeling strategy was chosen to reduce cancer cell proliferation and promote their death. We then obtained four lists of genes from the MSigDB repository (http://www.gsea-msigdb.org/gsea/msigdb/index.jsp) based on the two hallmarks

previously considered, each list representing a specific cellular pathway. The gene lists related to Apoptosis and TP53 represent the "Evasion of cell death" hallmark [Wong, 2011], while Kras pathway (up and down-regulated genes) is indicative of the "Unlimited replicative potential" hallmark [Jančík et al., 2010; Aubrey et al., 2016]. These choices are justified by the importance of those Hallmarks and pathways in Brest Cancer's formation and evolution. We retained only the genes that were significantly differentially expressed (DE) in these two lists from RNA-seq data of the MDA-MB-231 cell line, a metastatic triple-negative breast cancer subtype (TNBC), and MCF10A cell line, used as control [Carels et al., 2015].

The selected genes were analyzed considering the number of interactions (edges) of their respective proteins (vertices) in the interactome. The human interactome used in this report is from the intact-micluster.txt file (IntAct database, version updated December 2017 accessed on January 11, 2018, at ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/intact-micluster.txt). Proteins with edge numbers equal to or greater than 50 were selected as seeds to build the gene regulatory network. Those proteins are potential hubs, for which inhibition has been widely associated with regulatory network disruption [Carels et al., 2015].

We also added five genes to the analysis (HSP90AB1, YWHAB, VIM, CSNK2B, and TK1) [Carels et al., 2015] whose knockdown was shown to inihibit the cell growth and promote the cell death of MDA-MB-231 in vitro [Tilli et al., 2016].

We used the human interactome to define the connections between the proteins coded by the selected genes (network vertices). In case of a lack of a direct relationship between two vertices, we looked for possible intermediary vertices (up to three). We excluded intermediary vertices absent in the gene expression data or with low expression values in MDA-MB-231.

We enriched this network with transcriptional factors that regulate the selected vertices, i.e., differentially expressed hubs and intermediary proteins. We performed this analysis with the online tool TRRUST [Han et al., 2015].

The human interactome from IntAct defines the direction of the interactions (node A regulates node B), but not their function (activation or inhibition). For the definition of interaction functions, we used the Metacore algorithm [Ekins et al., 2007].

3.1.3 Construction of the Boolean network model

We constructed a directed graph model based on Boolean logic from scRNA-seq data. The vertices represent the constituent elements of the dynamic cellular model, and their connections are for the functional regulations acting between them [Emmert-Streib et al., 2014]. Boolean network modeling is among the simplest methods for dynamic modeling [Thomas, 1973], but at the same time with characteristics of reliability in providing insights into the dynamics of a system [Herrmann et al., 2012;Siegle et al., 2018]. We have translated the gene expression status of a gene into the value of a Boolean variable ($\boldsymbol{B}$), which can be True or False (1 or 0) based on RNA-seq data. Thus, for the $\boldsymbol{n}$ vertices of our network, we have:

$$\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \ldots, \boldsymbol{x_n}\}, \boldsymbol{x_i} \in \boldsymbol{B} \tag{3.1}$$

This formalization finds its justification considering that one can describe many biological processes, such as concentration levels, through the Hill-Function. For most of the Hill function coefficient values, the resulting curve is a sigmoidal curve, which can be approximated by a dichotomous step-function [Schwab et al., 2020].

The representation of this network's state in the discrete-state flow of time is a vector whose components are the network's vertices.

$$\vec{\boldsymbol{x}} = (\boldsymbol{x_1}(\mathbf{t}), \ldots\ldots, \boldsymbol{x_n}(\mathbf{t})) \tag{3.2}$$

and the passage from a certain point of the state space of the system to another is due to the regulatory action of the corresponding Boolean functions:

$$\boldsymbol{x_i}(\boldsymbol{t}+\mathbf{1}) = \boldsymbol{f_i}\big(\vec{\boldsymbol{x}}(\boldsymbol{t})\big), \boldsymbol{f_i}: \boldsymbol{B^n} \rightarrow \boldsymbol{B} \tag{3.3}$$

for $\boldsymbol{n}$ nodes of the network.

We decided to adopt a synchronous update mode, where all genes update their values simultaneously at consecutive time points:

$$\boldsymbol{T}\big(\vec{\boldsymbol{x}}_i^t, \vec{\boldsymbol{x}}_i^{t+1}\big) = \boldsymbol{T_1}(\boldsymbol{x}_1^t, \boldsymbol{x}_1^{t+1}) \wedge \ldots \wedge \boldsymbol{T_n}(\boldsymbol{x}_n^t, \boldsymbol{x}_n^{t+1}) \tag{3.4}$$

In the equation (3.4) where $\boldsymbol{T}\left(\vec{x}_i^t, \vec{x}_i^{t+1}\right)$ represents the transition function of the state of the network, all the genes in the network simultaneously make the transition from the state $\vec{x}_i^t$ to the next state $\vec{x}_i^{t+1}$ in transitions $\boldsymbol{T_1}, \boldsymbol{T_2}, \ldots, \boldsymbol{T_n}$ occorring in the system. Some reports state that asynchronous updating seems better to model biological systems [Schwab et al., 2020]. Nevertheless, synchronous dynamic evolution is computationally more efficient for the type of network used in this report and seems to represent the network's dynamic behavior in a very similar way [Schwab et al., 2020].

Identifying the rules of interaction among the different entities of the network is usually one of the most challenging tasks in studying gene regulatory network systems. Our choice was oriented towards the nested canalyzing functions [Hinkelmann and Jarrah, 2012], where multiple variables act simultaneously on the function, determining a mechanism of domination of one or a group of variables concerning the others based on their Boolean state. For example, in the expression **(A ∧ B) ∨ (C ∧ D)**, if **A ∧ B = 1**, the first two variables dominate and determine the expression value. If **(A ∧ B) = 0**, the expression value is defined by **(C ∧ D)**. Furthermore, it has been shown that nested canalizing functions are a good representation of biological regulations [Nikolajewa et al., 2007; Harris et al., 2002].

3.1.4 Single-cell RNA-seq data

The scRNA-seq data were obtained from the NCBI Gene Expression Omnibus database (accession number GSE 75688, accessed in March 2020). These data refer to the genomic expression profile of 11 patients with 549 cells analyzed. Most of those cells were malignants, while others were not. A large part of the latter were immune T-cells, immune B-cells, and myeloid immune cells. The cancer cells analyzed represented the four subtypes of breast cancer: luminal A, luminal B, HER2, and TNBC [Chung et al., 2017]. We used single-cell data for the analyzed network's corresponding genes, excluding data related to pooled samples (bulk RNA-seq). **(Supplementary material 1).** The four subtypes of breast cancer were present among the samples of the 11 patients: BC01_X and BC02_X for luminal A, BC03_X for luminal B, BC04_X,

BC05_X and BC06_X for HER2, BC07_X, BC08_X, BC09_X, BC10_X and BC11_X for TNBC. For BC03_X and BC07_X patients, there were metastatic lymph datasets corresponding to BC03lN and BC07LN. For the patient BC09_X, there was another single-cell RNA-seq (BC09Re_X). Note that patient BC05 is the only patient who received prior treatment (neoadjuvant chemotherapy and Herceptin).

As specified in the above description, the different types of breast cancer encountered in this report have already been identified in the dataset.

It is worth pointing out that for each patient, the model is associated with a specific group of cells. This approach can be conceptually made equivalent to the one defined as multi-cell pathway, and that the relatively high number of available cells analyzed allowed a correct use of the R "Binarize" application, used in this report for the extraction of the Boolean value.

3.1.5 Binarization of scRNA-seq data

Once the genes making up the network were found and its topology defined, and finally assigned the corresponding scRNA-seq values to each element of the network, the next operation necessary for the Boolean network modeling of the system was to binarize the gene expression values assigned to each single node, such as $\boldsymbol{f}: \mathbb{R} \longrightarrow \mathbb{B}$ using

$$\boldsymbol{f}(\boldsymbol{u}) = \begin{cases} \boldsymbol{0} & \boldsymbol{u} \leq \boldsymbol{t} \\ \boldsymbol{1} & \boldsymbol{u} \geq \boldsymbol{t} \end{cases} \tag{3.5}$$

where $\boldsymbol{t}$ is the separation threshold. This result was achieved through the use of the BASC-B algorithm (Binarization Across multiple SCales) [Hopfensitz et al., 2012]. The BASC algorithm considers as input values a sorted vector in ascending order $(\boldsymbol{u_1}, \dots, \boldsymbol{u_N}) \in \mathbb{R}$, and based on it, BASC defines a discrete, monotonically increasing step function ***f(x)*** with *N* steps and *N - 1* discontinuities:

$$\boldsymbol{f}(\boldsymbol{x}) = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{N}} \boldsymbol{u_i}\, \boldsymbol{I_{A_i}}(\boldsymbol{x}) \tag{3.6}$$

with $\boldsymbol{i} \in \{\boldsymbol{1}, \dots, \boldsymbol{N}\}$. Defining $\boldsymbol{d_i} = \boldsymbol{N} - \boldsymbol{1}$ as discontinuities, we have $\boldsymbol{A_i}$ as intervals defined as follow

$$A_i = \begin{cases} (0, d_i], & if\ i = 1 \\ (d_{i-1}, N], & if\ i = N \\ (d_{i-1}, d_i], & otherwise \end{cases} \quad (3.7)$$

and I_A as

$$I_{A_i}(x) = \begin{cases} 1, & if\ x \in A \\ 0, & otherwise \end{cases} \quad (3.8)$$

Once the step function ***f(x)*** is obtained using the output vector ordered in increasing order, the algorithm calculates additional step functions that approximate this function with a smaller number of discontinuities. The algorithm then finds the strongest discontinuity in each step function and estimates the strongest discontinuities' location and variation. This algorithm was implemented through the R Software package BiTrinA [Müssel et al., 2016].

3.1.6 Search for attractors

After defining the binarized RNA-seq values on each node of the network and establishing the rules that determine its dynamics, we sought the network's stable equilibrium state, i.e., the attractors, which can be either singleton (composed of a single state) or cyclic (composed of multiple states) [Huang et al., 2009]. The hypothesis under which one may consider the malignant state as a particular type of attractor [Huang et al., 2009; Creixell et al., 2012; Yu and Wang, 2016; Poret and Guziolowski, 2018] has oriented our investigations towards the localization and characterization of attractors in Boolean networks. Furthermore, basins of attraction include all the system states that evolve into a given attractor. They conceptually represent the epigenetic barriers that delimit the basin of attraction[Conforte et al., 2020]. We obtained the corresponding attractors matching a given gene network for each scRNA-seq dataset of the eleven patients with breast cancer [Chung et al., 2017]. Attractor analysis allowed us to highlight the key genes in each basin of attraction and how their inhibition could determine a change in cell fate by using the python Open Source software application "BooleanNet" [Albert et al., 2008].

We performed the following procedure to search for attractors from the available data:

- We used BooleanNet [Albert et al., 2008] to assess the logic functions assigned to each gene of our regulatory network and search for Boolean attractors.

- The Boolean values of the 103 genes making up the network were obtained by the binarization of scRNA-seq relative to each patient sample. This setting was the initial state of a trajectory that eventually evolved to a cyclic attractor.

- Considering that all the attractors obtained were cyclical for each cell analyzed, we assessed the behavior of every single gene in the network by noting whether they varied in their boolean value during the attractor cycle or if they kept a fixed Boolean value for the entire attractor cycle. In the first case, we indicated genes in each particular cell with an "X," in the second case with its Boolean value True or False.

- By grouping all cells according to their batch samples (BCXX_X) and their carcinogenic features for each patient, we selected only the genes that did not show variations in boolean values in any of the attractors for all cells, i.e., we kept their Boolean value True or False in most states of the attractor cycle, for at least 95% of the number of cells making up the group under analysis.

Figure 3.2 shows the pseudocode of the procedure described above

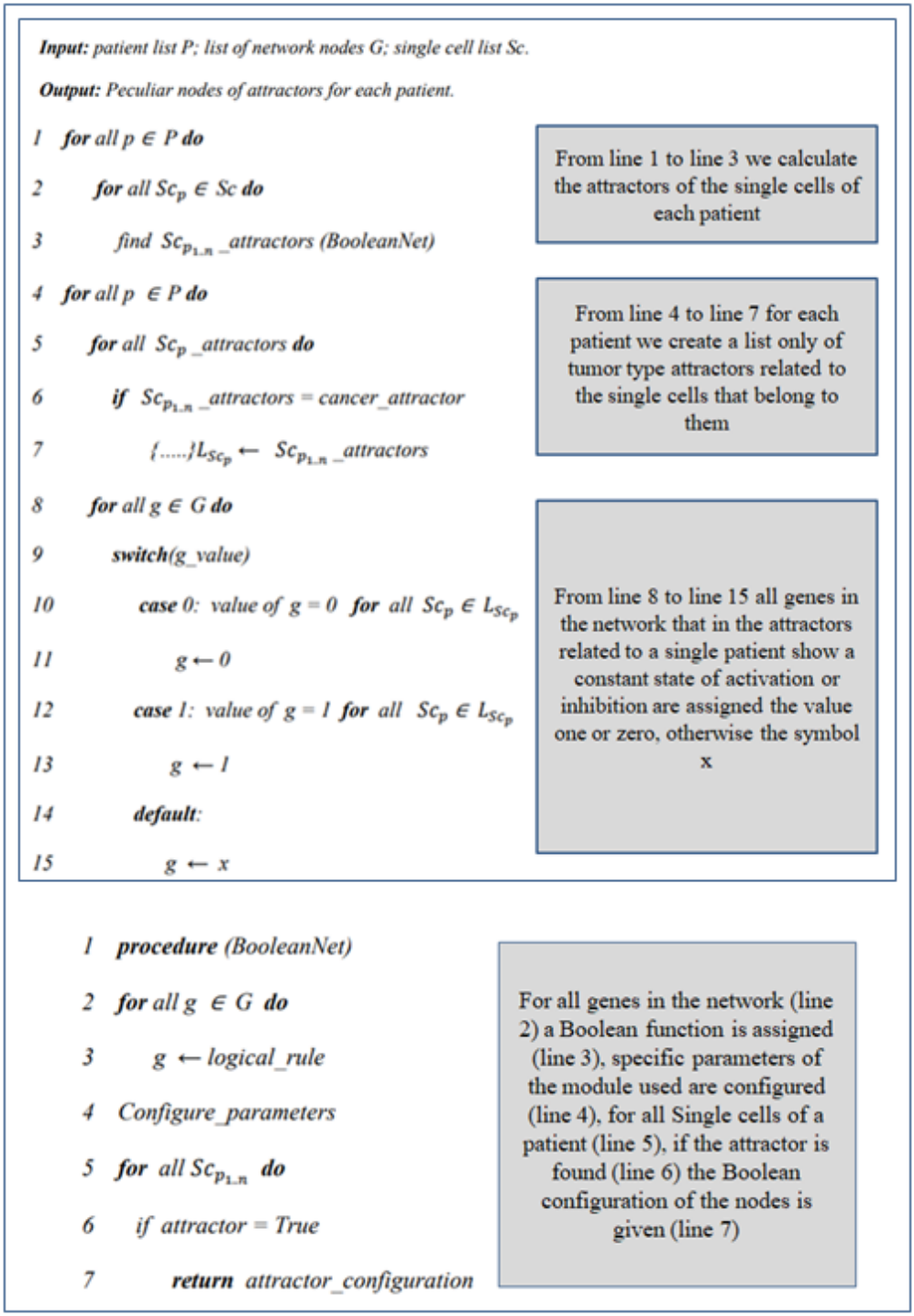

Fig. 3.2 Pseudocode of the procedure adopted for calculating the attractors of the single cells of each patient and the corresponding peculiar vertices.

Figure 3.3 below shows an intuitive diagram of the adopted procedure detailed in Figure 3.2

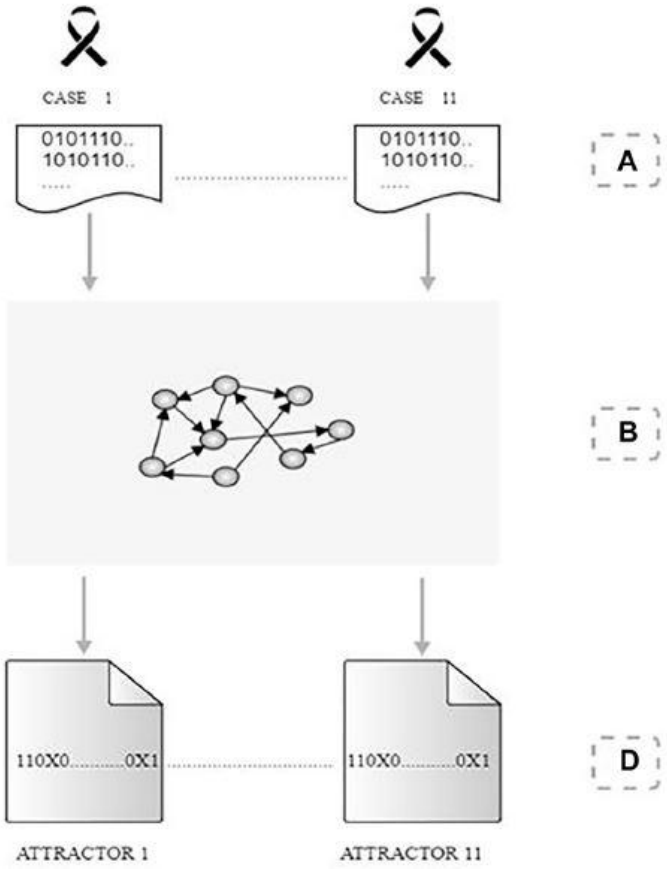

Fig 3.3: Procedure for identifying attractors. (A) we obtain a set of Boolean values for the cell samples of 11 patients considering a regulatory network of 103 genes. (B) Each patient's Boolean data was processed individually in the gene network to search for cell attractors. (C) For each detected attractor, the genes that did not change their Boolean value for the set of states that compose the cyclic attractor received the value "True" or "False" (1 or 0). The marker "X," on the other hand, highlights the genes that did not keep a single Boolean value in the set of states of the cyclic attractor.

3.2 Results

3.2.1 Breast cancer gene regulatory network

The process of choosing the elements (genes) constituting the gene regulatory network adopted in this report produced the following results.

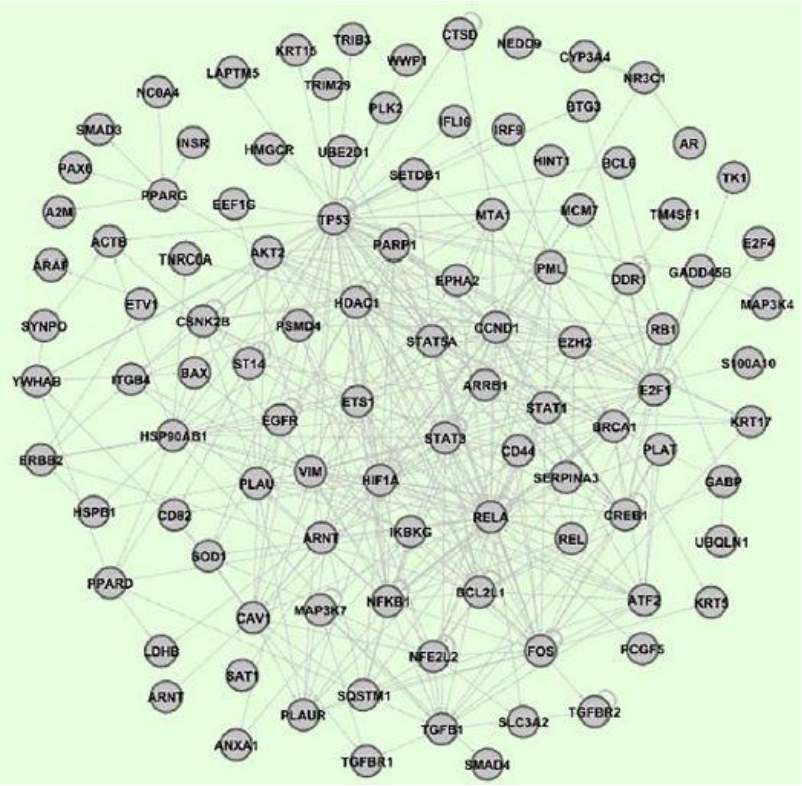

Fig. 3.4: Breast cancer gene regulatory network developed in this report. This network is composed by 103 nodes (genes).

First, we obtained 761 genes derived from the Broad Institute repository, divided into four lists related to the two cancer hallmarks used to build the network. The 761 genes were classified as follows **(Supplementary material 2)**:

- 161 related to the APOPTOSIS pathway.
- 200 related to the TP53 pathway.
- 200 related to the KRAS UP pathway.
- 200 related to the KRAS DOWN pathway.

In order to retain only differentially expressed genes, we compared the lists obtained with the RNA-seq data of the MDA-MB-231 and MCF10A cell lines, obtaining the following results:

- Because they were neither present in the gene expression data of MDA-MB-231 nor in the MCF10A one (i) 129 genes were excluded from the APOPTOSIS pathway list, leaving 32 genes;
- (ii) 191 genes were excluded from the KRAS_UP pathway list, leaving 9 genes;
- (iii) 192 genes were excluded from the KRAS_DN pathway list, leaving 8 genes;
- (iv) 164 genes were excluded from the TP53 pathway list, leaving 36 genes.

Among the genes retained, we selected only those that were differentially expressed, which resulted in a total of 51 genes **(Supplementary material 3):**

- 18 genes of the APOPTOSIS group, 9 Up and 9 Down.
- 7 genes of the KRAS_UP group, 3 Up and 4 Down.
- 4 genes of the KRAS_DN group, 0 Up and 4 Down.
- 22 genes of the TP53 group, 10 Up and 12 Down.

Based on the number of interactions in the interactome, 15 genes of the 51 were selected, from which 5 (*HSP90AB1*, *YWHAB, VIM, CSNK2B, TK1*), considered more relevant for the present research, have been added to the network [Carels et al., 2015]. As outlined above, (i) the vertice vertex connections obtained by comparison with IntAct human interactome, (ii) the inclusion of intermediate vertices, (iii) the enrichment of the network with transcriptional factors that regulate the selected vertices with the online tool TRRUST [Han et al., 2015], and (iv) the activation or inhibition of vertex inputs obtained with the Metacore algorithm [Karin, 2006]**(Supplementary material 4)**, allowed us to obtain a gene regulatory network consisting of 103 vertices (see Figure 3), and whose dynamics were regulated by nested canalyzing functions [Hinkelmann and Jarrah, 2012] **(Supplementary material 5).**

3.2.2 Binarization of scRNA-seq values

The 14 groups of scRNA-seq binarization values from the 11 patients belong to 4 types of breast cancer. They were divided and organized according to the following criterion: 26 single-cell datasets for the patient BC01_X, 56 for BC02_X, 37 and 55 for

BC03_X and BC03LN, 59 for BC04_X, 77 for BC05_X, 25 for BC06_X, 51 and 53 for BC07_X and BC07LN, 23 for BC08_X, 29 and 31 for BC09_X and BC09Re, 16 for BC10_X, 11 for BC11_X (see Figure 4). It is worth noting that the values relating to pooled single-cell present in each patient group were excluded from the count.

The gene expression values of every single cell of each patient were matched to the corresponding 103 genes making up the gene regulatory network and subsequently binarized using the BASC-B algorithm [Hopfensitz et al., 2012](**Supplementary material 6**).

3.2.3 Attractors search

For every single cell of the 14 groups representing the 11 patients of breast cancer, the 103 binarized values at each node of the gene regulatory network were processed by the previously described Boolean attractor search procedure [Albert et al., 2008]. The attractors obtained for malignant cells, stromal cells, immune B and T-cells, and myeloid cells are thus summarized as follow: (i) BC01_X: 19 malignant attractors, 2 stromal cell attractors, 5 no result; (ii) BC02_X: 49 malignant attractors, 7 no result; (iii) BC03_X: 15 malignant attractors, 7 immune B-cell attractors, 5 immune T-cell attractors, 10 no results; (iv) BC03LN_X: 6 malignant attractors, 35 immune B-cell attractors, 3 immune T-cell attractors, 11 no results; (v) BC04_X: 42 malignant attractors, 3 immune T-cell attractors, 2 immune Myeloid attractors, 12 no results; (vi) BC05_X: 74 malignant attractors, 3 no results; (vii) BC06_X: 6 malignant attractors, 2 stromal cell attractors, 6 immune B-cell attractors, 11 no results; (viii) BC07_X: 24 malignant attractors, 4 stromal cell attractors, 3 immune B-cell attractors, 4 immune T-cell attractors, 8 immune myeloid attractors, 8 no results; (ix) BC07LN_X: 24 malignant attractors, 19 immune B-cell attractors, 10 no results; (x) BC08_X: 15 malignant attractors, 6 stromal cell attractors, 2 no results; (xi) BC09_X: 2 malignant cell attractors, 1 immune B-cell attractors, 7 immune T-cell attractors, 15 immune myeloid attractors, 4 no results; (xii) BC09Re_X: 2 stromal cell attractors, 1 immune B-cell attractors, 20 immune T-cell attractors, 6 immune myeloid attractors, 2 no results;

(xiii) BC10_X: 11 malignant attractors, 2 stromal cell attractors, 2 immune myeloid attractors, 1 no results; and (xiv) BC11_X: 10 malignant attractors, 1 no results.

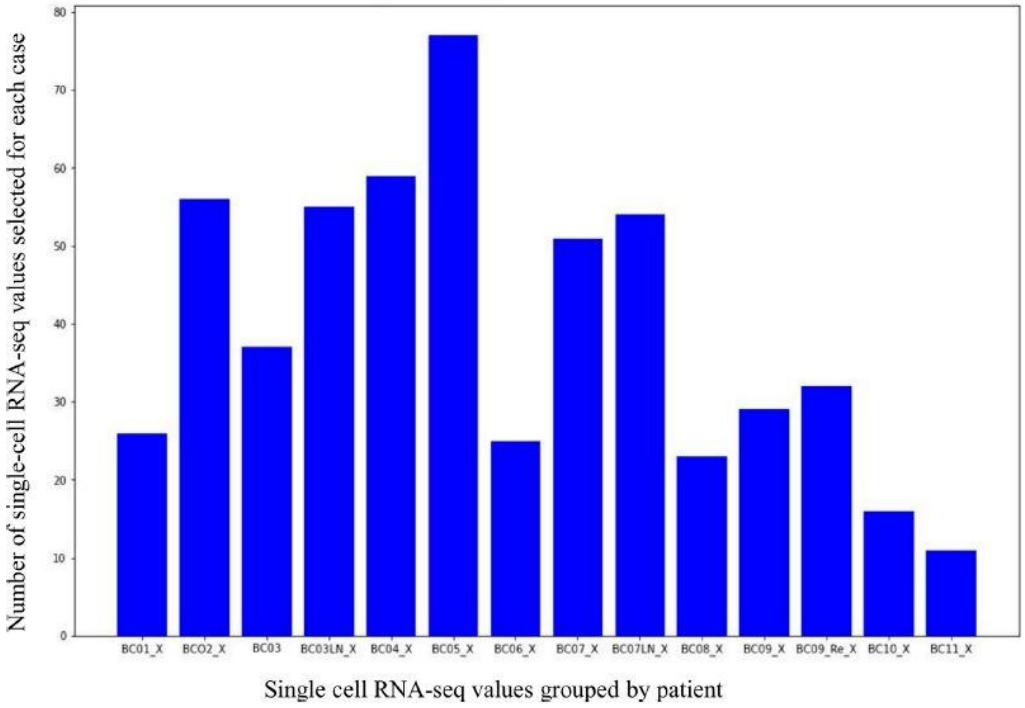

Fig 3.5: Distribution of the scRNA-seq data for each patient.

Based on these results, we decided to exclude patient data BC09_X and BC09Re due to the lack of specific tumor cell attractors. **(Supplementary material 7).**

The pie-charts in Figure 3.6 shows these results expressed as a percentage of the total Single-cell RNA-seq datasets available for each patient, highlighting the success rate in the search for specific attractors on tumor cells.

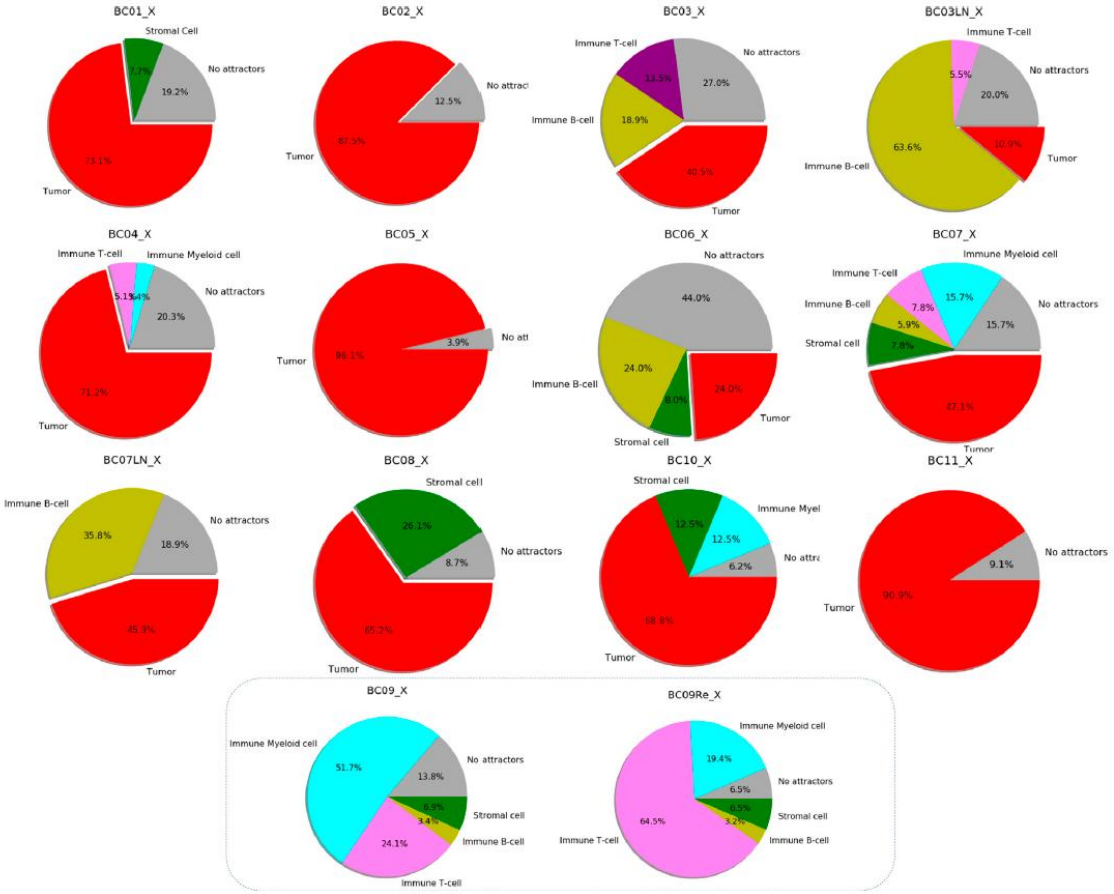

Fig 3.6: Specific categories of attractors representing a group of scRNA-seq data belonging to each patient's breast cancer sample. Every diagram shows the percentages of attractors encountered based on the total number of cells analyzed for different cell types. The different colors refer to the type of cells analyzed: red for cancer, green for stromal cells, yellow for immune B-cells, violet for immune T-cells, cyan for myeloid immune cells. Gray indicates the percentage of cells in which it was not possible to find any attractors. The last two pie charts do not indicate tumor type attractors (absence of red color).

We selected network genes based on tumor cells' attractors according to the following criterion: each patient's attractors kept their level of gene expression (or non-expression) constant for a particular gene, formalized respectively with the symbol "True" or "False". This criterion allowed us to formulate the following considerations on the results obtained:

- BAX is expressed in the attractors of all patients except BC03_X.

- EGFR is expressed in tumor cells' attractors for all patients except BC03_X and BC05_X.
- ERBB2 is expressed in the attractors of all patients, except BC03_X.
- ETV1 is expressed in the attractors of tumor cells for every patient, except BC03_X.
- IKBKG is expressed in the tumor attractors of all patients.
- MAP3K7 is expressed in the attractors of all patients except BC01_X.
- ST14 is expressed in the attractors of all patients except BC01_X.
- PLAT is expressed in patient attractors BC02_X, BC03_X, BC07_X, BC07LN_X, BC08_X, BC10_X, BC11_X.
- DDR1 is expressed in the tumor attractor of patient BC04_X.

These results are summarized in Figure 3.7.

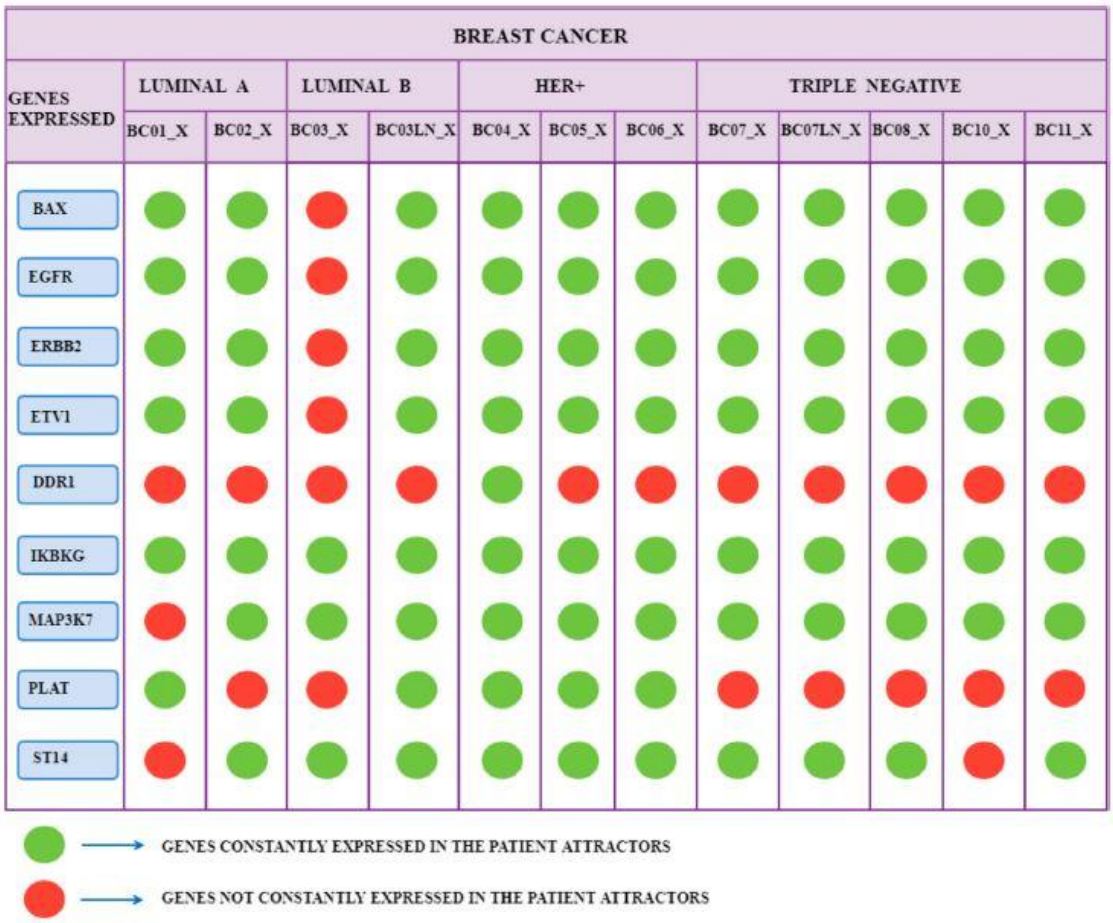

Fig. 3.7: The 12 scRNA-seq groups of breast cancer samples of 10 patients, translated into attractors, are divided into four subtypes of tumor: Luminal A and B, HER+, and

TNBC. The green color indicates that the corresponding gene has a constant Boolean value (True or False) for all the patient's attractors. The red color indicates that the state of the gene does not remain constant for all the attractors of a specific patient.

Interestingly, PLAT is never expressed in patients with breast cancer classified as TNBC. Unlike for patients of Luminal A and Luminal B in which the inactivity of PLAT affects 50% of patients, this characteristic covers all TNBC group cases (Figure 3.7).

Further considerations concern the comparison between the attractors of malignant cells with other types of cells from the same patient. For example, in the BC07LN_ patient sample, it is interesting to compare EEF1G in malignant and immune B-cells. In the first case, the gene is expressed in only 4.2% of the attractors detected (1/24), while in the second case, it is expressed in 36.9% of the attractors detected (7/19). Considering DDR1, the attractor rate of expression was 60% (9/15) in the patient sample BC08_X. For stromal cells, the attractor level of expression for the same gene is 100% (6/6).

3.3 Discussion

The widely spread use of Boolean networks to model gene regulatory network dynamics is well-established in the scientific community. Identifying attractors with this type of model enables the elucidation of long-term cell functioning, which corresponds to a particular phenotype in molecular biology. An attractor is a stable state of the cell. Starting from an initial point of the state space, the cell dynamics simulated by the model induce a sequence of states driven by the regulatory interaction established between the network nodes until reaching an equilibrium. This stable set of states manifests itself with the repetition of the configuration of the network in its Boolean values in a fixed or ciclica way. The initial point from where the trajectory started is part of the basin of attraction of a given attractor in which all the points (or state spaces) contained in it converge.

This work's central hypothesis is the interpretation of cancer phenotypes as basins of attraction in the epigenetic landscape [Huang et al., 2009]. Another central assumption is that the perturbation of a subset of genes can produce the transition from

one basin of attraction to another, which corresponds to another phenotype [Crespo et al., 2013]. Therefore, we modeled the dynamics of breast cancer through the identification and description of the attractors associated with a specific gene regulatory network in such a way as to be able to find out the essential genes that determine the formation of the basin of attraction. These essential genes are potential therapeutic targets.

In large gene regulatory networks (with more than 100 vertices, such as the one presented in this work), it is possible to adopt different approaches to define attractors to overcome the exponential growth of the state space size according to the increase in the number of network size. For example, one approach was to partition the network into small subnets, finding the attractors corresponding to these partitions and then combining them to build a stable state relative to the entire network [Hong et al., 2015]. Another approach is to configure network input vertices with initial Boolean values representing their gene expression level [Cho et al., 2016].

This report searched the network attractors that result from the topological features and logical functions attributed to each vertex, given the binarized scRNA-seq data available. The use of single cell data, allows a better description of the heterogeneous nature of cancer with a consequent better therapeutic outcome, unlike Bulk RNA-seq data which provide average expression levels of a cell population that may include tumor cells and other cell types. The attribution of a specific Boolean value to each vertex of the network, obtained through gene expression binarization, conditioned the initial conditions in the system's state space. These initial conditions were the starting points of a trajectory, driven by the topology and the logic functions characterizing the network, whose evolution ended when reaching an attractor.This strategy is very time efficient, avoiding the nonpolynomial complexity of other strategies to find network attractors [Hong et al., 2015]. The result obtained can be considered satisfactory given the percentage of attractors obtained.

From the analysis of the attractors obtained in this work, we extracted peculiar characteristics on several genes, demonstrating the need for a theranostics approach based on specific patient data. Key genes frequently expressed in attractors identified in this report were cited in the scientific literature related to breast cancer. They are BAX[Sturm et al., 2000], DDR1 [Belfiore et al., 2018], EGFR[Bhargava et al., 2005], ERBB2 [Vernimmen et al., 2003)], ETV1 [Ouyang et al., 2015], MAP3K7 [Zhou et al., 2017], PLAT [Theillet et al., 1993], ST14 [Kauppinen et al., 2010].

BAX pro-apoptotic protein is differentially expressed in breast tumors by the BAX gene. The tumor suppressor gene TP53 regulates the expression of BAX and its mediated apoptosis. A reduced level of BAX expression is an adverse prognostic factor in breast cancer [Sturm et al., 2000].

DDR1, a non-integrin collagen tyrosine kinase receptor, plays an essential role in cellular communication with the microenvironment. It is differentially expressed in several malignant tumors, playing an essential role in tumor progression, including breast cancer [Belfiore et al., 2018].

EGFR is an epidermal growth factor receptor protein. It is part of pathways that control several key biological processes like angiogenesis, cellular proliferation, and apoptosis avoidance. Indeed, it is worth highlighting the FDA approved GEFITINIB availability, an anticancer drug that acts as an EGFR tyrosine kinase inhibitor. In a sample of 175 breast cancer cases, there was EGFR amplification in 11 of them. On these 11, 10 (91%) had an EGFR protein overexpression detected by immunohistochemistry [Bhargava et al., 2005].

ERBB2, commonly referred to as HER2, encodes a member of the epidermal growth factor (EGF) receptor, a family of tyrosine kinase receptors. About 30% of invasive breast carcinomas overexpress this gene and are correlated with poor prognosis. HER2 encodes a 185kDa transmembrane receptor belonging to the EGFR group. The monoclonal antibody Trastuzumab effectively inhibits the growth of breast cancer tumors that overexpressed HER2 [Vernimmen et al., 2003].

The ETV1 protein (together with ETV4 and ETV5) forms the PEA3 subfamily of ETS transcription factors. The PEA3 group could be a tumorigenic factor in breast cancer. ETV1 expression is higher in TNBC tissues compared to normal tissues. Negative regulation of ETV1 can activate COP1 as a tumor suppressor in patients with TNBC [Ouyang et al., 2015].

MAP3K7 is an enzyme that is encoded by the MAP3K7 gene. This protein controls a series of cell functions like apoptosis and transcription regulation. Cell growth assessment performed by MTT assay showed an increase in MAP3K7 expression in breast cancer tissues compared with non-malignant breast tissue [Zhou et al., 2017]. Given the crucial role of this protein in other types of cancer [Rodrigues et al., 2015; Cheng et al., 2019; Washino et al., 2019], it would be interesting investigate in more detail its role in breast cancer.

PLAT encodes tissue-type plasminogen activator, a serine protease that transforms the proenzyme plasminogen to plasmin, an enzyme. Reports in the literature point out the amplification of PLAT in breast cancer. Literature reports indicate that 15.6% of breast cancer tumors present PLAT amplification [Theillet et al.,1993]. It is also interesting to note the impact on gene expression related to migration and invasion in breast cancer, especially PLAT, obtained from docosahexaenoic acid (DHA), which emerged in a recent study [Chénais et al., 2020].

ST14 encodes a matriptase protein. It is an epithelial-derived integral serine protease. The overexpression of this protein is associated with low tumor survival in node-negative breast cancer cases. It also seems that a coordinated overexpression of ST14 and other genes (MNP and MST1R) is associated with metastasis and poor breast tumor prognosis [Kauppinen et al., 2010].

IKBKG gene encodes the NF-KAPPA-B essential modulator (NEMO), a protein that is the regulatory subunit of the IKB kinase complex's inhibitor. This protein's overexpression may occur in cases of *inflammatory breast cancer* (IBC), a rare form of breast cancer characterized by a particular phenotype [Lerebours al., 2008]. As this protein is often highlighted in the literature for its role as a growth and progression factor in several types of cancer [Karin, 2006], it might be appropriate to thoroughly investigate its role in breast cancer development.

All those scientific evidence confirm the effectiveness of the approach proposed in this work to identify biomarkers and potential therapeutic targets. The present report also produced a detailed list of genes never expressed in the attractors obtained. An example is the ANXA1, never expressed in the attractors related to the breast cancer sample BC07_X. The level of expression of the protein produced by ANXA1, seems to indicate poor overall survival in TNBC [Gibbs and Vishwanatha, 2018]. Another example is the SMAD4, never expressed in the attractors of BC02_X. The protein produced by this gene is part of the SMAD family of transcription factor proteins, which acts on the TGF-β signal transduction. SMAD4 expression was lower in breast cancer tissue than in the surrounding breast epithelium [Stuelten et al., 2006]. These constantly not-expressed genes in tumor attractors can be used as biomarkers for diagnostics, predictive, and prognostic purposes [Lerebours et al., 2008], awaiting further research advances on the challenge of increasing gene-level expression using CRISP techniques [Matharu et al., 2019].

It is worth noting that even if we based the choice of genes that compose the network on differentially expressed genes from the MDA-MB-231 cell line, which is associated with the TNBC subtype, we succeed in obtaining attractors for other cancer subtypes as well. This result indicates that the method used to include intermediary vertices from the human interactome and related transcription factors is robust enough to capture key genes possibly involved in all major breast cancer subtypes.

It is significant to highlight another point that emerged in this report: the ERBB2 gene is a therapeutic target in breast cancer for which specific drugs exist [Gomez et al., 2008]. ERBB2 is constantly expressed in all patients analyzed in this report except for one, the patient BC03-X. For this reason, BC03-X may not need the type of therapeutic intervention related to ERBB2. This consideration allows us to place our method in the context of personalized medicine. Nevertheless, further specific algorithm development for defining the more appropriate therapeutic approach for each patient is needed.

Different settings in specific parts of the procedure illustrated in this report may be further studied. For example, one example of future work is to compare the results obtained with both asynchronous and synchronous models on the network dynamics. Another example of future work is to use specific logic functions for each node of the network instead of the nested canalyzing function approach used in this analysis.

3.4 Chapter conclusion

In this work, we model the complex dynamics of a gene regulatory system related to breast cancer using scRNA-seq data. We computed the attractors of the analyzed cells, as well as the genes related to attractor stability. Each group of cells belongs to a different patient, and a certan degree of differentiation between the various patients was found in the genes characterizing the attractors. This characterization drives therapeutic actions differentiated from patient to patient based on the analysis that emerged. These considerations allow us to frame the system developed in this report within the paradigm of personalized medicine. This work can be expanded in many ways. One significant advancement will be to define an algorithm to define optimal therapeutic interventions based on the analysis of the model. One crucial optimization parameter for this algorithm is to minimize the number of therapeutic interventions while providing maximum efficacy. Another contribution is to evaluate if asynchronous boolean

modeling can provide new insights compared to synchronous boolean modeling. Our group intends to explore those questions soon.

Optimizing Therapeutic Targets

Studying gene regulatory networks associated with cancer provides valuable insights for therapeutic purposes, given that cancer is fundamentally a genetic disease. However, as the number of genes in the system increases, the complexity arising from the interconnections between network components grows exponentially. In this study, using Boolean logic to adjust the existing relationships between network components has facilitated simplifying the modeling process, enabling the generation of attractors that represent cell phenotypes based on breast cancer RNA-seq data. A key therapeutic objective is to guide cells, through targeted interventions, to transition from the current cancer attractor to a physiologically distinct attractor unrelated to cancer. To achieve this, we developed a computational method that identifies network nodes whose inhibition can facilitate the desired transition from one tumor attractor to another associated with apoptosis, leveraging transcriptomic data from cell lines. To validate the model, we utilized previously published in vitro experiments where the downregulation of specific proteins resulted in cell growth arrest and death of a breast cancer cell line. The method proposed in this manuscript combines diverse data sources, conducts structural network analysis, and incorporates relevant biological knowledge on apoptosis in cancer cells. This comprehensive approach aims to identify potential targets of significance for personalized medicine.

Cancer is a disease characterized primarily by uncontrolled cellular proliferation. This dysregulation disrupts normal cellular homeostasis, leading to the emergence of distinctive traits known as "hallmarks of cancer," which are common across different tumor types [Hanahan D, 2022]. Carcinomas, a type of epithelial cell tumor, account for approximately 85% of all cancers and can affect various tissues in the human body. When these tumors occur in glandular tissue, they are specifically referred to as adenocarcinomas. Breast cancer falls into the category of adenocarcinomas. It is the

most prevalent neoplastic condition affecting women, with a global incidence of approximately 2,261,419 new cases and 684,996 deaths in 2020 [Sung, Hyuna et al., 2021].

In addition, treating this pathology gives rise to harmful adverse effects in patients. For instance, a study identified 38 distinct negative symptoms categorized into five groups that resulted from chemotherapy administration [Chan et al., 2017]. Therefore, developing new intervention strategies that can enhance therapies and minimize their unwanted side effects is crucial. We propose using a Boolean modeling approach for breast cancer to address this need. Cancer is a genetic disease with multifaceted ramifications [Vogelstein and Kinzler, 2004]. Cancer cells' DNA undergoes numerous alterations due to the oncogenic process, including single-base pair mutations, indels, and epigenetic modifications.

Cancer occurrence leads to network modifications, where the pathways involved are frequently intertwined to generate processes characteristic of tumor dynamics and progression [Yuan et al., 2017; Barillot et al., 2013; Feng et al., 2018]. Epigenetic changes and alterations in gene regulatory networks [Waddington, 1957] provide an opportunity for modeling cancer attractors [Huang et al., 2009]. This study builds upon a previous analysis [Sgariglia et al., 2021] of attractors identified within a gene regulatory network based on breast cancer data. Specifically, by incorporating a novel set of genes associated with apoptosis into the Boolean network, we identified new attractors resulting from target inactivation. This modeling enabled our gene model to transition towards a cell death phenotype, as observed in corresponding *in vitro* experiments [Tilli et al., 2016].

This paper presents an algorithm that optimizes the selection of network elements capable of inducing trajectories between attractors in the epigenetic landscape. Additionally, we have introduced an indicator that quantifies the network's response when inducing a trajectory from a malignant state to an apoptosis state through direct intervention on its vertices. We incorporated a set of genes representing the apoptosis process into the gene regulatory network associated with breast malignancy to achieve this. By manipulating the activation or inhibition state of each gene in this group, we assessed the effectiveness of network perturbations in transitioning the phenotype from malignancy to apoptosis. We calibrated the network based on (i) the typical gene expression level observed in the malignant attractor and (ii) the genes to be inhibited for inducing apoptosis in a malignant cell line, as determined from *in vitro* experiments. To

validate the system dynamics, we compared the results with the *in vitro* experiment [Tilli et al., 2016], where five genes were silenced to induce the death of a breast cancer cell line. This experimental data was used to evaluate the network's behavior upon the permanent silencing of specific targets.

We confirmed that the network structure derived from the interactome could drive the malignant attractor toward apoptosis by selectively silencing the same network vertices as those targeted in the *in vitro* experiment. The capacity of our model to replicate the conditions conducive to malignant cell death observed *in vitro* enabled us to optimize the selection of targets for transitioning the system dynamics from the malignant attractor to the apoptosis state. This optimization process involved utilizing specific analysis techniques to examine the network structure, enabling us to identify the vertices whose inhibition could mimic and enhance the outcomes achieved in the *in vitro* experiment.

4.1 Materials and Methods

The various stages involved in conducting this research are briefly outlined in Figure 4.1.

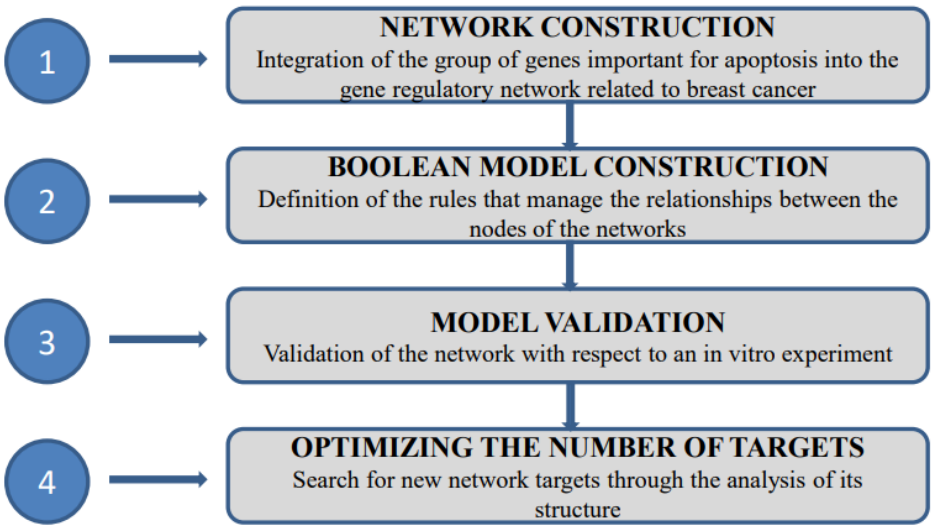

Fig. 4.1: Steps for Boolean network construction and dynamic emulation..

4.1.1 Network construction

We initiated our study using a gene regulation network established in a previous publication [Sgariglia et al., 2021]. Breast cancer RNA-seq data guided the selection of network nodes, and the chosen genes were linked to their respective hallmarks using the MSigDB repository. Notably, the two crucial hallmarks of cancer, namely "UNLIMITED REPLICATIVE POTENTIAL" and "EVASION OF CELL DEATH," were well-represented in the dataset. We introduced an additional set of 28 genes to supplement the initial network consisting of 103 genes [Sgariglia et al., 2021] to enlarge the network. Twenty-five incorporated nodes were for apoptosis-associated genes, exerting either inducing or inhibitory effects on this cellular process. Alongside these 25 novel vertices, two existing vertices from the previous study assume a crucial role in cellular apoptosis as constituents of the apoptotic cascade. Collectively, we refer to these 27 genes as apoptosis-related genes. This network enlargement was imperative to facilitate the modeling of the transition from the malignant state to the apoptosis one, which was induced by network perturbations through targeted inactivation of specific vertices. Protein-protein interactions were acquired from the IntAct interactome (IntAct database, version updated in December 2017) to establish the network connections between genes. The specific file used for obtaining the interactions was retrieved from the FTP link ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/intact-micluster.txt, accessed on January 11, 2018. The directionality of the connections and their regulatory nature (activation or inhibition) were determined by consulting the Metacore database [Ekins et al., 2007]. To verify the network's validity in replicating a tested biological scenario *in vitro*, our system was configured to reproduce the outcomes documented by Tilli et al. [Tilli et al., 2016]. Their experimental study demonstrated cell death in a cancer cell line (MDA-MB-231) by inhibiting five genes using RNA interference. To achieve this objective, we retrieved the RNA-seq data of two distinct cell lines, namely MCF10A and MDA-MB-231, from the Gene Expression Omnibus (GEO) repository available at https://www.ncbi.nlm.nih.gov/gds/. MDA-MB-231 is a malignant cell line derived from triple-negative breast cancer, while MCF10A served as the non-tumoral control in this study. For MCF10A, we obtained the following RNA-seq datasets: SRR2149928, SRR2149929, SRR2149930, SRR2870783, and SRR2872995. Regarding MDA-MB-231, we acquired the RNA-seq datasets:

ERR493677, ERR493680 (corresponding to the *body* portion), and ERR493678, ERR493679 (corresponding to the *protrusion* portion) of the cells.

Based on information obtained from the GEO repository, the cells were cultured on a polycarbonate transwell filter with 3-micrometer pores, allowing the formation of protrusions through the pores for 2 hours. Subsequently, the cells underwent a washing step, and both sides of the filter were lysed to extract RNA for further analysis. Through this protocol, the cells were fractionated into two distinct fractions: *protrusion* and *body* types. In the subsequent analysis, we employed a binary approach to categorize the up-regulated genes observed in each MDA-MB-231 RNA-seq dataset (both *body* and *protrusion*) compared to every MCF10A RNA-seq dataset. This approach introduced variability into the experiment and facilitated the assessment of system robustness.

To ensure the convergence of our network model with scale-free networks, which is characteristic of cell signaling pathways [Albert, 2005], we examined the degree distribution of the network vertices. To evaluate the network's structure, we compared its degree distribution with that of random graphs [P. Erdós and A. Rényi, 1959], Watts and Strogatz small-world networks [Watts DJ and Strogatz SH, 1998], and scale-free networks [Albert, 2005], all generated with the same number of vertices. To facilitate this comparison, we utilized the complementary cumulative distribution function (CCDF) as defined by equation 1. The CCDF provides the probability (F) of a vertex having a connectivity degree equal to or greater than a specified value. By analyzing these distributions, we could determine the degree of alignment between our network and these reference models.

$$\boldsymbol{F(k) = \sum_{k'=k}^{\infty} P(k')} \qquad (4.1)$$

4.1.2 Boolean model construction

After constructing our model's directed graph, we established the conditions necessary for its dynamic simulation by defining the transfer functions that govern the system's evolution at discrete time intervals. The objective was to guide the dynamic behavior of the network elements in a manner that faithfully replicated the observed conditions from the *in vitro* experiment conducted by Tilli et al. [Tilli et al., 2016]. By

incorporating these specific conditions, we aimed to ensure the accurate representation of the experimental findings within our model's dynamic framework. In systems biology, accurately deducing the interaction rules of a network poses a significant challenge. To address this complexity, we employed Boolean nested canalizing functions [Nikolajewa et al., 2006], where the function is influenced by the specific order in which variables are organized. A Boolean function is considered canalizing if a single input can solely determine the output. In cases where this input does not play the canalizing role, the other inputs are deemed responsible for fulfilling this function. By adopting the hierarchical structure of transfer functions, achieved through nested canalizing functions, we aimed to capture the behavior of biological systems more effectively [Kauffman, 1993; Shmulevich et al., 2003]. Furthermore, many network nodes exhibited a substantial number of inputs, emphasizing the need for a robust modeling approach. In this scenario, using nested canalizing functions offers increased system stability [Kauffman et al., 2004], which is crucial in managing the inherent noise observed in biological systems.

The utilization of nested canalizing functions and the need to align the model with biological facts enabled us to manually establish the transfer rules for each gene [Kauffman et al., 2004; Szallasi and Liang, 1998]. Opportunely, Harris [Harris et al., 2002] demonstrated that a significant portion of the gene updating rules fell under canalizing functions. Considering these considerations, we determined the number of inputs for the Boolean functions, as defined in Equation 4.2.

$$\boldsymbol{f(x) = f(x_1, \ldots., x_z)} \tag{4.2}$$

Considering a set of Boolean variables $\boldsymbol{x = \{x_1, x_2, \ldots, x_z\}}$, the input was defined as essential (see equation 4.3) if the condition of equation 4.3 was satisfied.

$$\boldsymbol{f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_z) \neq f(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_z)} \tag{4.3}$$

The essential inputs were defined as *canalizing* if there were values $\boldsymbol{a, b} \in \{\boldsymbol{0, 1}\}$ that satisfied equation 4.4 for all remaining combinations of variables $\boldsymbol{x - \{x_i\}}$, where *x* is a canalizing input value, and x_i is a canalized value.

$$\boldsymbol{f(x_1, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_z) = b} \tag{4.4}$$

A function with $\boldsymbol{z}$ essential input is defined as nested if it is z-times canalized with $\boldsymbol{x_1, \dots, x_z}$ canalizing inputs and $\boldsymbol{a_1, \dots, a_z}$ canalizing values, to which correspond the canalized value $\boldsymbol{b_1, \dots, b_z}$.

Nested canalizing functions are an extension of the canalizing function formalism [Nikolajewa et al., 2007], in which the order of the inputs is considered to assign the canalizing role. In addition, canalizing functions can be nested if it is possible to set them with $\boldsymbol{z}$ inputs and $\boldsymbol{z-1}$ Boolean operators AND ($\wedge$) or OR ($\vee$) with a priority proceeding from left to right. Thus, defining $(* \in \{\wedge, \vee\})$, we have equation 4.5.

$$\boldsymbol{f_i(x_1, x_2, \dots, x_z) = x_{i1} * \left(x_{i2} * \left(\dots (x_{iz-1} * x_{iz+1})\right)\right)} \tag{4.5}$$

In the implementation phase of this report, each input of the function was coupled to a single logical operator, which can be an $\wedge$ (and) or an $\vee$ (or). In light of these rules, the transfer functions of the network have been implemented according to Equation 4.6

$$\boldsymbol{node_c = (a_1 \vee a_2 \vee \quad \dots \vee a_n) \wedge \neg\left(b_1 \wedge \neg\left(b_2 \wedge \neg(\dots \wedge \neg(b_m))\right)\right)} \tag{4.6}$$

where $\boldsymbol{node_c}$ is the n^{th} node of the network, $\boldsymbol{a_i}$, $\boldsymbol{i = [1, n]}$, $\boldsymbol{n}$ represents the number of nodes with activation function on $\boldsymbol{node_c}$, and $\boldsymbol{b_j}$, $\boldsymbol{j = [1, m]}$, $\boldsymbol{m}$ is the number of nodes with inhibition function on $\boldsymbol{node_c}$. To create suitable conditions for the network to reproduce the results obtained in the *in vitro* experiment [Tilli et al., 2016], we applied some changes to the general scheme of the nested canalizing functions illustrated above in the nodes representing the *TP53*, *HIF1A*, *RELA*, *NFKB1*, *HDAC1*, *STAT3*, *BCL2*, *CASP3*, and *BRCA1* genes as shown in equations 4.7 to 4.11 where some input variables with activation or inhibition roles on $node_c$ ceased to be independent of the other elements of the function and assume a cumulative role for the final result, which the other nodes cannot replace.

$$\boldsymbol{node_{TP53,HIF1A,CASP3} = (a_1 \vee a_2 \vee, \dots, \vee a_n) \wedge (\neg b_1) \wedge (\neg b_2 \vee , \dots, \vee \neg b_m)} \tag{4.7}$$

$$\boldsymbol{node_{RELA} = (a_1) \wedge (a_2 \vee, \dots, \vee a_n) \wedge (\neg b_1) \wedge (\neg b_2 \vee \quad , \dots, \vee \neg b_m)} \tag{4.8}$$

$$\boldsymbol{node_{NFKB1}} = (\boldsymbol{a_1}) \wedge \left(\boldsymbol{a_2} \vee \ , \ldots, \boldsymbol{a_n}\right) \wedge (\neg \boldsymbol{b_1} \vee \neg \boldsymbol{b_2} \vee \ , \ldots, \vee \neg \boldsymbol{b_m}) \quad (4.9)$$

$$\boldsymbol{node_{HDAC1}} = (\boldsymbol{a_1}) \wedge (\boldsymbol{a_2} \vee \ , \ldots, \vee \boldsymbol{a_n}) \quad (4.10)$$

$$\boldsymbol{node_{STAT3,BCL2,BRCA1}} = \left(\boldsymbol{a_1} \vee \boldsymbol{a_2} \vee_{,} \ldots, \vee \boldsymbol{a_n}\right) \wedge (\neg \boldsymbol{b_1} \vee \neg \boldsymbol{b_2} \vee_{,\ldots,} \vee \neg \boldsymbol{b_m}) \quad (4.11)$$

After defining the constituent elements of the gene regulation network and analyzing its structure, the subsequent task was determining the appropriate mathematical formalism for the system's dynamic analysis. We opted to employ a directed graph model based on Boolean logic. Boolean network modeling represents one of the simplest methods for dynamic modeling while offering the advantage of reliably providing insights into system dynamics.

In this context, we considered a Boolean variable, denoted as $\boldsymbol{B}$, which takes on the value of True (1) or False (0) depending on whether a particular gene is up-regulated or not in the RNA-seq data of the MDA-MB-231 (malignant) cell line compared to the MCF10A (control) cell line. Consequently, for the $\boldsymbol{n}$ vertices within our network, we can express this relationship using equation 4.12.

$$\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \ldots., \boldsymbol{x_n}\}, \ \ \boldsymbol{x_i} \in \boldsymbol{B} \quad (4.12)$$

When time is represented as a discrete scalar value, the states of the network can be depicted as a vector with its components being the vertices of the network (equation 4.13).

$$\vec{\boldsymbol{x}} = \left(\boldsymbol{x_1}(\boldsymbol{t}), \ldots., \boldsymbol{x_n}(\boldsymbol{t})\right) \quad (4.13)$$

The trajectories of the system within the state space are then contingent upon the Boolean functions associated with the n vertices of the network (equation 4.14).

$$\boldsymbol{x_i}(\boldsymbol{t}+\boldsymbol{1}) = \boldsymbol{f_i}\left(\vec{\boldsymbol{x}}(\boldsymbol{t})\right), \boldsymbol{f_i}: \boldsymbol{B^n} \rightarrow \boldsymbol{B} \quad (4.14)$$

In this report, we used a synchronous update mode for the network vertices, wherein all vertices are updated simultaneously. While an asynchronous update mode may align better with biological realism, the choice of update mode is not crucial given the computational and conceptual advantages of synchronous updates and the enhanced system stability achieved through the utilization of nested canalized transfer functions [Kauffman et al., 2003]. In the synchronous update mode, the system's progress occurs in consecutive temporal states (equation 4.15).

$$\boldsymbol{T}\left(\boldsymbol{x_i^{\rightarrow t}}, \boldsymbol{x_i^{\rightarrow t+1}}\right) = \boldsymbol{T_1}\left(\boldsymbol{x_1^t}, \boldsymbol{x_1^{t+1}}\right) \wedge \ldots \wedge \boldsymbol{T_n}\left(\boldsymbol{x_n^t}, \boldsymbol{x_n^{t+1}}\right) \quad (4.15)$$

The goal of Boolean modeling is to identify the attractors expressed by the dynamics of the system. Attractors are stable gene activity patterns that represent the long-term behavior of the Boolean network and are interpreted as a specific cellular phenotype.

Attractors in a Boolean system can be subdivided into different classes. Examples are fixed-point attractors, characterized by a single state of the system (i.e., the Boolean configuration of network nodes) that persists indefinitely, and cyclic attractors, characterized by a sequence of states that repeat periodically. Each attractor is matched with a specific basin of attraction, composed of all the system states for which it represents the stable state at the end of their dynamic evolution.

4.1.3 Model validation

Initially, we compared MDA-MB-231 RNA-seq samples (two from the *body* and two from the *protrusion*) and each corresponding MCF10A sample. To achieve this, we employed the Reads per kilobase of transcript per Million reads mapped (RPKM) normalization process, as outlined in Pires [Pires et al., 2021], to normalize the read counts of the twenty paired RNA-seq samples. Subsequently, we subtracted each normalized value of the MCF10A RNA-seq sample from the corresponding MDA-MB-231 data. For positive values (indicating up-regulated genes in the malignant state), we

applied a logarithmic transformation based on equation 4.16, utilizing the pipeline described by Pires [Pires et al., 2021].

$$y = x*Log_2(x+1) \quad (4.16)$$

The procedure involved in this pipeline consists of classifying genes as up-regulated based on whether the logarithmic transformation of their differential expression surpasses a critical threshold. To determine this critical value, a Python script was employed to fit a Gaussian curve with a 95% confidence level to the data for a p-value of 0.025 (for more details, refer to Pires et al., 2021). In each of the twenty comparisons between MBA-MD-231 and MCF10A, genes that were identified as up-regulated through this process were assigned a value of "1", while the remaining genes were assigned a value of "0" (**Supplementary material S1**).

Input : V nodes of the Boolean network; N cell lines with body-fraction MDA-MD 231 RNA-Seq values ($B_{n,v}$); M cell lines with RNA-Seq values MDA-MD 231 of protrusion fraction ($P_{m,v}$); C cell lines of MCF10A type RNA-Seq values ($CO_{c,v}$);

Output : Boolean values of RNA-Seq

1 ***for*** *all* $n \in N$ ***do***
2 ***for*** *all* $v \in V$ ***do***
3 $B_{n,v} \leftarrow RPKM_{upper} = \frac{RC_g * 10^9}{L * (RC_{pc} - (\delta * RC_{pc}))}$
4 ***for*** *all* $m \in M$ ***do***
5 ***for*** *all* $v \in V$ ***do***
6 $P_{m,v} \leftarrow RPKM_{upper} = \frac{RC_g * 10^9}{L * (RC_{pc} - (\delta * RC_{pc}))}$
7 ***for*** *all* $c \in C$ ***do***
8 ***for*** *all* $v \in V$ ***do***
9 $CO_{c,v} \leftarrow RPKM_{upper} = \frac{RC_g * 10^9}{L * (RC_{pc} - (\delta * RC_{pc}))}$
10
11 ***for*** *all* $v \in V$ ***do***
12 ***for*** *all* $n,m \in (N, M)$ ***do***
13 ***for*** *all* $c \in C$ ***do***
14 $gene\_diff_{v,n,c} \leftarrow B_{n,v} - CO_{c,v}$
15 $gene\_diff_{v,m,c} \leftarrow P_{m,v} - CO_{c,v}$
16 $G_{v,n,c} \leftarrow gene\_diff_{v,n,c} + lg_2(gene_{diff\ v,n,c} + 1)$
17 $G_{v,m,c} \leftarrow gene\_diff_{v,m,c} + lg_2(gene_{diff\ v,m,c} + 1)$
18
19
20 $P\text{-}value \leftarrow 0.025$
21 ***for*** *all* $n,m \in (N, M)$ ***do***
22 ***for*** *all* $c \in C$ ***do***
23 $\mu_{n,c} = mean_{n,c}$
24 $\mu_{m,c} = mean_{m,c}$
25 $\sigma_{n,c} = standard\_deviation_{n,c}$
26 $\sigma_{m,c} = standard\_deviation_{m,c}$
27 $PDF_{n,c} = \frac{1}{\sigma_{n,c}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu_{n,c}}{\sigma})^2}$
28 $PDF_{m,c} = \frac{1}{\sigma_{m,c}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu_{m,c}}{\sigma})^2}$
29 $CDF_{n,c} = P(X \leq x),\ \textbf{for all}\ x_{n,c}$
30 $CDF_{m,c} = P(X \leq x),\ \textbf{for all}\ x_{m,c}$
31 $critical\_value_{n,c} = inverse.CDF_{n,c}(1 - 0.025)$
32 $critical\_value_{m,c} = inverse.CDF_{m,c}(1 - 0.025)$

Block 1
Apply the process of RPKM for RNA-seq normalization for all genes in the network. MDA-MD 231 body (lines 1-3), MDA-MD 231 protrusion (lines 4-6), MCF10A (lines 7-9).

Block 2
Subtract each normalized value of MCF10A RNA-seq from that of each MDA-MB-231 (lines 11-15), both for body (line 14) and protrusion (line 15). A log transform is applied (body type line 16, Protrusion line 17)

Block 3
For each of the groups formed by all the nodes of the network obtained in BLOCK 2, we established the threshold (line 20), for both cellular lines
For body and protrusion lines we obtained the average (lines 23-24), standard deviation (lines 25-26), normal distribution (lines 27-28), cumulative distribution function (lines 29-30), critical threshold value through the inverse of cumulative distribution function (lines 31-32).

```
33
34 for all v ∈ V do
35   for all n,m ∈ (N, M) do
36     for all c ∈ C do
37       if G_{v,n,c} ≥ critical_value_{n,c,v}
38           G_{v,n,c} ← 1
39         else:
40           G_{v,n,c} ← 0
41       if G_{v,m,c} ≥ critical_value_{m,c,v}
42           G_{v,m,c} ← 1
43         else:
44           G_{v,n,c} ← 0
```

Block 4
Each of the values detected (lines 34-36) from the difference of RNA-seq values (lines 16-17) is compared with the specific threshold of a given group (lines 37-41). If this value is greater than the threshold is assigned 1 (lines 38-42), otherwise it's assigned the Boolean value 0 (lines 40-44).

Fig. 4.2: Procedure used for binarization of RNA-seq values. BLOCK 1: normalization of RNA-seq values. BLOCK 2: subtraction of each normalized value of MCF10A RNA-seq from each value of MDA-MB-231data and log transformation application. BLOCK 3: determination of the threshold value for which to attribute a specific Boolean value. BLOCK 4: attribution of a Boolean value based on the critical value.

Using the BooleanNet library [Albert et al., 2008], we analyzed the binary values of each MDA-MB-231 RNA-seq normalized data in conjunction with the corresponding genes from every MCF10A sample within our network (**Supplementary material S2**). Our objective was to identify the attractors generated through the dynamic evolution of the network. The presence of the 27 apoptosis-related genes in the attractors of this initial configuration served as a reference point for evaluating the impact of subsequent network modifications (Figure 4.3).

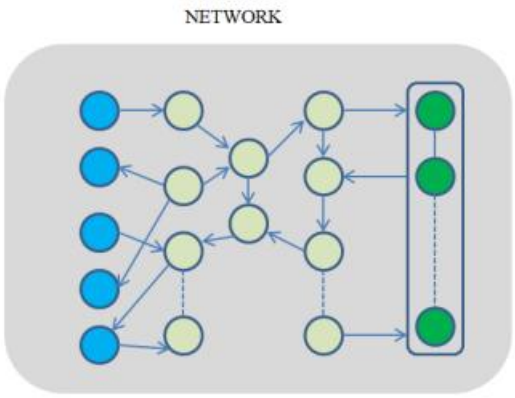

Fig. 4.3: Structure of the gene regulation network under study. The blue color indicates the five target genes, *CSNK2B*, *HSP90AB1*, *TK1*, *VIM*, *YWHAB*, which were inhibited in the *in vitro* experiment of Tilli et al. [Tilli et al., 2016]. The dark green nodes represent the apoptosis-related genes. The nodes in light green exemplify the rest of the network genes. There is no vertex inhibition in this network representation.

.

Subsequently, we conducted simulations to replicate the conditions of the *in vitro* experiment performed by Tilli et al. [Tilli et al., 2016], where the MDA-MB-231 cancer cell line's death was induced through transient inhibition of *TK1*, *VIM*, *YWHAB*, *CSNK2B*, and *HSP90AB1* genes using siRNA interference. For the sake of clarity, we will, below, refer to these five targets as *bench targets*. To emulate this *in vitro* experiment, we permanently inhibited the vertices corresponding to these bench targets in the dynamic evolution of the network, as shown in Figure 4.4.

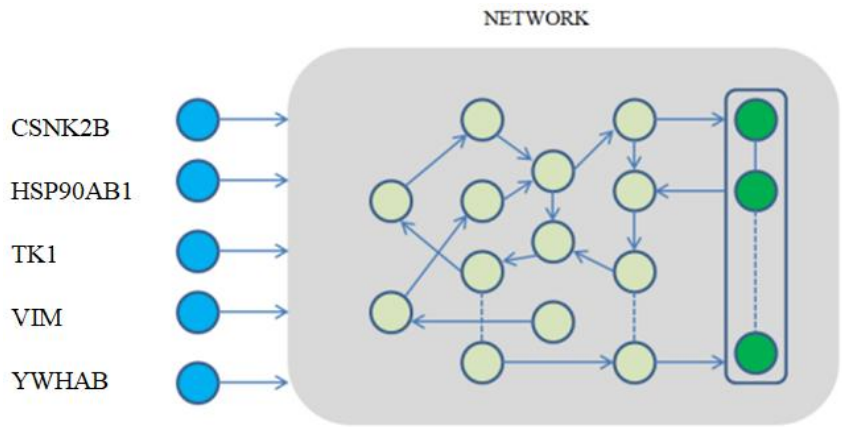

Fig. 4.4: The five bench targets (blue nodes) were set to zero (inhibition) for network dynamics emulation.

Figure 4.4 compares the activation or inhibition of the 27 apoptotic genes and their involvement in the attractors after inhibiting the five bench targets in the original network depicted in Figure 4.3. This comparison allowed us to evaluate the functional agreement between our model and the *in vitro* experiment conducted by Tilli [Tilli et al., 2016]. Furthermore, TP53 was permanently inhibited throughout the network simulation since mutations render it ineffective as a tumor suppressor in MDA-MB-231 [Yoshida et al., 2009].

4.1.4 Optimizing the number of targets

Based on the *in vitro* induction of cell death in MDA-MB-231 through the silencing of *CSNK2B*, *HSP90AB1*, *TK1*, *VIM*, and *YWHAB* [Tilli et al., 2016], and considering our understanding of the key 27 apoptosis-related genes, we present a methodology to identify genes capable of driving cancer cells towards programmed cell death.

To maximize the presence of the 27 apoptosis-related genes within the apoptosis attractor, with genes promoting apoptosis being activated and genes inhibiting it being deactivated, our objective was to identify the most effective target genes within the network structure. To achieve this, we examined the network's modularity based on the connectivity of its vertices. The Clauset-Newman-Moore greedy modularity maximization algorithm [Clauset et al., 2004] was used to identify the modular structure, considering the network as undirected. Modularity was calculated using equation 4.17, where c represents communities, L_c denotes the number of links within a community, Υ is the resolution parameter, and k_c is the sum of degrees within the community.

$$\boldsymbol{Q} = \sum_{\boldsymbol{c}\in\boldsymbol{C}}\left[\frac{\boldsymbol{L_c}}{|\boldsymbol{E}|} - \Upsilon\left(\frac{\boldsymbol{K_c}}{\boldsymbol{2|E|}}\right)^{\boldsymbol{2}}\right] \tag{4.17}$$

In the initial stage of the algorithm, each node is assigned to its own cluster, forming a partition. The algorithm then proceeds iteratively, merging pairs of clusters to increase modularity. The initial modularity value is negative, representing a singleton cluster, and gradually increases until reaching a positive peak, corresponding to the optimal solution found by the algorithm. Eventually, the modularity value returns to zero when all nodes are in the same community. In a backward process, the algorithm identifies the partition corresponding to the peak value. The implementation of this algorithm utilized the NetworkX Python library [Hagberg et al., 2008]. After detecting the communities in the network, we examined whether the 27 apoptosis-related genes were grouped or dispersed among these identified modules, with the possibility of forming a single community by the algorithm's criteria. This approach has previously been implemented in [Takeshi et al., 2014], where a modularized network was used to

map drug targets for cancer and identify modules that were the focus of therapeutic action.

Due to the observed clustering of apoptosis-related genes in the network structure, we searched nodes that could be bridges among these apoptosis-related clusters and the five bench targets. Identifying these bridge nodes could potentially shorten the path for target genes to replicate the outcomes of the *in vitro* experiment. To accomplish this, we employed the Dijkstra algorithm using the NetworkX library in Python [Hagberg et al., 2008] to find the shortest path between each of the five bench targets (*CSNK2B*, *HSP90AB1*, *TK1*, *VIM*, *YWHAB*) and every apoptosis-related gene.

A similar approach was adopted by George [George et al., 2006], wherein intermediate genes along the shortest path were identified as potential therapeutic targets. These genes were then ranked based on the number of shortest paths in which they were involved.

Utilizing the knowledge gained from the *in vitro* experiment, which demonstrated that inhibiting the five bench targets resulted in cell death of MDA-MB-231, we sought alternative vertices that, when inhibited, would activate the maximum number of genes within the apoptosis group. To accomplish this, we aimed to identify the smallest set of vertices that shared the common property of being involved in at least one shortest path between each of the five genes (*CSNK2B*, *HSP90AB1*, *TK1*, *VIM*, *YWHAB*) as starting nodes and any apoptosis-related gene as the final node (Figure 4.5).

Input: *List of in vitro experiment target genes G; list of apoptosis gene group A; list of network nodes V; adjacent list of the network T.*

Output: *New_target_list.*

```
1  for all g ∈ G do
2      for all a ∈ A do
3  C_ga = T.shortest_path_list_from_g_to_a
4          for all v ∈ V do
5              if v = g or v = a
6                  C_ga ← C_ga {v}
7  for all v ∈ V do
8      for all g ∈ G do
9          for all a ∈ A do
10             if v ∈ C_ga
11                 {...}L_g ← v
12 for all v ∈ V do
13     for all g ∈ G do
14             if v ∈ L_g
15                 x_v ← v + 1
16 New_target_list ← v ∈ max{...} x_v
```

Block 1
All the shortest paths from the in vitro experiment target gene g to each gene of the list of apoptosis group a are detected, obtaining a number of shortest paths equal to $g \times a$ (lines 1-3). In each of them the starting gene and the target gene are eliminated (lines 4-6).

Block 2
For each node of the network v, its presence is verified in the shortest paths detected in block 1 (lines 7-9). If the node is present in at least one of the paths existing between a gene g and all the genes of group a (lines 10), it is inserted in the specific list relating to the specific target gene g (lines 11).

Block 3
For each node of the network v (line 12), its presence is verified in each of the detected lists (lines 13-14) previously established (***Block 2***, line 11). For each list in which the v gene is present, the value of v is incremented by one (line 15). The nodes of network v with the highest x_v values are assigned to the final list "***New_target_list***" (line 16).

Fig.4.5: Pseudocode of target vertex determination in shortest paths. Lines 1-6 (block 1): Detection of the shortest paths via Python networkX library between the five bench targets and the apoptosis-related genes. Lines 7-11 (block 2): Creation of a list for each bench target gene containing the nodes detected by the algorithm on every shortest path. Line 12-17 (block 3): Insertion in the list of new targets of the genes present in the largest number of lists of the previous step. The asymptotic complexity of the algorithm is $O(\max\{C_{1,2}(G * A * V), C_3(V * G)\})$.

The methodology outlined in Figure 4.5 enabled us to identify novel target genes that can potentially influence the configuration of the apoptosis-related group. By "configuration," we refer to both the count of activated apoptosis-related genes that can

induce apoptosis and the consistency of this count across the RNA-seq comparisons conducted in this study.

To evaluate the impact of deactivating the newly identified target vertices using the shortest path strategy on the configuration of the apoptosis-related group, we examined the activation or inhibition status of the apoptosis-related genes across the twenty analyzed comparisons (ten for the *body* and ten for the *protrusion*, representing the two fractions of MDA-MB-231). This approach enabled us to compare the effects of inhibiting the new targets within the shortest paths to those obtained by inhibiting the original five bench targets.

4.2 Results

4.2.1 Gene regulatory network

The breast cancer regulatory network utilized in this study consists of the genes employed in a previous report for the computation of Boolean attractors [Sgariglia et al., 2021], which were further expanded by the inclusion of 25 additional genes (Fig. 6). These 25 genes, along with the two genes already present in the initial network, play a key role in the cellular apoptosis process.

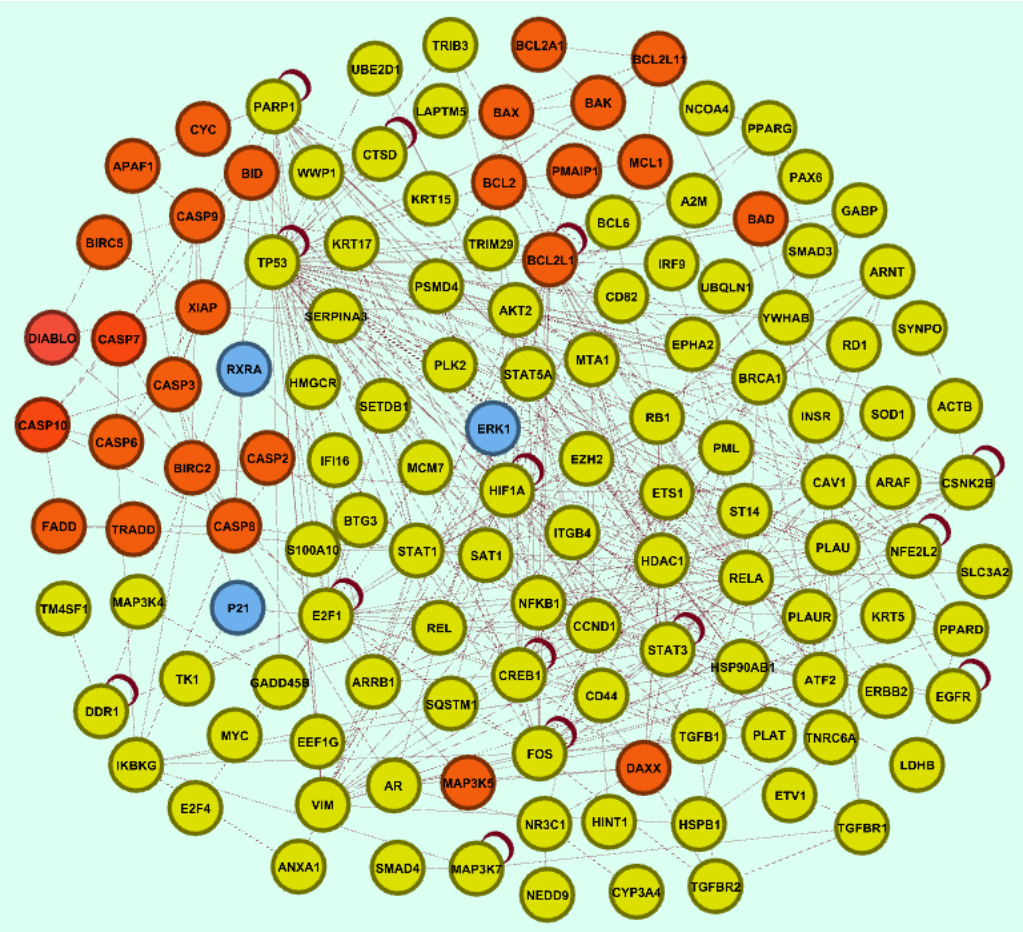

Fig 4.6: Breast cancer gene regulatory network used in this report. Twenty-five red nodes and the three blue nodes, not part of the apoptosis group, are the new vertices added to the network used in the previous report (yellow nodes). In the group of twenty-seven red genes related to apoptosis, two (*BAX* and *BCL2L1*) were already present in our earlier work network.

The network depicted in Figure 4.6 consists of **131** nodes and **494** edges. The system's dynamics are governed by Boolean transfer functions, where each node can act as either an activator or an inhibitor on the other nodes to which it is connected (**Supplementary materia S3**). The nature of these interactions was deduced using a dedicated software [Ekins et al., 2007], which also facilitated the integration of the genes from the apoptosis-related group into the pre-existing network.

4.2.2 Structural analysis of the network

We examined the structural features of the network by comparing it to well-known canonical network types. This comparison involved analyzing the degree distribution and comparing it to three distinct network types commonly described in the literature: The Erdos-Renyi network [P. Erdós and A. Rényi, 1959], the Watts-Strogatz network [Watts and Strogatz, 1998], and the Barabasi-Albert network [Albert, 2005].

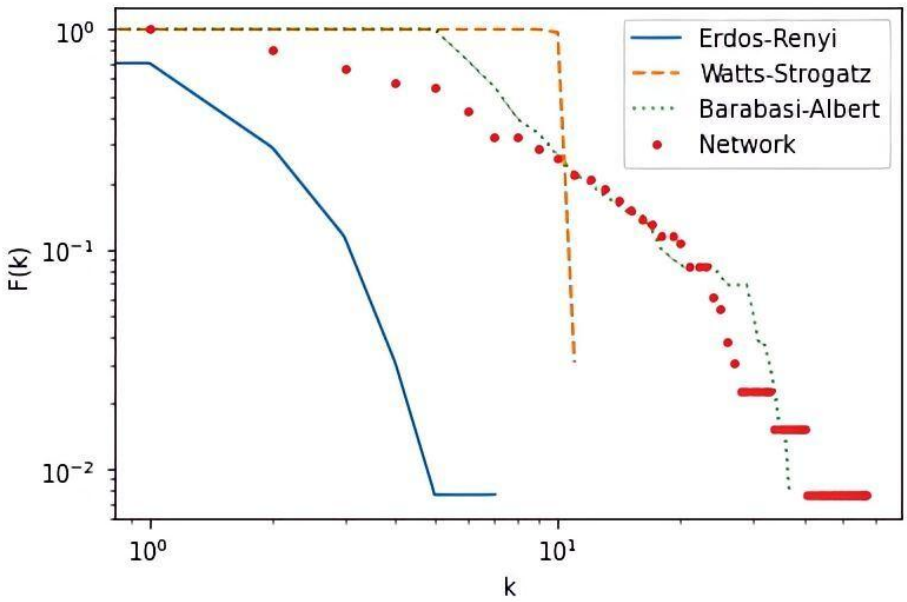

Fig. 4.7: Convergence of our model with three canonical networks. Log-log plot of Erdos and Renvi (blue line), Watts and Strogatz (orange line), and scale-free networks (green line) networks compared to the experimental network of this study (red).

By analyzing the plot presented in Figure 4.7, which depicts the complementary cumulative distribution function, we observed a striking resemblance between the network utilized in this study and the network model characterized by a power-law degree distribution. This finding aligns with the observations made by Albert [Albert, 2005], who noted that power-law distributions are commonly observed in various real networks, including those describing intricate biological systems like the network employed in our analysis.

4.2.3 Attractor analysis

Upon examining the gene regulation network depicted in Figure 4.3, specifically in its unaltered state without any vertex inhibition, we observed that the percentage of genes satisfying the requisite conditions for transitioning from the malignant state to apoptosis was 29.6% for *body* samples and 14.8% for *protrusion* samples. This finding indicates an unfavorable configuration for the cell apoptosis process, as only two genes from the Bcl-2 family demonstrated a pro-apoptosis role. Moreover, the fact that crucial genes such as *CASP3, CASP6, and CASP7*, which play a fundamental role in the intrinsic apoptosis pathway, were inactive further supported the unsuitability of the configuration. Additionally, *XIAP* and *DIABLO*, which serve as inhibitors of *CASP9* and the IAP family, failed to meet the necessary conditions for initiating an apoptosis process, as illustrated in Figure 4.8.

BODY		*PROTRUSION*	
	1-10		1-10
APAF1	F	APAF1	F
BAD	T	BAD	T
BAK	F	BAK	F
BAX	F	BAX	F
BCL2	T	BCL2	T
BCL2A1	T	BCL2A1	T
BCL2L1	F	BCL2L1	F
BCL2L11	F	BCL2L11	F
BID	F	BID	F
BIRC2	F	BIRC2	X
BIRC5	F	BIRC5	F
CASP10	T	CASP10	T
CASP2	T	CASP2	X
CASP3	F	CASP3	F
CASP6	F	CASP6	F
CASP7	F	CASP7	F
CASP8	T	CASP8	X
CASP9	F	CASP9	F
CYC	F	CYC	F
DIABLO	F	DIABLO	F
FADD	T	FADD	X
MP3K5	F	MP3K5	F
MCL1	T	MCL1	T
PMAIP1	F	PMAIP1	F
TRADD	F	TRADD	F
XIAP	T	XIAP	T
DAXX	F	DAXX	F

Fig. 4.8: Configuration of the apoptosis-related genes in the attractors of the twenty comparisons of two MDA-MB-231 sample types, *body*, and *protrusion*. The seven genes indicated in red favor a pro-apoptotic mechanism if the gene is inhibited. On the other hand, the 20 genes displayed in black favor apoptosis if genes are activated (**Supplementary material 1**). The *DAXX* gene is an exception, for which we did not find any characterization of its pro-apoptosis state in the literature. T (for True or activation), F (for False or inhibition), and X (for a continuous alternation between T and F) represent the boolean value of the 27 apoptosis-related genes within the detected attractors. T, F, and X background colors indicate if there are matches between the detected state and the ideal one of the corresponding genes for apoptosis induction. The green background means correspondence, while the orange one indicates a divergence.

To replicate the outcomes described in Tilli et al. [Tilli et al., 2016], we conducted *in silico* the inhibition of the five bench targets known to induce cell death in the MDA-MB-231 cell line *in vitro*. In the *body* samples, the percentages of genes in the apoptosis attractors that assume the state of inhibition or activation supporting cell death were 74.1% and 55.6% in the comparisons 1 to 6 and 7 to 10, respectively. As for the *protrusion* samples, we found that 41% of the genes in the apoptosis phenotype state were observed in comparisons 1 to 6, while in comparisons 7 to 10, the percentage rose to 74%. These results are depicted in Figure 4.9.

Comparing these results with those presented in Figure 4.8 (obtained without any gene inactivation in the network), there is a noticeable difference in both quantitative and qualitative aspects. Quantitatively, the percentages of genes aligned with the apoptosis state in the attractors are significantly higher, indicating a certain degree of representation of the *in vitro* experiment within the model. Qualitatively, the presence of the apoptotic state in the Bcl-2 and Caspase families and DIABLO and XIAP supports the expected outcomes of the bench experiment in the *body* 1-6 and *protrusion* 7-10 comparisons. However, in the *body* 7-10 and *protrusion* 1-6 comparisons, this alignment is only partially observed due to discrepancies in the RNA-seq profiles of the MCF10A cell lines used in these particular comparisons.

BODY			*PROTRUSION*		
	1-6	7-10		1-6	7-10
APAF1	F	F	APAF1	F	F
BAD	F	T	BAD	F	F
BAK	T	F	BAK	T	T
BAX	T	F	BAX	T	T
BCL2	F	T	BCL2	F	F
BCL2A1	F	T	BCL2A1	F	F
BCL2L1	F	F	BCL2L1	F	F
BCL2L11	T	F	BCL2L11	T	T
BID	F	F	BID	F	F
BIRC2	F	F	BIRC2	X	F
BIRC5	F	F	BIRC5	X	F
CASP10	T	T	CASP10	T	T
CASP2	T	T	CASP2	X	T
CASP3	T	T	CASP3	X	T
CASP6	T	T	CASP6	T	T
CASP7	T	T	CASP7	X	T
CASP8	T	T	CASP8	X	T
CASP9	T	T	CASP9	F	T
CYC	F	F	CYC	F	F
DIABLO	T	T	DIABLO	X	T
FADD	T	T	FADD	X	T
MP3K5	F	F	MP3K5	F	F
MCL1	F	T	MCL1	F	F
PMAIP1	T	T	PMAIP1	T	T
TRADD	F	F	TRADD	F	F
XIAP	F	F	XIAP	F	F
DAXX	F	F	DAXX	F	F

Fig. 4.9: Attractors obtained by inhibiting the five bench targets: *CSNK2B*, *HSP90AB1*, *TK1*, *VIM*, and *YWHAB*. Columns *BODY* 1-6 and 7-10 refer to the ten *body* samples, while columns *PROTRUSION* 1-6 and 7-10 refer to the ten *protrusion* ones.

4.2.4 Network modularity analysis

The network modularity analysis revealed that the apoptosis-related genes tended to form distinct clusters with clear functional characteristics. This observation is depicted in Figure 4.10, where the two different groups of apoptosis-related genes are illustrated.

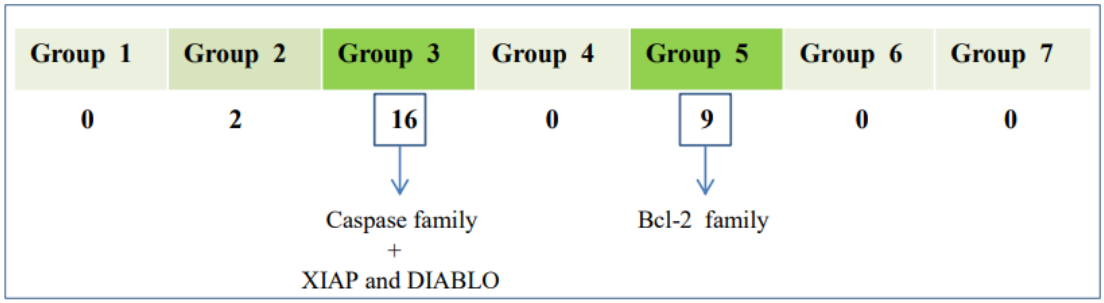

Fig. 4.10: Network modularization process through the Clauset-Newman-More algorithm with the number of components of the apoptosis group in each of the seven modules. Groups 3 and 5, highlighted in green, indicate the groups belonging to the same modules relevant to the cell apoptosis process (mainly Caspases and Bcl-2 families) [Pecorino, 2012].

The majority of apoptosis-related genes, approximately 92.5%, are found in Groups 3 and 5. Group 3 comprises the complete *Caspases*, *XIAP*, and *DIABLO* group, while Group 5 includes the entire Bcl-2 family. The modular distribution of these genes, as depicted in **Supplementary material S4**, demonstrates their tendency to cluster together. This characteristic is of great significance in the methodology employed in this study as it enables the utilization of a relatively small number of target vertices to activate these genes.

4.2.5 Shortest path evaluation

By examining the shortest path connecting the five bench targets with the apoptosis-related genes, we identified three genes present in at least one of these paths, namely HIF1A, XIAP, and BCL2. Since *XIAP* and *BCL2* are part of the apoptosis group, which serves as an indicator of the network's state, and they also serve as the final nodes in the shortest paths, we did not evaluate the effects of inhibiting these genes on the apoptosis attractor. However, since *HIF1A* is not a member of the apoptosis group, we replaced *XIAP* and *BCL2* with their respective input nodes, *STAT5A* and *BRCA1* (Figure 4.11).

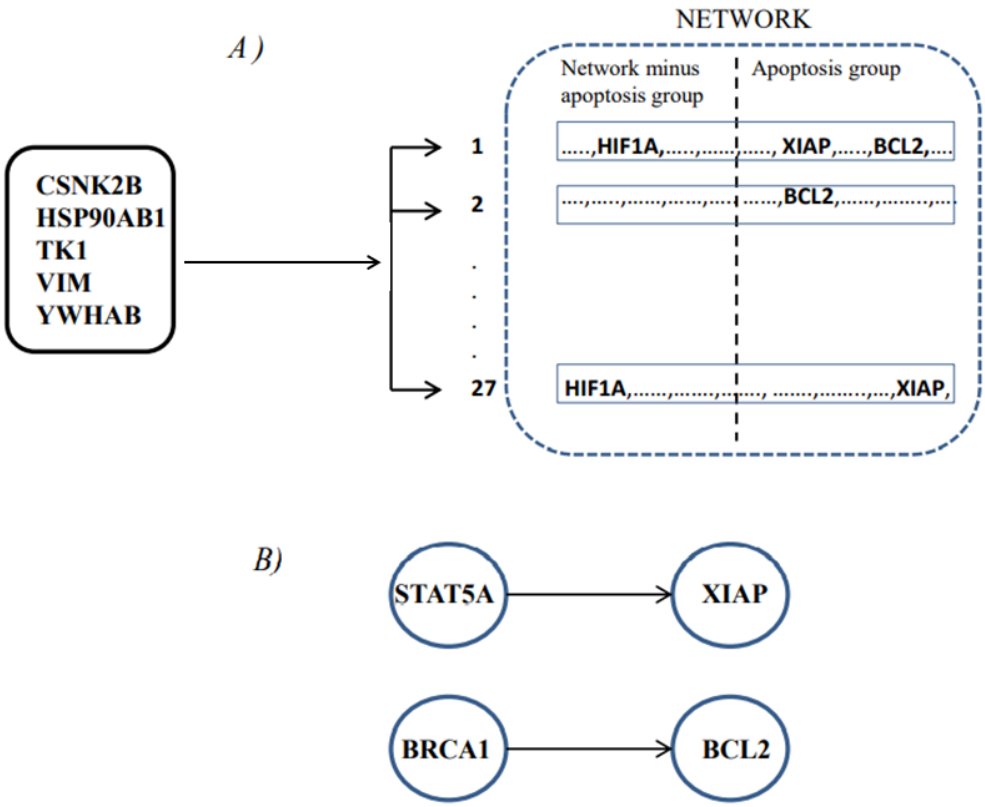

Fig. 4.11: Process of new target identification by the shortest search. **Panel A:** *HIF1A*, *XIAP*, and *BCL2* are present in at least one of the 27 shortest paths. From these three genes, only *HIF1A* did not belong to the apoptosis-related group**. Panel B:** *STA5A* and *BRCA1* represent the only input genes of *XIAP* and *BCL2*. Thus, *HIF1A*, *STAT5A*, and *BRCA1* were the optimized target nodes detected within the shortest paths between the five bench targets and the 27 apoptosis-related genes. These three vertices were excellent candidates to complete the new set of optimized target genes that trigger the network within the apoptosis state.

4.2.6 Optimizing the number of targets

Identifying the vertices with the highest centrality between the five bench targets and the 27 apoptosis-related genes allowed us to explore the impact of inhibiting the new targets on the apoptosis attractor and propose a novel approach for selecting therapeutic targets. By applying the algorithm outlined in Figure 4.5, we identified *STAT5A*, *BRCA1*, and *HIF1A* as highly central vertices. Hence, we targeted the inhibition towards these three genes instead of inhibiting CSNK2B, HSP90AB1, TK1,

VIM, and YWHAB. Combining these three genes successfully activated the apoptosis attractor in all sample comparisons. The outcomes achieved by inhibiting *HIF1A*, *STAT5A*, and *BRCA1* as substitutes for the five bench targets are given in Figure 4.12.

BODY		*PROTRUSION*	
	1-10		1-10
APAF1	F	APAF1	F
BAD	F	BAD	F
BAK	T	BAK	T
BAX	T	BAX	T
BCL2	F	BCL2	F
BCL2A1	F	BCL2A1	F
BCL2L1	F	BCL2L1	F
BCL2L11	T	BCL2L11	T
BID	T	BID	T
BIRC2	F	BIRC2	F
BIRC5	F	BIRC5	F
CASP10	T	CASP10	T
CASP2	T	CASP2	T
CASP3	T	CASP3	T
CASP6	T	CASP6	T
CASP7	T	CASP7	T
CASP8	T	CASP8	T
CASP9	T	CASP9	T
CYC	F	CYC	F
DIABLO	T	DIABLO	T
FADD	T	FADD	T
MP3K5	F	MP3K5	F
MCL1	F	MCL1	F
PMAIP1	T	PMAIP1	T
TRADD	F	TRADD	F
XIAP	F	XIAP	F
DAXX	F	DAXX	F

Fig. 4.12: Activation or inhibition status detected in the 27 genes constituting the apoptosis-related group in the attractors of 10 *body* (1-10) and ten *protrusion* (1-10) sample comparisons by inhibiting *HIF1A*, *STAT5A*, and *BRCA1* instead of *CSNK2B*, *HSP90AB1*, *TK1*, *VIM*, and *YWHAB*.

In Figure 4.12, we demonstrated the simultaneous induction of cell apoptosis in both *body* and *protrusion* types for the *Bcl-2* and *Caspase* gene families across all combinations of RNA-seq data. Notably, *BID*, a pro-apoptosis member of the *Bcl-2* family, is activated in this simulation. The activation of BID plays a crucial role in activating downstream Caspases by directly activating *BAX* and *BAK*. This activation is absent in simulating the apoptosis attractor using the five bench genes (Figure 4.8).

Similar considerations apply to *DIABLO*, which exhibited inconsistent expression in the six RNA-seq protrusion combinations from 1 to 6 (Figure 4.8).

Consequently, we simulated the induction of the network into the apoptosis attractor using various gene inhibition strategies. By inhibiting the five bench targets, we achieved a configuration conducive to apoptosis in a significant portion of the genes within this group (shown in green on the graph) (Figure 4.13). The percentage of genes activated in the apoptosis-related group was notably higher (74.1%) compared to scenarios without bench target inhibition (first column). When *HIF1A*, *STAT5A*, and *BRCA1* were selected as targets for inhibition, the proportion of genes activated in the apoptosis group further increased (77.8%) and remained consistent regardless of the MDA-MB-231 fraction or the MCF10A RNA-seq data used to identify the up-regulated genes in MDA-MB-231 (Figure 4.13, right column).

Based on the findings above, it might be inferred that the transition from sample-specific malignant basins of attraction in *body* and *protrusion* samples, respectively, occurred towards a unified basin of attraction that signified a generalized state of cellular apoptosis, which was achieved by inhibiting *HIF1A*, *STAT5A*, and *BRCA1* (Figure 4.14).

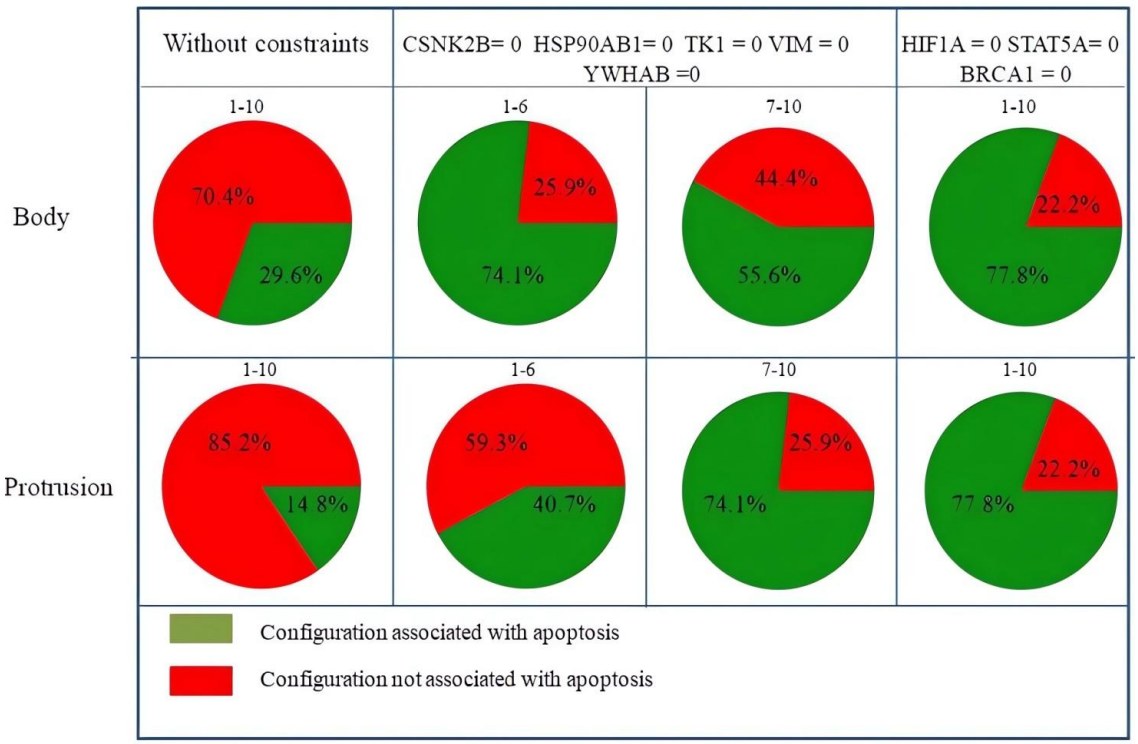

Fig. 4.13: The results detected on the *body* and *protrusion* fractions of MDA-MB-231 used in this report, respectively. The columns identify which network genes were kept silenced in the dynamic simulation of the model. The green color, as opposed to the red one, indicates the percentage of the genes of the apoptosis group presenting a stable configuration of activation or inhibition associated with the attractor of the apoptosis phenotype.

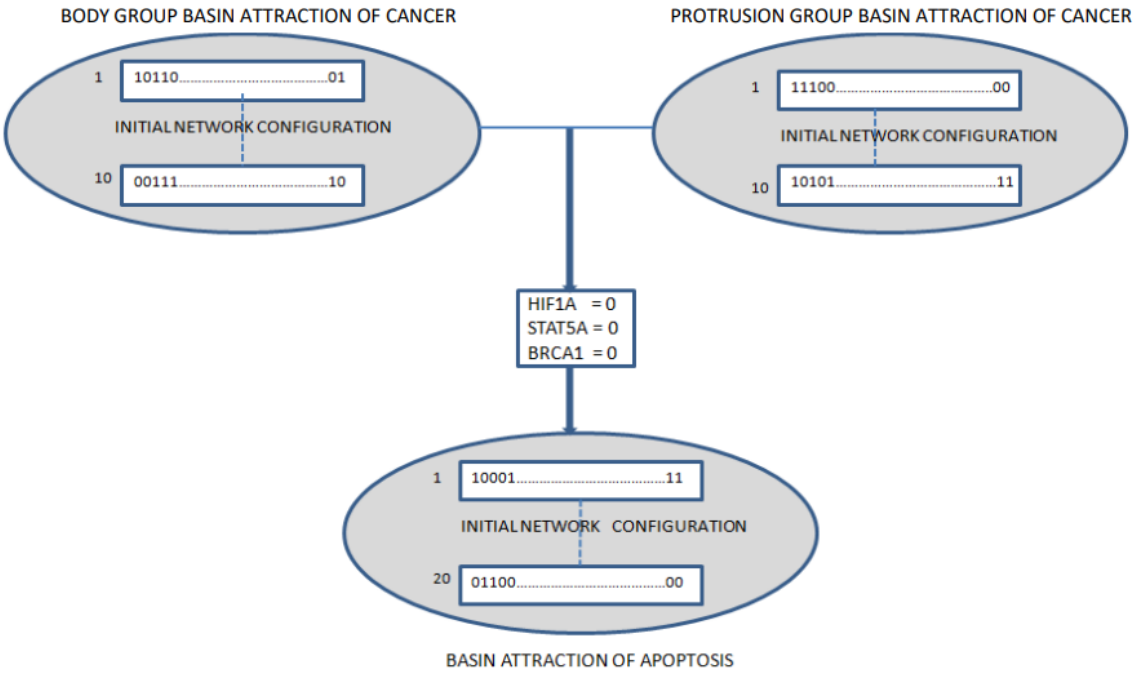

Fig. 4.14: Boolean description of the transition from two basins of attraction representing the malignant cellular state of *body* and *protrusion* to a single basin of attraction of the cellular apoptosis state, obtained by inhibiting *HIF1A*, *STAT5A*, and *BRCA1*.

4.3 Discussion

As a multifaceted disease, cancer is influenced by numerous factors that cannot be comprehensively understood solely through molecular analysis. Consequently, there is a growing inclination to integrate molecular data with the dynamic characteristics of biological networks, employing computational and mathematical modeling techniques to gain deeper insights into the underlying biological mechanisms driving its progression [Kitano, 2002].

The choice between quantitative and qualitative modeling approaches depends on the nature of the available data. Quantitative modeling, which involves ordinary differential equations and requires kinetic parameters, becomes challenging and feasible only for gene regulatory networks of limited scale [Karlebach and Shamir, 2008]. In contrast, qualitative Boolean network modeling provides a viable alternative, allowing for relatively straightforward dynamic simulation of complex biological systems [Schwab et al., 2020]. This approach proves beneficial in exploring regulatory interactions in protein expression [Dahlhaus et al., 2016] and developing strategies for therapeutic interventions [Bloomingdale et al., 2018]. It is also important to note that since the model proposed in this work is a Boolean-type model, we implicitly assume that the values of the system components are binary and Boolean functions govern their interaction. Such a description of the system dynamics, termed qualitative, necessarily implies a loss of the functional detail of the system that a quantitative methodology can provide instead. In addition, having chosen a synchronous rather than asynchronous network update system, giving preference to the deterministic nature of interactions an easy interpretation of results, a rough approximation was accepted in the timing mechanisms of the system elements, at the expense of the stochastic nature of these interactions.

In addition to the documented characteristics, we conducted functional compatibility checks to validate the Boolean model used in this study against the results obtained from an *in vitro* experiment [Tilli et al., 2016] involving silencing five genes using siRNA. This experiment induced apoptosis in the MDA-MB-231 cell line. Our data show that our system can generate functionally compatible outcomes by inhibiting the same genes as in the *in vitro* experiment. Thus, we successfully replicated the behavior of an actual biological system within the Boolean dynamics of the gene regulatory network implemented in our research.

In this study, we utilized a Boolean network that represents a set of up-regulated genes in breast cancer to identify attractors corresponding to specific cellular phenotypes. We further assessed the compatibility of the Boolean network with an existing biological system by comparing it with an *in vitro* experiment [Tilli et al., 2016]. The assignment of Boolean values to the network nodes followed the algorithm depicted in Figure 4.2. This algorithm facilitated the Booleanization of RNA-seq values based on the gene expression variations between malignant and non-malignant cell lines. By considering the gene expression differences across different cell lines, we were

able to perform dynamic simulations of our model on a substantial number of comparisons, yielding valuable insights. Indeed, rather than solely relying on a single control for the four malignant samples (two *body* and two *protrusion*) to Booleanize the RNA-seq values and identify attractors, we expanded our approach by incorporating five samples of the non-malignant cell line MCF10A. By incorporating these additional samples, we derived gene expression differences that enabled us to perform a more comprehensive analysis. This strategy resulted in twenty combinatorial comparisons, significantly enhancing the numerical significance of network configurations and enabling the application of the procedures outlined in this study.

Considering cancer phenotype as basins of attraction in the epigenetic landscape [Crespo et al., 2013], this report aimed to cause the transition from a basin of attraction of malignant type to that of apoptosis [Shi et al., 2010] through the perturbation of a subset of genes belonging to the network. For this purpose, we developed an algorithm (Fig. 5) that optimizes the choice of the network elements to produce a transition from one specific phenotype to another.

The approach of exploring the relationship existing between the results of an *in vitro* experiment, the insertion of a specific group of genes for apoptosis into the system, and the investigation of the network structure through the analysis of shortest paths between the five bench targets and the apoptosis-related group, represents an innovative strategy for clinical applications to increase patient benefit in personalized approaches of cancer therapies.

Including the apoptosis-related gene group within the network was a reference to evaluate the induction of cell death attractors through vertex inhibition. The results obtained in this study, depicted in Figure 4.13, schematically illustrate the effectiveness of this methodology. By inhibiting the three genes *HIF1A*, *STAT5A*, and *BRCA1*, we observed a transition in the system dynamics from malignant basins of attraction to those associated with cell apoptosis across all analyzed samples. The comparison of this result (Figure 4.12) with that obtained by reproducing the *in vitro* experiment (Figure 4.9), shows the robustness of the data obtained by applying the procedure of Figure 4.5. Inhibiting the five genes (*CSNK2B*, *HSP90AB1*, *TK1*, *VIM*, *YWHAB*) described in Tilli et al. [Tilli et al., 2016] promotes a configuration of the 23 apoptosis-related genes conducive to apoptosis. However, the quantitative uniformity of this configuration varies among different comparisons. In the *body* sample, lines 1 to 6 exhibited a greater inclination towards apoptosis, whereas lines 7 to 10 in the *protrusion* sample displayed

a higher favorability towards apoptosis (Figure 4.9). Conversely, when inhibiting the three genes (*HIF1A*, *STAT5A*, *BRCA1*) as detected through the procedure outlined in Figure 4.5, a more consistent profile of activated apoptosis-related genes was observed across all comparisons (Figure 4.12). Despite the divergent attractors between the body and protrusion samples, the induction of apoptosis by inhibiting *HIF1A*, *STAT5A*, and *BRCA1* underscores the method's robustness.

According to Figure 4.11, the target genes identified through the procedure outlined in Figure 4.5 should ideally be *HIF1A*, *XIAP*, and *BCL2*, considering our objective of identifying target genes capable of activating or inhibiting the 27 apoptosis-related genes, regardless of their specific configuration. Consequently, we decided to avoid designating *XIAP* and *BCL2* as targets. One possible alternative was substituting them with their respective input nodes, *STAT5A* and *BRCA1*, which do not belong to the apoptosis group. Therefore, it cannot be ruled out that the combined inhibitory effect on the *HIF1A*, *XIAP*, and *BCL2* genes may significantly induce apoptosis in cancer cells.

The role of *HIF1A*, *STAT5*, and *BRCA1* is well documented in tumors. *HIF1A* encodes the HIF-1α protein, whose level is regulated by hypoxia and other mechanisms, and is part of the heterodimeric transcription factor HIF-1. HIF-1α has crucial roles in many tumorigenic processes, such as epithelial-mesenchymal transition (EMT), metastasis, cancer cell metabolism, and angiogenesis [Yang et al., 2008; Wang et al., 2021; Sharma et al., 2002]. Interestingly, there is a crosstalk between the HIF-1 and p53 pathways to determine cell fate depending on hypoxic conditions [Zhou et al., 2015; Wang et al., 2019]. Therefore, targeting the HIF-1 signaling in cancer can be a promising therapeutic strategy [Sharma et al., 2002].

The transcriptional factor STAT5 is a member of the JAK-STAT (Janus kinase/Signal transducer and activator of transcription) pathway, which is altered in many tumors. Activated STAT5 upregulates the expression of genes involved in cell proliferation, invasion, angiogenesis, and the inhibition of apoptosis [Halim et a., 2020]. The exact role of STAT5 in breast cancer is still under debate. The STAT5 activation in tumor macrophages by derived factors from breast cancer cells led to the expression of anti-tumor immune stimulatory genes [Jesser et al., 2021]. On the other hand, it was shown that STAT5a, an isoform of STAT5, could confer resistance to doxorubicin [Li et al., 2021] and combined PI3K/mTOR and JAK2/STAT5 pathways inhibition induced cell death in triple-negative breast cancer [Britschgi et al., 2012].

BRCA1 is an essential gene in DNA repair and cell cycle regulation. When mutated, the risk of developing many cancers significantly increases, especially for breast and ovarian tumors [Fu et al., 2022]. Several studies have shown increased brain metastasis frequency in patients carrying *BRCA1* mutations [Ratner et al., 2019; Zavitsanos et al., 2018]. Another fundamental role of this gene is the maintenance of genomic stability [Roy et al., 2012]. Therefore, *BRCA1* is essential to tissue homeostasis.

More specifically, several reports have established the relationship between the inhibition of *HIF1A*, *STAT5A*, and *BRCA1* genes and the induction of apoptosis in the MDA-MB-231 cell line. In the case of *HIF1A*, suppressing its expression using siRNA has been shown to inhibit cell growth and enhance apoptosis [Zeng et al., 2014]. Inhibition of *STAT5A* has been correlated with reduced metastasis and growth of breast cancer tumor cells [Medler et al., 2016]. Additionally, the knockdown of *STAT5A* restores cellular sensitivity to TRAIL-induced apoptosis [Yoshida et al., 2009]. As for *BRCA1*, its RNAi-mediated silencing, along with miR-342 transfection, has been found to increase the percentage of apoptotic cells [Crippa et al., 2016]. Furthermore, *BRCA1*-depleted MDA-MD-231 cells exhibited heightened susceptibility to proteasome inhibitors [Gu et al., 2014]. Considering the known functions and the consequences of the deregulation of these three genes in cell homeostasis, our study underscores the impact of inhibiting them on promoting apoptosis induction in the MDA-MB-231 cell line. It is important to note the interplay between HIF-1 and p53 pathways to determine cell death under hypoxic conditions [Zhou et al., 2015, Wang et al., 2019]. The inhibition of HIF1A could favor p53 in its apoptotic roles. However, in our model, p53 was inactivated since it is mutated in this cell line and not working as a tumor suppressor (Hui et al., 2006). Therefore, other mechanisms need to be investigated more deeply in the future.

The outcomes presented in this study hinged on the fine-tuning of the transfer functions (eqs. 6-10) to align the model with the *in vitro* experiment [Tilli et al., 2016]. However, the concurrence observed between the *in vitro* results, and the computational simulation indicated a satisfactory level of model representativeness, warranting its potential for future optimization and application in therapeutic scenarios. Consequently, it becomes feasible to integrate specific experimental findings with computational hypotheses formulated to tackle therapeutic challenges associated with cancer.

It is worth noting that the identified therapeutic targets are the results obtained by running the algorithm presented in Figure 4.5 on the boolean network model constructed and validated by our group. The results obtained through their inhibition show that their choice is necessary and sufficient to achieve optimization in qualitative terms of the performance obtained from the in vitro experiment taken as a reference in this report. All this does not exclude the possibility of not having considered other important therapeutic targets that have emerged in other contexts.

The main objective of our research was to identify therapeutic targets on which an inhibition action is capable of causing a change in the state trajectory of the system, consequently producing a change in the system's final target attractor.

However, the use of Boolean gene regulatory networks in some research areas, such as pharmacogenetics [Hemedan et al., 2022], can be challenging in identifying the complicated mechanisms between the genome, its products (RNAs and proteins), and the cellular-level response to drugs.

Because of the complex interactions that exist among molecules involved in a carcinogenic process, a perturbation analysis method such as the one we used in our research can be useful in dealing with such complexity, proposing specific interventions on the system by guiding and facilitating the subsequent choice of therapy useful for the purpose. Indeed, once therapeutic targets have been identified, there is the possibility of pharmacologically acting on them directly or through signaling pathways in which they are involved, through drugs currently in use.

Another therapeutic possibility of greater complexity is using siRNA molecole encapsulated in nanoparticles specific to the identified targets.

The outcomes presented in this study are derived from the analysis of data obtained from specific biological samples. The growing abundance of such information on distinct pathological conditions of cancer highlights the versatility of our model in accommodating various configurations of the same disease. The positioning of the method developed in this study within personalized medicine reflects its capacity to address individualized approaches to cancer treatment.

4.4 Chapter conclusion

In this research, we implemented a new computational method for optimizing the number of potential targets for breast cancer. We constructed a Boolean Gene Regulatory Network Model of a breast cancer tumor and validated it using RNA-seq data from tumoral cell lines. We achieved these results by integrating experimental data with those obtained from an extensive literature search in Boolean gene regulatory networks, for which the analysis of the corresponding attractors allowed the identification of potential therapeutic targets. In future work, we intend to apply our method to actual patient data to validate our results in the context of personalized medicine.

CHAPTER 5

Discussion

The results shown in the specific sections of the thesis demonstrate how the goals pursued in the introductory section were achieved. Potential therapeutic targets for breast cancer were identified using the implemented model, whose biological compatibility was also verified by emulating an in vitro laboratory activity.

The various steps performed to achieve these results offer several interpretive insights into the methodology adopted.

The different types of data used in the two stages of the research, scRNA-seq in the first stage and RNA-seq values from in vitro cell lines in the second stage, and the different data binarization methodology adopted between the two different distinct stages described in the thesis, did not prevent the production of biologically relevant results in the various stages of the research analyzed.

The procedure for constructing the gene regulatory network related to breast cancer, built from scratch in the first stage of the research and used as a starting point in the second stage, is the result of intensive research into the functional relationships existing between the various genes comprising the system, culminating in the use of software capable of providing in detail the bibliographic justifications for the links between the individual genes detected. The size of the network obtained justifies the accuracy of the search carried out in choosing the elements of which it is composed while at the same time allowing an accurate analysis of the problem under investigation. In the proposed method with the network that was implemented, and through Boolean formalism, we investigated the epigenetic landscape representing the cell under certain conditions. We provided a way to identify the critical elements of the system to drive the cell state within it.

However, it is important to consider that there are still many genes and proteins whose functions, interactions, and regulatory logic have not been discovered yet, with the consequence that the implemented network model is not complete and contains many uncertainties that will be filled with the advancement of research in this specific field. Using the Boolean network allowed the dynamic analysis of a gene regulation network with a large number of nodes. This task would have been difficult o, through a

quantitative analysis, for example, by using differential equations. However, this advantage must be contrasted with an approximation in the description of the analyzed biological phenomenon that Boolean networks provide. Describing a gene's activity level through two states, active or inhibited, provides an extreme approximation of biological reality and lends itself to a potential excessive simplification of the analyzed phenomenon.

The success achieved using Boolean networks in this research illustrates how the organization of network structure played a more important role than the kinetic details of the individual interaction [Albert and Thakar, 2014]. This suggests the hypothesis that the model implemented in this research can be used as a foundation of regulatory models where more detailed continuous models can be built as kinetic information of quantitative experimental data becomes available.

A further point of reflection is the updating method adopted in the dynamic analysis of the network. Synchronous updating, in which all network elements are updated simultaneously, although it provides an advantage from a computational point of view, represents a simplification of what happens in biological reality [Schwab et al., 2020]. For this reason, an asynchronous network updating method could represent a valid alternative to investigate.

It is also interesting to reflect on the adequacy of nested canalizing functions adopted on the network nodes. The advantages of using this type of function are well documented [Zhou et al., 2013], along with the benefit of applying the same function calculation rule to all nodes in the network. However, precisely this generalization in use represents a limit in describing the complexity and variety of the system, suggesting further study of the method adopted in assigning Boolean functions to the individual nodes of the network.

The results obtained in this research show the capability of the method used in the silicon model and the dynamics of the analyzed cell. Effective control of biological systems can be achieved by controlling a limited number of distinct variables [Borriello and Daniels, 2021]. Considering the limited number of nodes in the network on which an inhibitory action was acted upon in the silicon model producing through the formalism of Boolean dynamical networks a direction of cancer cell fate, it can be said that the method described in this thesis confirmed this claim.

However, it is important to note that the control obtained over the system is also due to the direct intervention of modifying nested canalizing functions on some key

network vertices. While this type of intervention may be relatively easy in not very large networks, it may not represent an easy application in Boolean networks of considerable size, suggesting further study in this regard, for example, by implementing automation techniques that would allow the easy adaptation of the type of function applied to the nodes according to the type of attractor to be reached.

The use of increasingly complex and specific gene expression data for certain pathological situations represents a potential motivation for further and more in-depth development of the method implemented in this research toward use in the context of precision medicine.

CHAPTER 6

Conclusion

Changing cell fate from its natural course represents a challenge of considerable complexity and involves several interesting implications, especially in the therapeutic field.

The research pathway described in this dissertation represents a potentially valuable contribution to the therapeutic approach for breast cancer.

The interdisciplinary approach described in this dissertation resulted in a biological model closely mimicking an actual experimental situation.

Using modeling techniques already frequently used in systems biology, such as Boolean gene regulatory network modeling, together with original procedural methodologies, we identified specific genes capable of modifying the fate of a cancer cell line.

The use of the Boolean formalism allowed the analysis of the system dynamics of a network with a significant number of nodes despite the approximation that it entails in the description of biological phenomena. This exercise made a computationally efficient approach to network dynamic analysis possible despite the large number of nodes involved. Analysis of network dynamics on systems like the one implemented in our research with differential equations is undoubtedly more complex to model due to the need to infer the required parameters. This point is of fundamental importance, considering that incorporating the network with a significant number of specific genes for a given cell phenotype represents a key point of the methodology implemented. Moreover, using results obtained in the laboratory as a reference allowed the validation of our approach, providing a procedural paradigm applicable in other similar contexts.

In addition, the use of laboratory results obtained in the laboratory to validate the implemented model and as reference for an optimization of the results through the procedure we adopted could be used in contexts similar to the one described in this thesis.

REFERENCE

Albert, István, et al. "Boolean network simulations for life scientists." *Source code for biology and medicine* 3 (2008): 1-8.

Albert, Réka. "Scale-free networks in cell biology." *Journal of cell science* 118.21 (2005): 4947-4957.

Albert, Reka, and Juilee Thakar. "Boolean modeling: a logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions." *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 6.5 (2014): 353-369.

Akram, Muhammad, et al. "Awareness and current knowledge of breast cancer." *Biological research* 50 (2017): 1-23.

Aubrey, Brandon J., Andreas Strasser, and Gemma L. Kelly. "Tumor-suppressor functions of the TP53 pathway." *Cold Spring Harbor perspectives in medicine* 6.5 (2016): a026062.

Baedke, Jan. "The epigenetic landscape in the course of time: Conrad Hal Waddington's methodological impact on the life sciences." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44.4 (2013): 756-773

Barillot, Emmanuel, et al. *Computational systems biology of cancer*. Boca Raton, FL: CRC Press, 2013.

Barbuti, Roberto, et al. "A survey of gene regulatory networks modelling methods: from differential equations, to Boolean and qualitative bioinspired models." *Journal of Membrane Computing* 2.3 (2020): 207-226.

Belfiore, Antonino, et al. "A novel functional crosstalk between DDR1 and the IGF axis and its relevance for breast cancer." *Cell adhesion & migration* 12.4 (2018): 305-314.

Bhargava, Rohit, et al. "EGFR gene amplification in breast cancer: correlation with epidermal growth factor receptor mRNA and protein expression and HER-2 status and absence of EGFR-activating mutations." *Modern pathology* 18.8 (2005): 1027-1033.

Bloomingdale, Peter, et al. "Boolean network modeling in systems pharmacology." *Journal of pharmacokinetics and pharmacodynamics* 45 (2018): 159-180.

Borriello, Enrico, and Bryan C. Daniels. "The basis of easy controllability in Boolean networks." *Nature communications* 12.1 (2021): 5227.

Britschgi, Adrian, et al. "JAK2/STAT5 inhibition circumvents resistance to PI3K/mTOR blockade: a rationale for cotargeting these pathways in metastatic breast cancer." *Cancer cell* 22.6 (2012): 796-811.

Byrne, J. A., et al. "Producing primate embryonic stem cells by somatic cell nuclear transfer." *Nature* 450.7169 (2007): 497-502.

Cao, Jiguo, Xin Qi, and Hongyu Zhao. "Modeling gene regulation networks using ordinary differential equations." *Next generation microarray bioinformatics: methods and protocols* (2012): 185-197.

Carels, Nicolas, Tatiana Tilli, and Jack A. Tuszynski. "A computational strategy to select optimized protein targets for drug development toward the control of cancer diseases." *PLoS one* 10.1 (2015): e0115054.

Chang, Rui, Robert Shoemaker, and Wei Wang. "Systematic search for recipes to generate induced pluripotent stem cells." *PLoS computational biology* 7.12 (2011): e1002300.

Chaffey, Nigel. "Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. Molecular biology of the cell. 4th edn." (2003): 401-401.

Chénais, Benoît, et al. "Transcriptomic response of breast cancer cells MDA-MB-231 to docosahexaenoic acid: Downregulation of lipid and cholesterol metabolism genes and upregulation of genes of the pro-apoptotic ER-stress pathway." *International Journal of Environmental Research and Public Health* 17.10 (2020): 3746.

Cheng, Jin-Shiung, et al. "The MAP3K7-mTOR axis promotes the proliferation and malignancy of hepatocellular carcinoma cells." *Frontiers in Oncology* 9 (2019): 474.

Cho, Kwang-Hyun, et al. "Cancer reversion, a renewed challenge in systems biology." *Current Opinion in Systems Biology* 2 (2017): 49-58.

Cho, Sung-Hwan, et al. "Attractor landscape analysis of colorectal tumorigenesis and its reversion." *BMC systems biology* 10 (2016): 1-13.

Chung, Woosung, et al. "Single-cell RNA-seq enables comprehensive tumor and immune cell profiling in primary breast cancer." *Nature communications* 8.1 (2017): 15081.

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

Cornelius, Sean P., William L. Kath, and Adilson E. Motter. "Realistic control of network dynamics." *Nature communications* 4.1 (2013): 1942.

Creixell, Pau, et al. "Navigating cancer network attractors for tumor-specific therapy." *Nature biotechnology* 30.9 (2012): 842-848.

Crespo, Isaac, and Antonio Del Sol. "A general strategy for cellular reprogramming: the importance of transcription factor cross-repression." *Stem Cells* 31.10 (2013): 2127-2135.

Crespo, Isaac, et al. "Predicting missing expression values in gene regulatory networks using a discrete logic modeling optimization guided by network stable states." *Nucleic acids research* 41.1 (2013): e8-e8.

Crespo, Isaac, et al. "Detecting cellular reprogramming determinants by differential stability analysis of gene regulatory networks." *BMC systems biology* 7 (2013): 1-14.

Crippa, E. "miR-342 overexpression results in a synthetic lethal phenotype in BRCA1-mutant HCC1937 breast cancer cells. Oncotarget 7, 18594–18604." (2016).

Dahlhaus, Meike, et al. "Boolean modeling identifies Greatwall/MASTL as an important regulator in the AURKA network of neuroblastoma." *Cancer letters* 371.1 (2016): 79-89.

Davidson, Eric, and Michael Levin. "Gene regulatory networks." *Proceedings of the National Academy of Sciences* 102.14 (2005): 4935-4935.

Del Sol, Antonio, and Noel J. Buckley. "Concise review: A population shift view of cellular reprogramming." *Stem Cells* 32.6 (2014): 1367-1372.

Dezonne, R. S., et al. "Derivation of functional human astrocytes from cerebral organoids. Sci Rep 7: 45091." (2017).

Ding, Shengchao, and Wei Wang. "Recipes and mechanisms of cellular reprogramming: a case study on budding yeast Saccharomyces cerevisiae." *BMC systems biology* 5 (2011): 1-14.

D'Urso, Agustina, and Jason H. Brickner. "Mechanisms of epigenetic memory." *Trends in genetics* 30.6 (2014): 230-236.

Dzutsev, Amiran, et al. "Microbes and cancer." *Annual review of immunology* 35 (2017): 199-228.

Ekins, Sean, et al. "Pathway mapping tools for analysis of high content data." *High content screening: A powerful approach to systems cell biology and drug discovery* (2006): 319-350.

Emmert-Streib, Frank, Matthias Dehmer, and Benjamin Haibe-Kains. "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks." *Frontiers in cell and developmental biology* 2 (2014): 38.

Feng, Yixiao, et al. "Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis." *Genes & diseases* 5.2 (2018): 77-106.

Forrest ARR, Kawaji H, Rehli M, Kenneth Baillie J, de Hoon MJL, Haberle V, et al. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462–70.

Franceschini, Andrea, et al. "STRING v9. 1: protein-protein interaction networks, with increased coverage and integration." *Nucleic acids research* 41.D1 (2012): D808-D815.

Friedman, Nir, et al. "Using Bayesian networks to analyze expression data." *Proceedings of the fourth annual international conference on Computational molecular biology*. 2000.

Fu, Xiaoyu, et al. "BRCA1 and breast cancer: molecular mechanisms and therapeutic strategies." *Frontiers in cell and developmental biology* 10 (2022): 813457.

George, Richard A., et al. "Analysis of protein sequence and interaction data for candidate disease gene prediction." *Nucleic acids research* 34.19 (2006): e130-e130.

Gibbs, Lee D., and Jamboor K. Vishwanatha. "Prognostic impact of AnxA1 and AnxA2 gene expression in triple-negative breast cancer." *Oncotarget* 9.2 (2018): 2697.

Goldberg, Aaron D., C. David Allis, and Emily Bernstein. "Epigenetics: a landscape takes shape." *Cell* 128.4 (2007): 635-638.

Gomez, Henry L., et al. "Efficacy and safety of lapatinib as first-line therapy for ErbB2-amplified locally advanced or metastatic breast cancer." *Journal of Clinical Oncology* 26.18 (2008): 2999-3005.

Gu, Yuexi, et al. "Suppression of BRCA1 sensitizes cells to proteasome inhibitors." *Cell death & disease* 5.12 (2014): e1580-e1580.

Hagberg, Aric, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX*. No. LA-UR-08-05495; LA-UR-08-5495. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

Halim, Clarissa Esmeralda, et al. "Involvement of STAT5 in Oncogenesis." *Biomedicines* 8.9 (2020): 316.

Halley-Stott, Richard P., Vincent Pasque, and J. B. Gurdon. "Nuclear reprogramming." *Development* 140.12 (2013): 2468-2471.

Hamazaki, Takashi, et al. "Concise review: induced pluripotent stem cell research in the era of precision medicine." *Stem Cells* 35.3 (2017): 545-550.

Han, Heonjong, et al. "TRRUST: a reference database of human transcriptional regulatory interactions." *Scientific reports* 5.1 (2015): 11432.

Hanahan, Douglas, and Robert A. Weinberg. "Hallmarks of cancer: the next generation." *cell* 144.5 (2011): 646-674.

Hanahan, Douglas. "Hallmarks of cancer: new dimensions." *Cancer discovery* 12.1 (2022): 31-46.

Harris, Stephen E., et al. "A model of transcriptional regulatory networks based on biases in the observed regulation rules." *Complexity* 7.4 (2002): 23-40.

Hemedan, Ahmed Abdelmonem, et al. "Boolean modelling as a logic-based dynamic approach in systems medicine." *Computational and Structural Biotechnology Journal* 20 (2022): 3161-3172.

Hemmi, Jacob J., Anuja Mishra, and Peter J. Hornsby. "Overcoming barriers to reprogramming and differentiation in nonhuman primate induced pluripotent stem cells." *Primate Biology* 4.2 (2017): 153-162.

Herrmann, Franziska, et al. "A boolean model of the cardiac gene regulatory network determining first and second heart field identity." (2012): e46798.

Hinkelmann, Franziska, and Abdul Salam Jarrah. "Inferring biologically relevant models: nested canalyzing functions." *International Scholarly Research Notices* 2012 (2012).

Hong, Changki, et al. "An efficient steady-state analysis method for large boolean networks with high maximum node connectivity." *PloS one* 10.12 (2015): e0145734.

Hopfensitz, Martin, et al. "Multiscale binarization of gene expression data for reconstructing Boolean networks." *IEEE/ACM transactions on computational biology and bioinformatics* 9.2 (2011): 487-498.

Hou, Pingping, et al. "Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds." *Science* 341.6146 (2013): 651-654.

Huang, Sui, et al. "Cell fates as high-dimensional attractor states of a complex gene regulatory network." *Physical review letters* 94.12 (2005): 128701.

Huang, Sui, Ingemar Ernberg, and Stuart Kauffman. "Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective." *Seminars in cell & developmental biology*. Vol. 20. No. 7. Academic Press, 2009.

Hui, L., et al. "Mutant p53 in MDA-MB-231 breast cancer cells is stabilized by elevated phospholipase D activity and contributes to survival signals generated by phospholipase D." *Oncogene* 25.55 (2006): 7305-7310.

Hwang, Byungjin, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines." *Experimental & molecular medicine* 50.8 (2018): 1-14.

Jančík, Sylwia, et al. "Clinical relevance of KRAS in human cancers." *BioMed Research International* 2010 (2010).

Jesser, Emily A., et al. "STAT5 is activated in macrophages by breast cancer cell-derived factors and regulates macrophage function in the tumor microenvironment." *Breast Cancer Research* 23 (2021): 1-17.

Jonsson, Pall F., and Paul A. Bates. "Global topological features of cancer proteins in the human interactome." *Bioinformatics* 22.18 (2006): 2291-2297.

Karin, M. "Nuclear factor-kappaB in cancer development and progression. Nature [Internet]. 2006 May 25; 441 (7092): 431–6."

Karlebach, Guy, and Ron Shamir. "Modelling and analysis of gene regulatory networks." *Nature reviews Molecular cell biology* 9.10 (2008): 770-780.

Kauffman, Stuart A. *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA, 1993. pp. 447-449

Kauffman, Stuart, et al. "Genetic networks with canalyzing Boolean rules are always stable." *Proceedings of the National Academy of Sciences* 101.49 (2004): 17102-17107.

Kauffman, Stuart, et al. "Random Boolean network models and the yeast transcriptional network." *Proceedings of the National Academy of Sciences* 100.25 (2003): 14796-14799.

Kauffman, Stuart. "Homeostasis and differentiation in random genetic control networks." *Nature* 224.5215 (1969): 177-178.

Kauppinen, Jaana M., et al. "ST14 gene variant and decreased matriptase protein expression predict poor breast cancer survival." *Cancer epidemiology, biomarkers & prevention* 19.9 (2010): 2133-2142.

Kawser Hossain, Mohammed, et al. "Recent advances in disease modeling and drug discovery for diabetes mellitus using induced pluripotent stem cells." *International Journal of Molecular Sciences* 17.2 (2016): 256.

Kitano, Hiroaki. "Systems biology: a brief overview." *science* 295.5560 (2002): 1662-1664.

Lang, Alex H., et al. "Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes." *PLoS computational biology* 10.8 (2014): e1003734.

Lerebours, Florence, et al. "NF-kappa B genes have a major role in inflammatory breast cancer." *BMC cancer* 8 (2008): 1-11.

Li, Dan, et al. "Oxytocin receptor induces mammary tumorigenesis through prolactin/p-STAT5 pathway." *Cell Death & Disease* 12.6 (2021): 588.

Liberzon, Arthur, et al. "The molecular signatures database hallmark gene set collection." *Cell systems* 1.6 (2015): 417-425.

Mall, Moritz, and Marius Wernig. "The novel tool of cell reprogramming for applications in molecular medicine." *Journal of Molecular Medicine* 95 (2017): 695-703.

Martin, Gail R. "Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells." *Proceedings of the National Academy of Sciences* 78.12 (1981): 7634-7638.

Mason, Chris, and Peter Dunnill. "A brief definition of regenerative medicine." (2008): 1-5.

Matharu, Navneet, et al. "CRISPR-mediated activation of a promoter or enhancer rescues obesity caused by haploinsufficiency." *Science* 363.6424 (2019): eaau0629.

Medler, Terry R., et al. "HDAC6 deacetylates HMGN2 to regulate Stat5a activity and breast cancer growth." *Molecular Cancer Research* 14.10 (2016): 994-1008.

Mitra, Mithun K., et al. "Delayed self-regulation and time-dependent chemical drive leads to novel states in epigenetic landscapes." *Journal of The Royal Society Interface* 11.100 (2014): 20140706.

Mori, Tomoya, and Tatsuya Akutsu. "Attractor detection and enumeration algorithms for Boolean networks." *Computational and Structural Biotechnology Journal* 20 (2022): 2512-2520.

Müssel, Christoph, et al. "BiTrinA—multiscale binarization and trinarization with quality analysis." *Bioinformatics* 32.3 (2016): 465-468.

Nasti, Lucia. *Verification of robustness property in chemical reaction networks*. Diss. University of Pisa, Italy, 2020.

Nikolajewa, Swetlana, Maik Friedel, and Thomas Wilhelm. "Boolean networks with biologically relevant rules show ordered behavior." *Biosystems* 90.1 (2007): 40-47.

Nykter, Matti, et al. "Critical networks exhibit maximal information diversity in structure-dynamics relationships." *Physical review letters* 100.5 (2008): 058702.

Okita, Keisuke, et al. "Generation of mouse induced pluripotent stem cells without viral vectors." *Science* 322.5903 (2008): 949-953.

Ouyang, Mao, et al. "COP1, the negative regulator of ETV1, influences prognosis in triple-negative breast cancer." *BMC cancer* 15 (2015): 1-10.

Pecorino, Lauren. *Molecular biology of cancer: mechanisms, targets, and therapeutics*. Oxford university press, 2021.

P. Erdós and A. Rényi, "On the evolution of random graph I", *Publicationes Mthematicae Debrecen* 6:290(1959).

Phillips MA, Burrows JN, Manyando C, van Huijsduijnen RH, Van Voorhis WC, TNC W. Malaria. Nat Rev Dis Prim. Macmillan Publishers Limited. 2017;3:17050.

Pires, Jorge Guerra, et al. "Galaxy and MEAN Stack to create a user-friendly workflow for the rational optimization of cancer chemotherapy." *Frontiers in Genetics* 12 (2021): 624259.

Poret, Arnaud, and Carito Guziolowski. "Therapeutic target discovery using Boolean network attractors: improvements of kali." *Royal Society open science* 5.2 (2018): 171852.

Rackham, Owen JL, et al. "A predictive computational framework for direct reprogramming between human cell types." *Nature genetics* 48.3 (2016): 331-335.

Raman, Karthik. *An introduction to computational systems biology: systems-level modelling of cellular networks*. Chapman and Hall/CRC, 2021. pp. 75

Ratner, Elena, et al. "Increased risk of brain metastases in ovarian cancer patients with BRCA mutations." *Gynecologic Oncology* 153.3 (2019): 568-573.

Rodrigues, Lindsey Ulkus, et al. "Coordinate loss of MAP3K7 and CHD1 promotes aggressive prostate cancer." *Cancer research* 75.6 (2015): 1021-1034.

Roy, Rohini, Jarin Chun, and Simon N. Powell. "BRCA1 and BRCA2: different roles in a common pathway of genome protection." *Nature Reviews Cancer* 12.1 (2012): 68-78.

Sabatier, Renaud, Anthony Goncalves, and Francois Bertucci. "Personalized medicine: present and future of breast cancer management." *Critical reviews in oncology/hematology* 91.3 (2014): 223-233

Saliba, Antoine-Emmanuel, et al. "Single-cell RNA-seq: advances and future challenges." *Nucleic acids research* 42.14 (2014): 8845-8860.

Schwab, Julian D., et al. "Concepts in Boolean network modeling: What do they all mean?." *Computational and structural biotechnology journal* 18 (2020): 571-582.

Seah, Yu Fen Samantha, et al. "Induced pluripotency and gene editing in disease modelling: perspectives and challenges." *International Journal of Molecular Sciences* 16.12 (2015): 28614-28634.

Sgariglia, D., Conforte, A. J., de Carvalho, L. A. V., Carels, N., & da Silva, F. A. B. (2018). Cellular reprogramming. *Theoretical and Applied Aspects of Systems Biology*, 41-55. DOI:10.1007/978-3-319-74974-7_3

Sgariglia, Domenico, et al. "Data-driven modeling of breast cancer tumors using Boolean networks." *Frontiers in Big Data* 4 (2021): 656395.

Sgariglia, Domenico, et al. "Optimizing therapeutic targets for breast cancer using Boolean network models." *Computational Biology and Chemistry* (2024): 108022.

Sharma, Abhilasha, Sonam Sinha, and Neeta Shrivastava. "Therapeutic targeting hypoxia-inducible factor (HIF-1) in cancer: cutting gordian knot of cancer cell metabolism." *Frontiers in genetics* 13 (2022): 849040.

Shi, Yonghong, et al. "Role and mechanism of hypoxia-inducible factor-1 in cell growth and apoptosis of breast cancer cell line MDA-MB-231." *Oncology letters* 1.4 (2010): 657-662.

Shin, Dongkwan, and Kwang-Hyun Cho. "Critical transition and reversion of tumorigenesis." *Experimental & Molecular Medicine* 55.4 (2023): 692-705.

Shmulevich, Ilya, et al. "Steady-state analysis of genetic regulatory networks modelled by probabilistic Boolean networks." *Comparative and functional genomics* 4.6 (2003): 601-608.

Siegle, Lea, et al. "A Boolean network of the crosstalk between IGF and Wnt signaling in aging satellite cells." *PLoS One* 13.3 (2018): e0195126.

Siminovitch, Louis, Ernest A. McCulloch, and James E. Till. "The distribution of colony-forming cells among spleen colonies." (1963).

Somogyi, Roland, and Carol Ann Sniegoski. "Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation." *complexity* 1.6 (1996): 45-63.

Stuelten, Christina H., et al. "Smad4-expression is decreased in breast cancer tissues: a retrospective study." *BMC cancer* 6 (2006): 1-10.

Sturm, Isrid, et al. "Impaired BAX protein expression in breast cancer: mutational analysis of the BAX and the p53 gene." *International journal of cancer* 87.4 (2000): 517-521.

Su, Hang, et al. "Decoding early myelopoiesis from dynamics of core endogenous network." *Science China Life Sciences* 60 (2017): 627-646.

Sung, Hyuna, et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." *CA: a cancer journal for clinicians* 71.3 (2021): 209-249.

Suzuki H, Forrest ARR, van Nimwegen E, Daub CO, Balwierz PJ, Irvine KM, et al. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat Genet. 2009;41(5):553–62.

Szallasi, Zoltan, and Shoudan Liang. "Modeling the normal and neoplastic cell cycle with realistic Boolean genetic networks: their application for understanding carcinogenesis and assessing therapeutic strategies." *Pacific Symposium on Biocomputing*. Vol. 3. 1998.

Tabernero, Josep, et al. "First-in-humans trial of an RNA interference therapeutic targeting VEGF and KSP in cancer patients with liver involvement." *Cancer discovery* 3.4 (2013): 406-417.

Takahashi, Kazutoshi. "Cellular reprogramming." *Cold spring harbor perspectives in biology* 6.2 (2014): a018606.

Takahashi, Kazutoshi, and Shinya Yamanaka. "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." *cell* 126.4 (2006): 663-676.

Hase, Takeshi, et al. "Identification of drug-target modules in the human protein–protein interaction network." *Artificial Life and Robotics* 19 (2014): 406-413.

Theillet, Charles, et al. "FGFRI and PLAT genes and DNA amplification at 8p 12 in breast and ovarian cancers." *Genes, Chromosomes and Cancer* 7.4 (1993): 219-226.

Thomas, René. "Boolean formalization of genetic control circuits." *Journal of theoretical biology* 42.3 (1973): 563-585.

Tilli, Tatiana M., et al. "Validation of a network-based strategy for the optimization of combinatorial target selection in breast cancer therapy: siRNA knockdown of network targets in MDA-MB-231 cells as an in vitro model for inhibition of tumor development." *Oncotarget* 7.39 (2016): 63189.

Vernimmen, D., et al. "Different mechanisms are implicated in ERBB2 gene overexpression in breast and in other cancers." *British journal of cancer* 89.5 (2003): 899-906.

Vogelstein, Bert, and Kenneth W. Kinzler. "Cancer genes and the pathways they control." *Nature medicine* 10.8 (2004): 789-799.

Waddington, Conrad Hal. "Organisers and genes." *Organisers and genes.* (1940).

Waddington CH. An introduction to modern genetics. New York: The Macmillan Company;
1939.

Waddington C. The strategy of the genes: a discussion of some aspects of theoretical biology.

London: George Allen and Unwin; 1957. 262 pp

Waddington CH. Towards a theoretical biology. Nature. 1968;218(5141):525–7.

Wakayama, Teruhiko, et al. "Differentiation of embryonic stem cell lines generated from adult somatic cells by nuclear transfer." *Science* 292.5517 (2001): 740-743.

Wang, Ping, et al. "Modeling the regulation of p53 activation by HIF-1 upon hypoxia." *FEBS letters* 593.18 (2019): 2596-2611.

Wang, Lingling, Shizhen Zhang, and Xiaochen Wang. "The metabolic mechanisms of breast cancer metastasis." *Frontiers in Oncology* 10 (2021): 602416.

Washino, Satoshi, et al. "Loss of MAP3K7 sensitizes prostate cancer cells to CDK1/2 inhibition and DNA damage by disrupting homologous recombination." *Molecular Cancer Research* 17.10 (2019): 1985-1998.

Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of 'small-world'networks." *nature* 393.6684 (1998): 440-442.

Wong, Rebecca SY. "Apoptosis in cancer: from pathogenesis to treatment." *Journal of experimental & clinical cancer research* 30 (2011): 1-14.

Xiao, Yufei. "A tutorial on analysis and simulation of boolean gene regulatory network models." *Current genomics* 10.7 (2009): 511-525.

Yamanaka, Shinya, and Helen M. Blau. "Nuclear reprogramming to a pluripotent state by three approaches." *Nature* 465.7299 (2010): 704-712.

Yang, Muh-Hwa, and Kou-Juey Wu. "TWIST activation by hypoxia inducible factor-1 (HIF-1): implications in metastasis and development." *Cell cycle* 7.14 (2008): 2090-2096.

Yoshida, Tatsushi, et al. "Repeated treatment with subtoxic doses of TRAIL induces resistance to apoptosis through its death receptors in MDA-MB-231 breast cancer cells." *Molecular Cancer Research* 7.11 (2009): 1835-1844.

Yu, Chong, and Jin Wang. "A physical mechanism and global quantification of breast cancer." *PloS one* 11.7 (2016): e0157422.

Yuan, Ruoshi, et al. "Cancer as robust intrinsic state shaped by evolution: a key issues review." *Reports on Progress in Physics* 80.4 (2017): 042701.

Zavitsanos, Peter J., et al. "BRCA1 mutations associated with increased risk of brain metastases in breast cancer: a 1: 2 matched-pair analysis." *American journal of clinical oncology* 41.12 (2018): 1252-1256.

Zeng, Yan, et al. "Inhibition of STAT5a by Naa10p contributes to decreased breast cancer metastasis." *Carcinogenesis* 35.10 (2014): 2244-2253.

Zhou, Chunxia, et al. "Long noncoding RNA HOTAIR, a hypoxia-inducible factor-1α activated driver of malignancy, enhances hypoxic cancer cell proliferation, migration, and invasion in non-small cell lung cancer." *Tumor Biology* 36 (2015): 9179-9188.

Zhou, L. L., et al. "MicroRNA-143 inhibits cell growth by targeting ERK5 and MAP3K7 in breast cancer." *Brazilian Journal of Medical and Biological Research* 50 (2017).

Zhou, Joseph Xu, et al. "Discrete gene network models for understanding multicellularity and cell reprogramming: From network structure to attractor landscapes landscape." *Computational Systems Biology: From Molecular Mechanisms to Disease: Second Edition*. Elsevier Inc., 2013. 241-276.

Zickenrott, S., et al. "Prediction of disease–gene–drug relationships following a differential network analysis." *Cell death & disease* 7.1 (2016): e2040-e2040.

Publisced Papers

Sgariglia, D., Conforte, A. J., de Carvalho, L. A. V., Carels, N., & da Silva, F. A. B. (2018). Cellular reprogramming. *Theoretical and Applied Aspects of Systems Biology*, 41-55. . https://doi.org/10.1007/978-3-319-74974-7_3

Sgariglia D, Conforte AJ, Pedreira CE, Vidal de Carvalho LA, Carneiro FRG, Carels N, Silva FABD. Data-Driven Modeling of Breast Cancer Tumors Using Boolean Networks. Front Big Data. 2021 October 20;4:656395. https://doi/10.3389/fdata.2021.656395.

- Sgariglia, D., Gonçalves Carneiro, F. R., Vidal de Carvalho, L. A., Pedreira, C. E., Carels, N., & Barbosa da Silva, F. A. Optimizing therapeutic targets for breast cancer using Boolean network models . Computational Biology and Chemistry (2024): 108022 DOI: 10.1016/j.compbiolchem.2024.108022