ALMA MATER STUDIORUM · UNIVERSITY OF BOLOGNA

School of Science
Department of Physics and Astronomy
Master Degree in Physics

Modeling cell differentiation using dynamical systems on graphs

Supervisor:
Prof. Armando Bazzani

Co-supervisor:
Prof. Fabio Luciani

Submitted by:
Riccardo Scheda

Academic Year 2019/2020

Sommario

La cellula vivente è un sistema complesso governato da molti processi che non sono ancora stati compresi: il processo di differenziazione cellulare è uno di questi. La differenziazione cellulare è il processo in cui le cellule di un tipo specifico si riproducono e danno origine a diversi tipi di cellule. La differenziazione cellulare è regolata dai cosiddetti Gene Regulatory Networks (GRN). Un GRN è una raccolta di regolatori molecolari che interagiscono tra loro e con altre sostanze nella cellula per governare i livelli di espressione genica di mRNA e proteine. Kauffman propose per la prima volta nel 1969 di modellare GRN attraverso le cosiddette Random Boolean Networks (RBN). I RBN sono reti in cui ogni nodo può avere solo due possibili valori: 0 o 1, dove ogni nodo rappresenta un gene in GRN che può essere ″on″ oppure ″off″. Queste reti possono modellizzare i GRN perché l'attività di un nodo rappresenta il livello di espressione di un gene nell'intera regolazione.

In questo lavoro di tesi ci avvaliamo di un modello matematico per sviluppare e riprodurre una possibile rete di regolazione genica per il processo di differenziazione cellulare.

Abstract

Real living cell is a complex system governed by many process which are not yet understood: the process of cell differentiation is one of these. Cell differentiation is the process in which cells of a specific type reproduces themselves and give arise to different type of cells. Cell differentiation is governed by the so called Gene Regulatory Networks (GRNs). A GRN is a collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins. Kauffman proposed for the first time in 1969 to model GRN through the so called Random Boolean Networks (RBN). RBNs are networks in which each node can have only two possible values: 0 or 1, where each node represent a gene in GRN which can be ”on” or ”off”. These networks can model GRNs because the activity of one node represents the expression level of one gene among the whole regulation.

In this thesis work we make use of a mathematical model to develop and reproduce a possible Gene Regulatory Network for the process of cell differentiation.

Contents

List of Figures

Introduction

The cell is a paradigmatic example of complex system governed by many processes which are not yet understood: the process of cell differentiation is one of these. Cell differentiation is the dynamical process in which stem cells reproduce and give arise to different type of cells. Waddington in 1957 [1] proposed to model cell differentiation with an epigenetic landscape in which lay different type of cells. He portrayed the epigenetic landscape as an inclined surface with a cascade of branching ridges and valleys , which in the context of cell lineage selection, represent the series of "either/or" fate choices made by a developing cell. This epigenetic landscape can be seen as a potential in a physical system in which different type of cells are attracted by the different wells of this potential. Now, it is well known that cell differentiation is governed by the so called Gene Regulatory Networks (GRNs). A GRN is a collection of molecular regulators that interact with each others and with other substances in the cell to define the gene expression levels of mRNA and proteins.

Recent studies on omics data propose ways to measure epigenetic landscape for different cells[2][3][4], using stochastic processes. In these works it is proposed a probabilistic "pseudo-potential" to quantify the epigenetic landscape for a genetic network regulating cell fate, where the elevation of the surface is inversely related to the likelihood of occurrence of a particular state in phase space. In this formulation a stochastic potential energy landscape is characterized for a genetic network, based on a Hartree mean-field approximation of the underlying master equation.

We propose a new approach that highlights the possibility of the existence of a fitness landscape potential as an emergent property from the dynamics of interacting genetic networks. Our idea is to measure epigenetic landscape starting from mod-

eling GRNs through Random Boolean Networks (RBNs), which were introduced for the first time by Stuart Kauffman in 1969 [6]. RBNs are networks in which each node can have only two possible values: 0 or 1, where each node represents a gene in GRN which can be ”on” or ”off”[7]. The evolution of the state of the network is given by some boolean functions, depending on the connectivity of the nodes. So each node will have one boolean function which defines the next state during the descrete evolution.

The model proposed, based on these random networks, focuses in the representation of GRNs for cell differentiation, and aims to find a structure of fitness landscape. The main concept of the model is that one type of cell is governed by a specific type of GRN. In this sense, since cell differentiation concerns multiple types of cells, we want to construct a random network made by multiple and interacting clusters, which simulate the activity and the gene expression in GRN and so the proces of differentiation. The idea is based on the fact that the production of a specific type of cell may hinibit the raise of one other and viceversa. This this can be a seen as a bistable dynamical system. To bistable dynamical systems can be associated a potential which presents two different stationary states, and in the analysis of this model we can find that constructing a network with two interacting clusters, they present a bistable nature. Thus, using Kramer Transition Rate Theory, from this bistability we can make an estimate for a double well potential, recovering the theory of epigenetic landascape of Wadington.

In Chapters 1 and 2 we make a biological introduction of Cell Differentiation and of Gene Regulatory Network. In Chapter 3 we present the concept of Random Boolean Networks, a model proposed for the first time by Kauffman to model Gene Regulatory Networks. In Chapter 4 we propose a new model based on Random Boolean Networks but with some differences, where networks can have also hinibitory links between clusters. In Chapter 5 we make a briefly introduction on Kramer Transition Rate Theory. In Chapter 6 we make a numerical analysis of the thoretical model proposed in 4 and show the results.

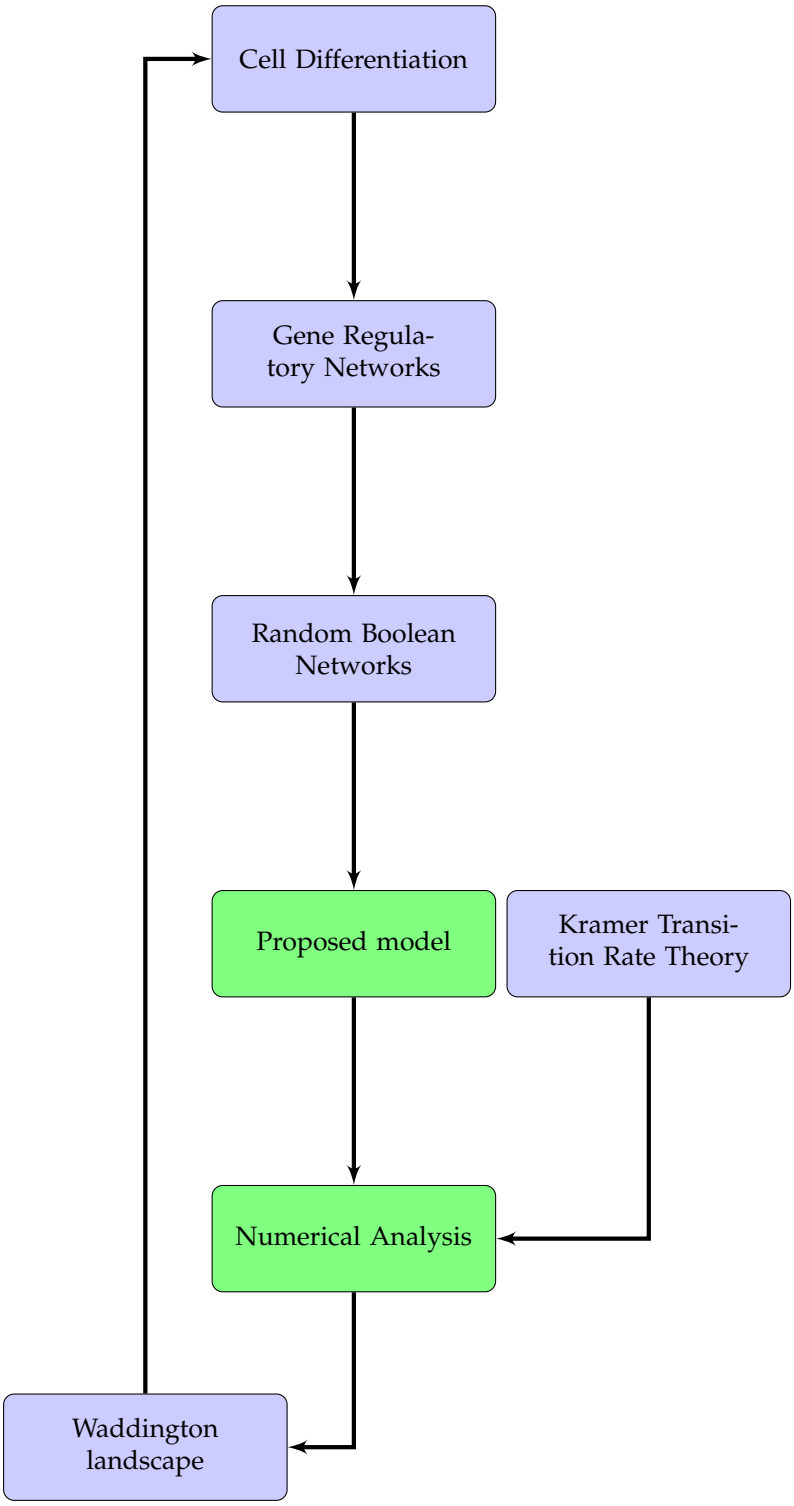

Figure 1: *Road map of the thesis work. First, we start introducing the process of cell differentiation and Gene Regulatory Networks. Then, we introduce the main theoretical concepts among Random Boolean Networks. Finally we present our theoretical model. At the end, we make analysis of the model with numerical simulations, and using Kramer Theory we try recover the Waddington epigenetic landscape.*

Chapter 1

Cell Differentiation

In this Chapter we explain cell differention and the role of genetic networks among this process.

1.1 Definition

Cell differentiation is the process whereby stem cells become progressively more specialized. The differentiation process occurs both during the development of a multicellular organism and during tissue repair and cell turnover in the adulthood. Gene expression, and therefore its regulatory mechanisms, plays a critical role in cell differentiation. Stem cells are undifferentiated biological cells which can both reproduce themselves, self-renewal ability, and differentiate into specialized cells, potency. The principles underlying cellular differentiation remain among the most enigmatic in biology. We are required to explain the spontaneous generation of a multiplicity of cell types from the single zygote, to deduce a natural tendency of a system to become increasingly heterogeneous, then to stop differentiating.

Among the important characteristics of cell differentiation are:

- initiation of change;
- stabilization of change after cessation of stimulus;

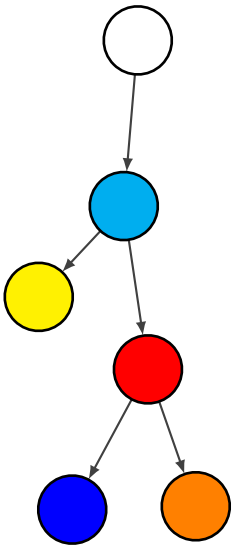

Figure 1.1: *Schematic representation of cellular differentiation.*

- the efficacy of many substances, exogenous and endogenous, as inductive stimuli;
- progressive limitation in the number of developmental pathways open to any small region of the embryo;
- restricted periods during which a cell is competent to respond to an inductive stimulus; the discreteness of cell types, that is, the mutually exclusive constellations of properties by which cells differ;
- a requirement for a minimal and preferably heterogeneous cell mass to initiate differentiation in many instances, and to maintain it in some;
- the occurrence of metaplasia between undifferentiated cell types, or from an undifferentiated type to a specialized type, but the lack of metaplasia (the isolation) between specialized cell types;
- cessation of differentiation.

Cells are thought to differ due to differential expression of, rather than structural loss of, the genes. Differential activity of the genes raises at least two questions which are not always carefully distinguished: the capacity of the genome to behave in more than one mode; and mechanisms which insure the appropriate assignment of these modes to the proper cells.

Within multicellular organisms, tissues are organized in communities of cells that work together to carry out a specific function. The exact role of a tissue in an organism depends on what types of cells it contains. For example, the endothelial tissue that lines the human gastrointestinal tract consists of several cell types. Some of these cells absorb nutrients from the digestive contents, whereas others secrete a lubricating mucus that helps the contents travel smoothly. However, the multiple cell types within a tissue don't just have different functions. They also have different transcriptional programs and may well divide at different rates. Proper regulation of these rates is essential to tissue maintenance and repair. Stem cells typically have the capacity to mature into many different cell types. *Transcription factors* (TF), which are proteins that regulate which genes are transcribed in a cell, appear to be essential to determining the pathway particular stem cells take as they differentiate. For example, both intestinal absorptive cells and goblet cells arise from the same stem cell population, but divergent transcriptional programs cause them to mature into dramatically different cells. Whenever stem cells are called upon to generate a particular type of cell, they undergo an asymmetric cell division. With asymmetric division, each of the two resulting daughter cells has its own unique life course. In this case, one of the daughter cells has a finite capacity for cell division and begins to differentiate, whereas the other daughter cell remains a stem cell with unlimited proliferative ability. Waddington in 1957 [1] proposed to model cell differentiation with an epigenetic landscape in which lay different type of cells. He portrayed the epigenetic landscape as an inclined surface with a cascade of branching ridges and valleys , which in the context of cell lineage selection, represent the series of "either/or" fate choices made by a developing cell (see figure 1.2).

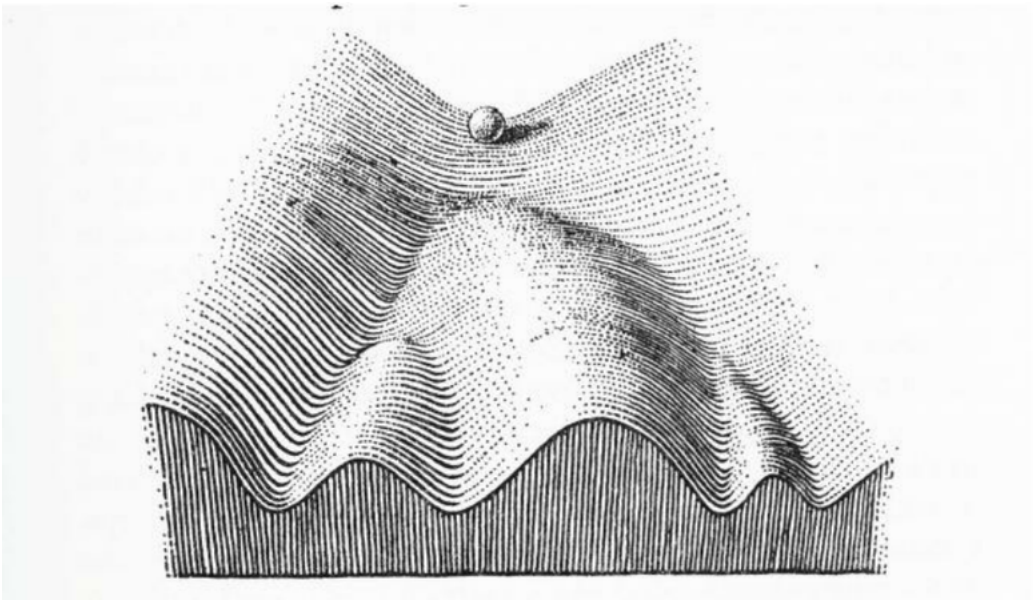

Figure 1.2: *The "epigenetic landscape" proposed by Wadington shows a ball rolling down valleys on an inclined surface, as a visual metaphor for branching pathways of cell fate (image from [1]).*

The process of cell differentiation can be seen by a physical point of view: a multistable dynamical system governed by a multi well potential in which different cell fates correspond to different stationary states of the system.

1.2 The role of Gene expression in cellular differentiation

Gene expression is a complex process regulated at several stages in the synthesis of proteins. In addition to the DNA transcription regulation, the expression of a gene may be controlled during RNA processing and transport (in eukaryotes), RNA translation, and the post-translational modification of proteins. This gives rise to genetic regulatory systems structured by networks of regulatory interactions between DNA, RNA, proteins and other molecules: a complex network termed as a *gene regulatory network* (GRN) (see Chapter 2). Some kind of proteins are the transcription factors that

bind to specific DNA sequences in order to regulate the expression of a given gene. The power of transcription factors resides in their ability to activate and/or repress transcription of genes. The activation of a gene is also referred to positive regulation, while the negative regulation identifies the inhibition of the gene[11]. The regulation of gene expression is essential for the cell, because it allows to control the internal and external functions of the cell. Furthermore, in multicellular organisms, gene regulation drives the processes of cellular differentiation and morphogenesis, leading to the creation of different cell types that possess different gene expression profiles, and these last therefore produce different proteins that have different ultrastructures that suit them to their functions. Therefore, with few exceptions, all cells in an organism contain the same genetic material, and hence the same genome. The difference between the cells are emergent and due to regulatory mechanisms which can turn on or off genes. Two cells are different, if they have different subsets of active genes. In the next Chapter we will see in details Gene Regulatory Networks and the principles for modeling them.

Chapter 2

Gene Regulatory Networks

2.1 Definition

Gene regulation controls the expression of genes and, consequently, all cellular functions. Gene expression is a process that involves transcription of the gene into mRNA, followed by translation to a protein, which may be subject to post-translational modification [13]. The transcription process is controlled by transcription factors (TFs) that can work as activators or inhibitors. TFs are themselves encoded by genes and subject to regulation, which altogether forms complex regulatory networks. Cells efficiently carry out molecular synthesis, energy transduction, and signal processing across a range of environmental conditions by networks of genes, which we define broadly as networks of interacting genes, proteins, and metabolites [14]. Formally speaking, a *gene regulatory network* or *genetic regulatory network* (GRN) is a collection of DNA segments in a cell which interact with each other (indirectly through their RNA and protein expression products) and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed into mRNA. In general, each mRNA molecule goes on to make a specific protein (or set of proteins). In some cases this protein will be structural, and will accumulate at the cell-wall or within the cell to give it particular structural properties.

These networks control biological process of all organisms. The complex control systems underlying development have probably been evolving for more than a billion

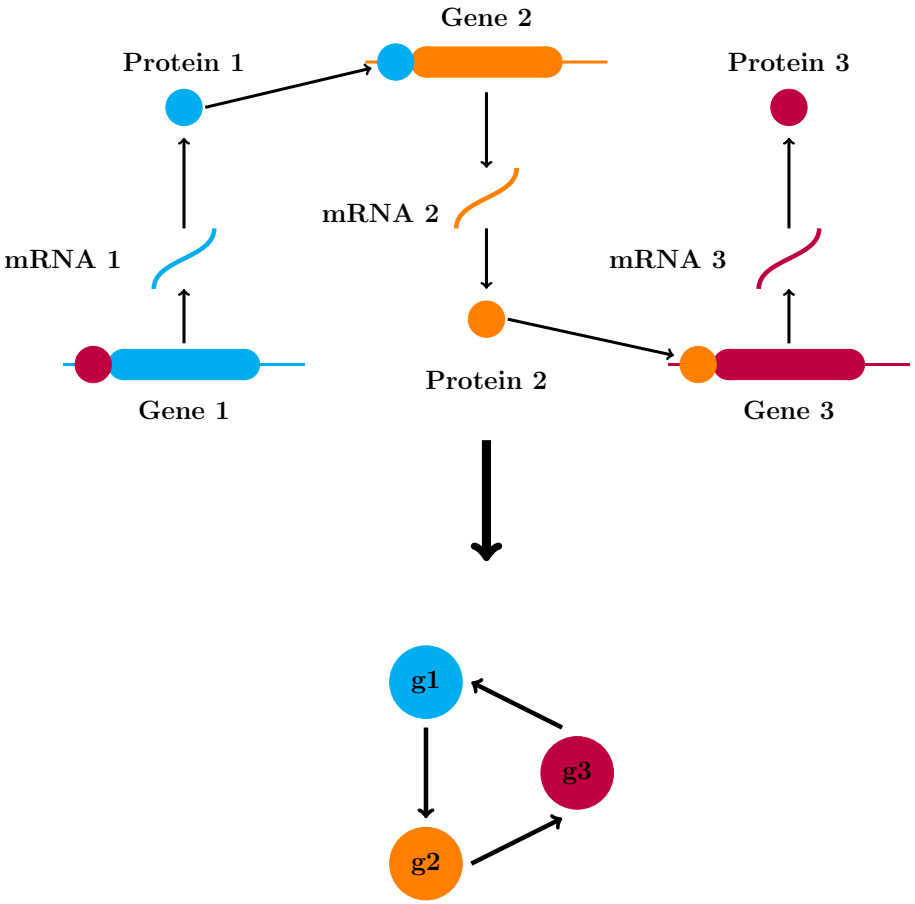

Figure 2.1: *Schematic representation of a Gene Regulatory Network.*

years. They regulate the expression of thousand of genes in any given biological process. They are essentially hardwired genomic regulatory codes, the role of which is to specify the sets of genes that must be expressed in specific spatial and temporal patterns. In physical terms, these control system consist of many thousands of modular DNA sequences. Each module receives and integrates multiple inputs, in the form of regulatory proteins (*activators* and *repressors*) that recognize specific sequences within them. The end result is the precise transcriptional control of the associated genes. Functional linkages between these particular genes, and their associated regulatory modules, define the core networks underlying development. They explain exactly how genomic sequence encodes the regulation of expression of the sets of genes that generate patterns and execute the construction of multiple states of differentiation.

The regulatory genome is a logic processing system: every regulatory module contained in the genome receives multiple inputs and processes in ways that can be mathematically represented as combinations of logic functions.

Definitive regulatory functions emerge only from the architecture of intergenic linkages, and these functions are not visible at the level of any individual genes. So gene regulatory networks can be determined only by experimental molecular biology in which the functional meaning of given regulatory sequences is directly determined.

GRNs have a complex structure: they are inhomogeneous compositions of different kinds of subnetworks, each performing a specific kind of function. Some subnetworks are used in many processes.

In principle, mathematical modeling of GRN dynamics can provide a theoretical foundation for understanding cell heterogeneity and gene expression dynamics, by quantitatively linking molecular-level regulatory mechanisms with observed cell states. However, due to the molecular complexity of gene regulatory mechanisms, it remains challenging to integrate such models with single-cell data.

2.2 Modelling GRNs

Mathematical models can account for (and at least partially reproduce) observed cellular heterogeneity in two primary ways. First, gene network models are multi-

stable dynamical systems, meaning a given network has the potential to reach multiple stable states of gene expression. These states arise from the dynamic interplay of activation, inhibition, feedback, and nonlinearity [6] [23]. Second, some mathematical models inherently treat cellular noise. This noise, or stochasticity, is modeled in various ways depending on assumptions about the source [15] [16]. Discrete, stochastic models of gene regulation, which track discrete molecular entities, regulatory-protein binding kinetics, and binding states of promoters controlling gene activity, have formed the basis of biophysical theories of gene expression noise due to so-called intrinsic molecular noise [16] [17]. Such stochastic gene regulation mechanisms have also been incorporated into larger regulatory network models using the formalism of stochastic biochemical reaction networks, and have been utilized to explore how molecular fluctuations can cause heterogeneity within phenotype-states and promote stochastic transitions between phenotypes [18] [19].

The quantitative landscape of cellular states is another concept that is increasingly utilized to describe cellular heterogeneity. Broadly, the cellular potential landscape (first conceptualized by Waddington [4][1]) is a function in high-dimensional space (over many molecular observables, typically expression levels of different genes), that quantifies the stability of a given cell state. In analogy to potential energy (gravitational, chemical, electric, etc.), cell states of higher potential are less stable than those of lower potential. The landscape concept inherently accounts for cellular heterogeneity, since it holds that a continuum of states is theoretically accessible to the cell, with low-potential states (in "valleys") more likely to be observed than high-potential states. The landscape is a rigorously defined function derived from the dynamics of the underlying gene network model, according to some choice of mathematical formalism [20][4].

Stochastic modeling of gene network dynamics has been employed in various forms for analysis of single cell measurements. However, few existing analysis methods utilize discrete-molecule, stochastic models, which fully account for intrinsic gene expression noise and its impact on cell-state, to aid in the interpretation of noisy distributions recovered from single cell RNA sequencing data [2][3]. There exists an opportunity to link such biophysical, stochastic models, which reproduce intrinsic noise

and cell heterogeneity in silico, to single cell datasets that characterize cell heterogeneity in vivo. In particular, the landscape of heterogeneous cell states computed from discrete stochastic models can be directly compared to single-cell measurements.

2.3 Principle of the models

GRNs may be interpreted as an idealized dynamical system of model genes with directional links (transcription factors), updating their state in parallel, according to the combinatorial logic of their inputs, Kauffman's Random Boolean Networks [6][24]. Gene activity at the molecular scale consists of discrete events occurring concurrently. Variable protein concentrations can be accounted for by genes being on for some fraction of a given time span. The RBN idealization is arguably a valid starting point for gaining insights into gene network dynamics. In a cell type's gene expression pattern over a span of time (i.e. its space- time pattern), a particular gene may, broadly speaking, be either on, off, or changing. If a large proportion of the genes are changing, chaotic dynamics, the cell will be unstable. On the other hand, dynamics that settles to a pattern where a large proportion of the genes are permanently on or o (frozen) may be too in exible for adaptive behavior. Cells constantly need to adapt their gene expression pattern in response to a variety of hormone and growth/di erentiation factors from nearby cells. The definition of a cell type may be more correctly expressed as a set of closely related gene expression patterns, allowing an essential measure of exibility in behavior.

In the next Chapter we will concentrate on the Random Boolean Networks proposed by Kauffmann in 1969.

Chapter 3

Random Boolean Networks

In this chapter we explain the basic concepts of Random Boolean Network proposed for the first time by Kauffman. In our new model (see Chapter 4 we will take the first basic concepts of Ranom Boolean Networks and we will deviate from Kauffman model.

3.1 Random Boolean Networks

Random Boolean networks (RBNs) were introduced in 1969 by S. Kauffman as a simple model of genetic systems [6]. Each gene was represented by a node that has two possible states, "on" (corresponding to a gene that is being transcribed) and "off" (corresponding to a gene that is not being transcribed). There are altogether N nodes, and each node receives input from K randomly chosen nodes, which represent the genes that control the considered gene. Furthermore, each node is assigned an *update boolean function* that prescribes the state of the node in the next time step, given the state of its input nodes. This update function is chosen from the set of all possible update functions according to some probability distribution. Starting from some initial configuration, the states of all nodes of the network are updated in parallel. Since configuration space is finite and since dynamics is deterministic, the system must eventually return to a configuration that it has had before, and from then on it repeats the same sequence of configurations periodically.

3.2 The model

Let's consider a network of N nodes. The state of each node at a time t is given by $\sigma_i(t) \in \{0,1\}$ with $i = 1,\ldots,N$. The N nodes of the network can therefore together assume 2^N different states. The number of incoming links to each node i is denoted by k_i and is drawn randomly independently from the distribution $P(k_i)$. The dynamical state of each $\sigma_i(t)$ is updated synchronously by a Boolean function Λ_i [35]:

$$\Lambda_i : \{0,1\}^{k_i} \rightarrow \{0,1\}$$

An update function specifies the state of a node in the next time step, given the state of its K inputs at the present time step. Since each of the K inputs of a node can be on or off, there are $M = 2^K$ possible input states. The update function has to specify the new state of a node for each of these input states. Consequently, there are 2^M different update functions. For example let's consider a network with $K = 1$, so all the functions Λ_i receives the input from one single node. In general each element receives inputs from exactly K nodes, so we have a dynamical system defined from:

$$\sigma_i(t+1) = \Lambda_i(\sigma_{i_1}(t), \sigma_{i_2}(t), ..., \sigma_{i_K}(t)). \tag{3.1}$$

So, the randomness of these network appears at two levels: in the connectivity of the network (which node is linked to which) and the dynamics (which function is attributed to which node).

3.3 Dynamics

All nodes are updated at the same time according to the state of their inputs and to their update function. Starting from some initial state, the network performs a trajectory in state space and eventually arrives on an *attractor*, where the same sequence of states is periodically repeated. Since the update rule is deterministic, the same state must always be followed by the same next state. If we represent the network states by points in the 2^N-dimensional state space, each of these points has exactly one "output", which is the successor state. We thus obtain a graph in state space. The size

or length of an attractor is the number of different states on the attractor. The basin of attraction of an attractor is the set of all states that eventually end up on this attractor, including the attractor states themselves. The size of the basin of attraction is the number of states belonging to it. The graph of states in state space consists of unconnected components, each of them being a basin of attraction and containing an attractor, which is a loop in state space. The transient states are those that do not lie on an attractor. They are on trees leading to the attractors.

Let us illustrate these concepts by studying the small $K = 1$ network shown in Figure 3.1, which consists of 4 nodes:

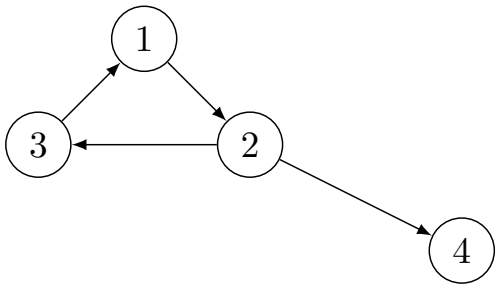

Figure 3.1: *A small network with* $N = 4$ *and* $K = 1$.

If we assign to the nodes 1,2,3,4 the functions invert, invert, copy, copy, an initial state 1111 evolves in the following way:

$$1111 \rightarrow 0011 \rightarrow 0100 \rightarrow 1111$$

This is an attractor of period 3. If we interpret the bit se- quence characterizing the state of the network as a number in binary notation, the sequence of states can also be written as

$$15 \rightarrow 3 \rightarrow 4 \rightarrow 15$$

The entire state space is shown in Figure 3.2:

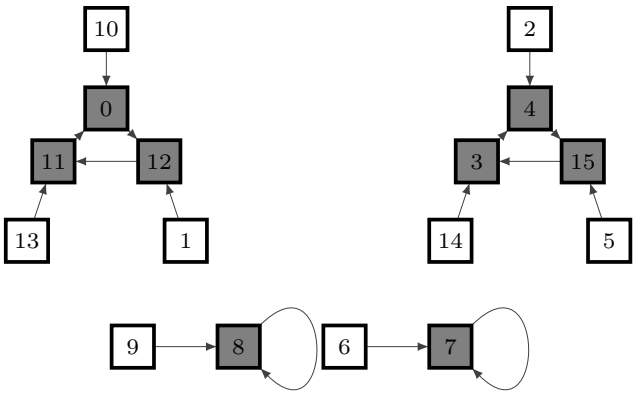

Figure 3.2: *The state space of the network shown in Figure 3.1, if the functions copy, copy, invert, invert are assigned to the four nodes. The numbers in the squares represent states, and arrows indicate the successor of each state. States on attractors are shaded.*

There are 4 attractors, two of which are fixed points (i.e., attractors of length 1). The sizes of the basins of attraction of the 4 attractors are $\Omega_1 = 6, \Omega_2 = 6, \Omega_3 = 2, \Omega_4 = 2$. If the function of node 1 is a constant function, fixing the value of the node at 1, the state of this node fixes the rest of the network, and there is only one attractor, which is a fixed point.

3.4 Dynamical phase transitions

In RBNs, as well as in many dynamical systems, three phases can be distinguished: *ordered, chaotic, and critical* [29]. These phases can be identified with different methods, since they have several unique features. these dynamical phases is related to "sensitivity to initial conditions", "damage spreading", and "robustness to perturbations" which are different ways of measuring the stability of a network. We can "mutate", "damage" or "perturb" a node of a RBN by flipping its state. We can also change a

connection between two nodes, or in the lookup table of a node. Since nodes affect other nodes, we can measure how much a random change affects the rest of the network. In other words, we can measure how the damage spreads. This can be done by comparing the evolution of a "normal" network and a "perturbed" network. In the ordered regime, usually the damage does not spread: a "perturbed" network "returns" to the same path of the "normal" network. This is because changes cannot propagate from one green island to another. In the chaotic phase, these small changes tend to propagate through the network, making it highly sensitive to perturbations [22]. An other feature is the convergence versus divergence of the trajectories in state space of the network dynamics. In the ordered phase, similar states tend to converge to the same state. In the chaotic regime, similar states tend to diverge. At the edge of chaos, nearby states tend to lie on trajectories that neither converge nor diverge in state space. Living systems, or computing systems, need certain stability to survive, or to keep information; but also flexibility to explore their space of possibilities. This has lead people to argue that life and computation occur more naturally at the edge of chaos or at the ordered regime close to the edge of chaos [25][22].

Very early in the studies of RBNs, people realized in simulations that the networks with $K \leq 2$ were in the ordered regime, and networks with $K \geq 3$, were in the chaotic regime. In Figure 3.3 we can appreciate characteristic dynamics of RBNs in different phases. We can identify phase transitions in RBNs in different ways. The main idea is to measure the effect of perturbations, the sensitivity to initial conditions, or damage spreading. This is analogous to Lyapunov exponents in continuous dynamics. The phase transitions can be statistically or analytically obtained. Derrida and Pomeau were the first to determine analytically that the critical phase (edge of chaos) was found when $K = 2$ [26][35][28]. following the model of Kauffman, RBN which most represent biological GRNs are those wich has $K = 2$ [25], because in the frozen phase (where $K = 1$) networks are too simple to represent real regulatory networks; while in the caothic phase (where $K = 3$) the time scales of the networks cycles grow exponentially, which is not biologically pheasible[34]. In Chapter 6 we will see the differences in $K = 1$ networks and $K = 2$ networks.

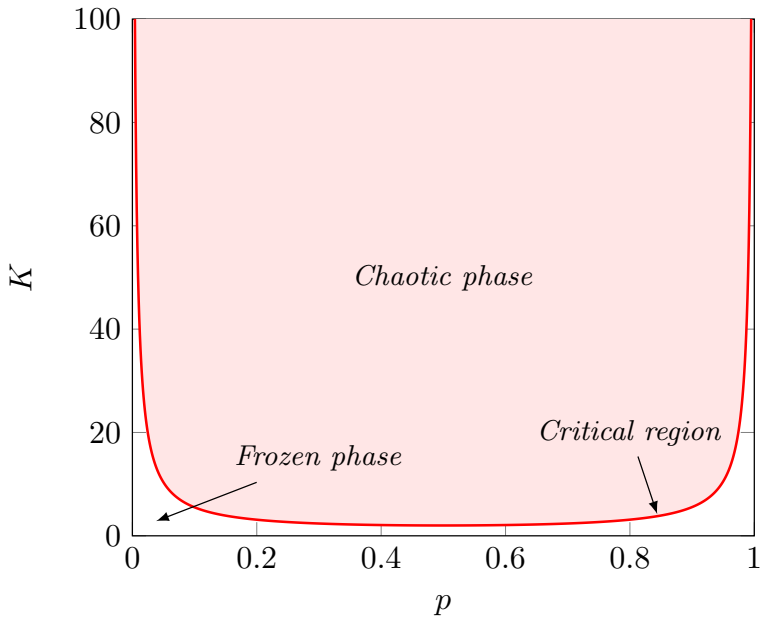

Figure 3.3: *Phase diagram for the $N - K$ model. The shaded area corresponds to the chaotic phase, whereas the white region corresponds to the chaotic phase. The curve separating both regions is the critical phase (image from[8]).*

3.5 Attractor jumps

In Kauffman model, attractors in the state space are considered as gene regulatory networks of different cells, where different attractors represent cells of different type, and where for example cancer cells lay in one specific attractor [9][10]. Now, if we suppose that different cell types lay in different attractors, we suppose that the jump from an attractor to one other is given by a perturbation in the binary sequence of the genes. So for example we take the previous network, and consider that we are in the state 12 of the first attractor:

$$12 \rightarrow 11 \rightarrow 0 \rightarrow 12$$

And suddenly we change the state of the third node from 0 to 1:

$$1100 \rightarrow 1110$$

We change the system to have the state 14 and so we jump into the second attractor:

$$14 \rightarrow 3 \rightarrow 4 \rightarrow 15 \rightarrow 3$$

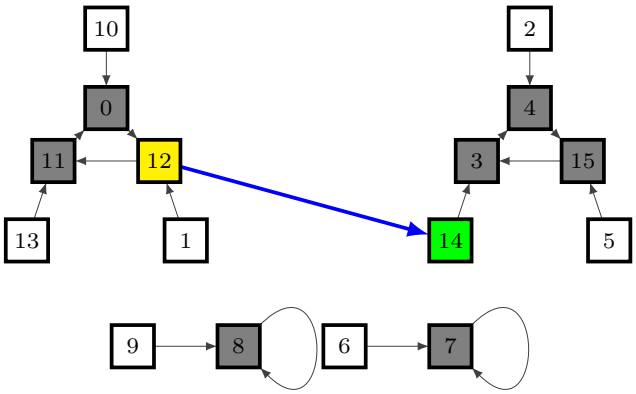

Figure 3.4: *Jump from one attractor to one other in the state space.*

So the branching pathways of differentiation between attractors in a RBN in the ensembles create a directed graph showing which attractors can be perturbed to reach which attractors.

In fact, if we consider all the possible stochastic perturbations in the binary sequence of the genes, we can get the all the possible transitions between the attractors, and what we obtain is an other network, where the nodes are the attractors and the frequencies of transitions can be used to build a random walk on this network[10][9][32].

3.6 Deviation from Kauffman model

Kauffman model on boolean networks focuses on the concept of attractors in the state space, where each attractor may represent a different type of cells, and the tran-

sition from an attractor to one other is totally stochastic and reversible process since they are disconnected components. Our model deviates from the concept of attractors, and focuses on the fact that Gene Regulatory Networks may be composed by multiple connected subnetworks (which we will call *clusters*), in which each cluster may represent a type of cell. The idea is to construct random networks with clusters which inhibit each other causing transitions of activity from one to oneother. Now, since these networks are randomly constructed, we decided to use an ensemble approach, considering the average behavior of these networks. Further, considering an ensemble of these networks, we can make use of Kramer Transition Rate Theory in order to recover the Waddington epigenetic landscape for two different type of cells.

Chapter 4

The proposed new model

In this Chapter we present our new theoretical model of Gene Regulatory Networks based on Random Boolean Networks, in order to estimate a theoretical fitness potential for cell differentiation. Kauffman model on boolean networks focuses on the concept of attractors in the state space (see Chapter 3), where each attractor may represent a different type of cells, but in the state space attractor networks are disconnected components. Our model deviates from the concept of attractors, and focuses on the fact that Gene Regulatory Networks may be composed by multiple connected subnetworks (which we will call *clusters*), in which each cluster may represent a type of cell. The idea is to construct random networks with clusters which inhibit each other causing transitions of activity from one to oneother.

4.1 Assumptions for the theoretical model

The dynamical model is a schematic representation of the activity of gene regulatory networks introduced in Chapter 2. We have to discuss the assumptions to define the model from a biological point of view: the main criticism to a model is that its assumptions cannot be justified by the biological mechanisms. Our goal is to model the genetic activity related to a differentiation process of a cell: i.e. this activity is a stable long term activity whose stability is probably controlled by biochemical mechanisms (i.e. methylation processes). Then we assume that this evolution is possible due to the

competition of different genetic activities through dynamical mechanisms that can be triggered by the external environmental signals. In particular we assume:

- the long term genetic activity is determined by the presence of small genetic networks that have a stable active dynamical state;
- there exists a control mechanism: the subnetworks have control nodes that prevent the arise of the active state in the subnetwork if there are set to the inactive state;
- once the active state has been established in a subnetwork it remains stable in time without any stimulus, except if an inhibitory stimulus change the state of control nodes;
- the stability and the controllability properties of a subnetwork depends from the existence of loops in the subnetwork: a loop may be related to the activation of metabolic cycles in the cell that define the cell behavior;
- each node of a subnetwork may represent the state of a gene that is connected and regulates the activation of other genes;
- the cell differentiation mechanism is defined by the competition of different subnetworks that interact in a inhibitory way;

The complexity of the model is not a fundamental issue since we want to point out universal behaviors: first of all the existence of bistability or bifurcation phenomena for simple model and the definition of control parameters.

4.2 The model

Here we studied the model dynamics in different situations using mathematical methods. The main idea is understand the dynamics of the models to point out the universal properties that are robust and could explain the experimental data. The biological meaning of control parameters is a fundamental task to apply the model to predict the results of new experiments.

We consider a physical system that can be described by a weighted interaction network among nodes that can assume different dynamical states (in the case of a gene network the states $\sigma \in \{0, 1\}$ and we have models similar to spin models). The interaction structure is defined by signed adjacency matrix $A_{ij} \in \{-1, 0, 1\}$ where the sign refers to a cooperative or antagonist interaction between the connected nodes. In the simplest case, we introduce a stochastic dynamics using the probability $p_i(\sigma, t)$ that the node i is in the state σ (we assume $\sigma > 0$) at time t: in a deterministic approach $p_i(\sigma, t) = \delta(\sigma - \sigma(t))$ to denote that the node assume the state $\sigma = 1$ with probability one. The evolution of a deterministic model can be described by the equation

$$\sigma_i(t+1) = \Phi_i(\sigma(t)) = \Theta\left(\sum_j A_{ij}\sigma_j(t)\right) \tag{4.1}$$

where $\Theta(x) \in [0, 1]$ is a threshold sigmoidal function (we assume $A_{ii} = 0$ to avoid self loops).

Remark: the dynamics is a information diffusion on the network. If we consider the linear system

$$\zeta_i(t+1) = \sum_j A_{ij}\zeta_j(t)$$

where ζ_i are non negative integers we have an equivalent dynamics since $\sigma_i = \Theta(\zeta_i)$ and it is possible to study the linear system to derive some properties of the initial system. For example the relaxation time to the solution $\sigma_i = 1 \; \forall \; i$ is for a given initial condition $\sigma_j^0 = \delta_{jk}$ is $t = n$ such that the matrix A^n has positive entries along the whole k-th column. This means that for each node i there is a walk of length n from the initial node k to i.

We also assume a cause-effect relation so that A_{ij} is a directed graph. The deterministic model is a Hopfield network (each node has at least an input and an output link; the environment nodes has only output links) and one could study the equilibrium states and their stability. An equilibrium condition as follows is characterized as follows: for each i let

$$Q_i(t) = \sum_j A_{ij}\sigma_j(t)$$

then $Q_i > 0$ if $\sigma_i > 0$ and vice versa. Then $A_{ij} \geq 0$(i.e. A_{ij} is a connectivity matrix for a directed network) implies that the non trivial equilibrium is $\sigma_i = 1$: if $\sigma_k = 0$ for

some $k \in K$ then we have

$$\sum_{j \notin K} A_{kj}\sigma_j = 0$$

so $A_{kj} = 0$ for all $j \notin K$ and the network is disconnected. Then we have the trivial solution $\sigma_i = 0$. For each equilibrium solution σ^* we have a stability basin

$$S_{\sigma^*} = \left\{ \sigma \mid \lim_{t\to\infty} \sigma(t) = \sigma^* \right\}$$

If S_{σ^*} defined neighborhood of σ^* the solution is stable or if $S_{\sigma^*} = \{\sigma^*\}$ the solution completely unstable. The stability of the origin depends on the existence of a Ljapounov function: let introduce the network activity

$$\Sigma(t) = \sum_i \sigma_i(t) = \sum_i \Theta(Q_i(t-1)) \geq \Sigma(t-1)$$

since if each node has at least one input link, $A_{ij} = 1$ implies $\sigma_j(t-1) \Rightarrow \sigma_i(t) = 1$ and the activity cannot decrease. The solution $\sigma_i = 0$ is completely unstable. If there would exists an equilibrium solution with $\sigma_k = 0$ for some k then we define S_A the set of nodes s.t.

$$i \in S_A \quad \Rightarrow \quad \sigma_i = 0$$

(obviously $\sigma_k \in S_A$). Let $S_{\bar{A}}$ the complement of S_A, the network dynamics implies

$$0 = \sum_j A_{ij}\sigma_j = \sum_{j \notin S_A} A_{ij}\sigma_j = 0 \qquad \text{if} \quad i \in S_A$$

so that $A_{ij} = 0$ if $i \in S_A$ and $j \in S_{\bar{A}}$: i.e. there is not a cause-effect connection between $S_{\bar{A}}$ and S_A and the state $\sigma_i = 1$ for $i \in S_{\bar{A}}$ is an equilibrium state. Therefore we have as many equilibrium states as many partitions S_A and $S_{\bar{A}}$ there exist such that S_A triggers the activity of $S_{\bar{A}}$ but not vice versa. For any initial condition $\sigma_i(0) = \delta_{ik}$ the possible evolution are a periodic orbit or an equilibrium state: one can detect all the equilibrium conditions by σ^* by the condition

$$\sigma_i^* = 1 \quad \text{if} \quad \sigma_i(t) = 1 \quad \text{for some} \quad t \geq 0$$

The equilibrium states are a semigroup: let σ^a and σ^b two equilibrium states the

$$\sigma^a \cup \sigma^b = \sigma^c$$

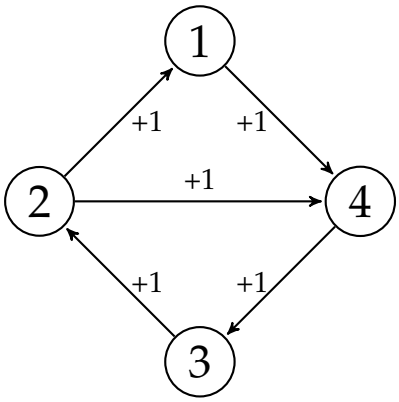

Figure 4.1: *Example of random boolean network.*

is still an equilibrium. An example: if there exist a one directional loop γ in the network and there is no output link from γ to the remaining nodes of the network then $\sigma_i = 1$ for any $i \in \gamma$ is an equilibrium. If the loop is simple (each node has a one input link and one output link) the equilibrium is neutral since any change $\sigma_i = 1 \to \sigma_i = 0$ creates a periodic orbit (the total activity is constant). But if we we add a link to the loop then we get a stable solution since a single node can trigger the activity of two nodes and the equilibrium is an attractive stationary state (see Figure 4.1). If a node is accidentally set to zero this anomaly propagates in the loop, until it reaches the node 4 where it is annihilated by the activity of the node (2). The average lifetime of a single perturbation is the average path length to propagate to the node (4) from the initial node (therefore it depends from the loop length or in case of presence of many loops, the average path length is computed considering independent loops).

The boolean network models the propagation of information. By studying the stability problem of the solution $\sigma_i = 1$ it is convenient to introduce the dual dynamics:

$$\sigma_i^c(t+1) = \Theta\left(\prod_{j\sim i} A_{ij}\sigma_j^c(t)\right) = \prod_{j\sim i} A_{ij}\sigma_j^c(t)$$

where $\sigma_i^c = 1 - \sigma_i$ is the dual state of the node and the product is restricted to the nodes connected to i ($A_{ij} \neq 0$): i.e. the node (4) takes the state $\sigma^c = 1$ only if both

the nodes (1) and (2) in that state at previous time. This dynamics is valid for any configuration of the network and the state $\sigma^c = 1$ moves on the network until it reaches an absorbing state for which

$$\prod_{j\sim i} \sigma_j^c(t) = 0 \quad \forall\, i$$

For a given stable equilibrium σ^γ state associated to a loop γ any environmental perturbation that set to zero a activity of a node will destroy the equilibrium after a time equal to the number of the loop nodes minus one. For example in the figure there are two loops $((1) \to (2) \to (3) \to (4))$ and $((2) \to (4) \to (3))$ if we set to zero the node (4) after three iterations all the nodes will be in the zero state. The two loops are not independent since one loops contains the other). On the contrary if we set to zero the node (1) one loop remains active. This remark allows to introduce the concept of *control node*: a node is a control node if its state is able to force the state of the whole network. The effect of a thermal bath could be introduced by assuming that the state of a node is defined as random variable that takes value $\sigma_i(t) = 1$ with probability $p_i(t)$ where

$$p_i(t+1) = \Theta_T\left(\sum_j A_{ij}\sigma_j(t)\right) \tag{4.2}$$

and $\Theta_T(x)$ is a logistic function

$$\Theta_T(x) = \frac{1}{2}\left(1 + \mathrm{tgh}(x/T - \epsilon)\right)$$

where ϵ measures to tendency of the network to be in the idle state when no stimulus is present. The logistic function is a generic sigmoidal function we do not expect that the specific form of $\Theta_T(x)$ is critical for the results.

Remark: since the values of x are quantized to integer in any case, if $\epsilon > T^{-1}$ the idle state is statistically attractive so ϵ could define a critical temperature for the network activation. We recover the deterministic dynamics for $T \to 0$. As a stochastic process we have a Markov process (since the realization of the variable $\sigma_i(t+1)$ depends only on the present state $\sigma_j(t)$ of the network [33]. The dynamics (4.2) is a Markov field: the realization of the variable σ_i depends only from the present state of the network (and not from past states) and only from the states of the connected

nodes $A_{ij} \neq 0$. The last condition (Markov field) means that the realizations of $\sigma_i(t)$ and $\sigma_j(t-1)$ are independent if the nodes are not connected. The transition probabilities depend from the state of the network and one derives the average dynamics

$$< \sigma_i > (t+1) = p_i(t+1) = \left\langle \Theta_T \left(\sum_j A_{ij} \sigma_j(t) \right) \right\rangle \simeq \Theta_T \left(\sum_j A_{ij} p_j(t) \right) \tag{4.3}$$

Then we have two possibilities: if the total average network activity tends to increase

$$\bar{\Sigma}(t+1) = \sum_i p_i(t+1) = \sum_i \Theta_T \left(\sum_j A_{ij} p_j(t) \right) > \bar{\Sigma}(t) \tag{4.4}$$

the equilibrium solution $\sigma_i = 1$ is attractive, on the contrary we have a an average tendency to decrease the network activity. The situation is illustrated in the fig.

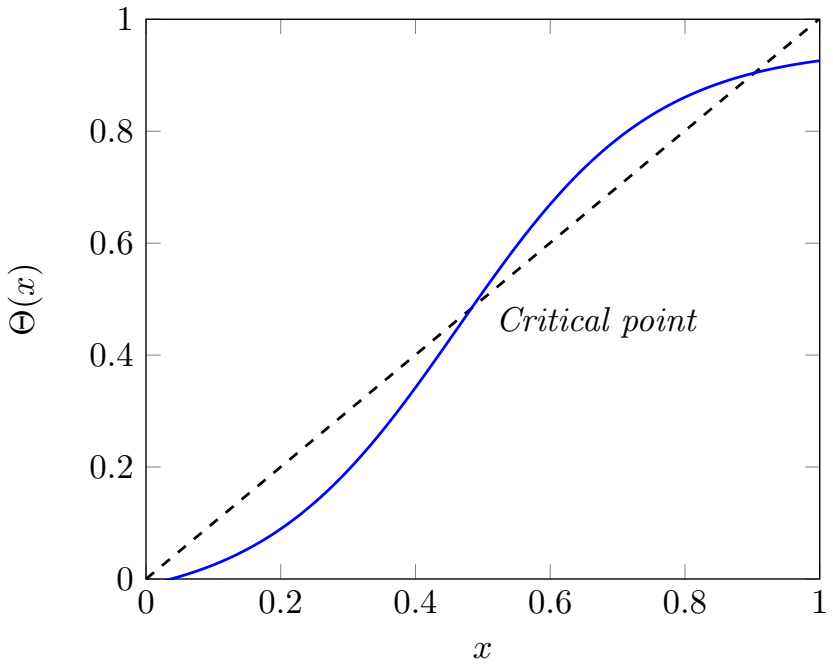

Figure 4.2: *Possible behavior for the condition (4.4); the units are arbitrary and scale with the network dimension.*

Remark: the mean field approximation apply when the $\Theta_T(x)$ can be approximated by a linear function locally: i.e. the fluctuations are small enough to approximate the

function by a linear function in the whole fluctuation range. This is certainly not true when we have fat tail fluctuations.
Except for a small initial region, the condition (4.4) can be satisfied up to a critical value of the network activity Σ (if the temperature is not too big), so that the average activity tends to increase. But if the activity is below the critical value then the network activity tend to decrease and the stability of the solution $\sigma_i = 1$ is lost. A connected network tends to be more stable since the quantities $\sum_j A_{ij}\sigma_j$ increase. This picture is clearly an approximation since we neglect the fluctuation effects: if the fluctuations are big (this depends also on the connectivity matrix) we may have a fast transition between the two possible regime and a correction of the critical value. The critical vale is a consequence of the sigmoidal behavior of the $\Theta_T(x)$ function and its depends on the temperature and on the ϵ values. In presence of fluctuations and of two dynamical regimes (active and non active) we expect that the network activity may switch from one regime to another with a characteristic time scale (Kramer transition rate Theory, see Chapter 5). The transition may be triggered by large fluctuations that are both consequence of rare events (in such a case the probability should be exponentially small with respect the activity) but also depend on the network structure (the presence of hub nodes that can change the activity of many nodes amplifies the effect of small fluctuations (i.e. the change of the hub node state) and may introduce fat tail statistic in the fluctuation distribution).

4.3 Double cluster networks

Let us consider the existence of competitive networks (see Figure 4.3) that are linked by inhibitory links: if the first network is in an exited state the second network should be completely switched off for a stable equilibrium.

If we start with all the node states set to one we create a frustrated situation, otherwise the network choose one of the two possible stable states. In such a case the presence of an environmental noise could induce the transition to one state to another. An external forcing breaks the symmetry.

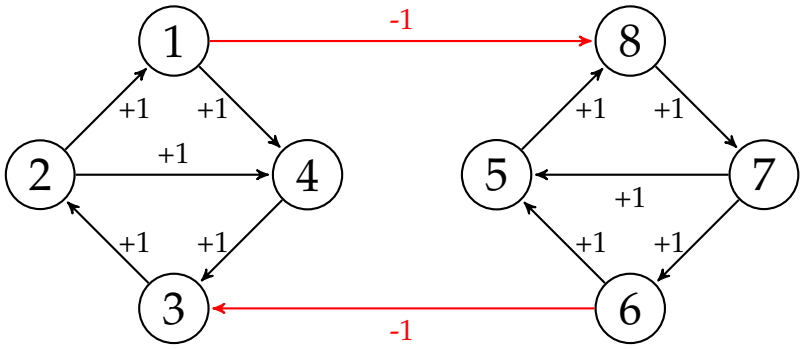

Figure 4.3: *Example of a network composed by two competitive subnetworks.*

We expect a transition phase as a function of the temperature: increasing the temperature the node states tends to be independent, but under a threshold the system should choose a stationary state.

The system can be generalized to consider the interactions of different cooperative networks (possibly with different internal structure) that are connected by inhibitory links (in the case of connection with excitatory links we join the subnetwork in a single one). We can introduce a metadynamics where $\nu_k(t)$ is the state of the k subnetwork and we have a relation

$$\nu_k(t+\Delta t) - \nu_k(t) = \phi(\nu_k(t)) - \gamma \left(H_{kj}\nu_j(t)\right) \tag{4.5}$$

where $H_{hk} \geq 0$ is an inhibitory connectivity matrix. $\phi(\nu_k(t))$ describes the tendency of the cluster to increase its activity and γ the average decreasing of the activity due to the presence of other clusters. This is an effective equation: ν_k should describe the cluster activity (i.e. it could be the time-average activity of the nodes assuming that the sub-network could be considered in a stationary state). Indeed the evolution time scale Δt could be assumed $\Delta t \gg 1$ so that the sub-network states are relaxed to a stationary states. The structure of attraction basins of the stable states could be related to a potential in the state space if

$$\nu_k(t+1) - \nu_k(t) = -\frac{\partial}{\partial \nu_k}\left[\frac{\gamma}{2}\sum_{ij}\nu_i H_{ij}\nu_j + \sum_j V(\nu_j)\right]$$

where

$$\phi(\nu) = -\frac{\partial V}{\partial \nu}$$

Since $\phi(\nu) \geq 0$ $V(\nu)$ is increasing. Then we introduce the energy

$$E = \frac{\gamma}{2} \sum_{ij} \nu_i H_{ij} \nu_j + \sum_j V(\nu_j)$$

and the equilibrium are the critical points of the energy. Moreover

$$\begin{aligned} E(t+1) - E(t) &\simeq (\nu_k(t+1) - \nu_k(t)) \frac{\partial}{\partial \nu_k} \left[\frac{\gamma}{2} \sum_{ij} \nu_i(t) H_{ij} \nu_j(t) + \sum_j V(\nu_j(t)) \right] \\ &= -\frac{1}{2} \frac{\partial}{\partial \nu_k} \left[\frac{\gamma}{2} \sum_{ij} \nu_i(t) H_{ij} \nu_j(t) + \sum_j V(\nu_j(t)) \right]^2 \end{aligned}$$

Therefore the energy is a Ljapounov function and the system equilibria are defined by the critical points of the Energy function corresponding to local minima and maxima. Remark: the existence of the Energy implies that H_{ij} is symmetric negative defined

$$\frac{\partial^2 E}{\partial \nu_j \partial \nu_i} = \frac{\partial^2 E}{\partial \nu_i \partial \nu_j}$$

The stochastic effect has to be introduce but it is possible a thermodynamics approach and a thermodynamics equilibrium exists according to the Maxwell-Boltzmann distribution and the detailed balance condition. This means that the whole network does not satisfy this condition, but the metadynamic network realizes a reversible Markov process.

In the next Chapter we introduce Kramer Transition Rate Theory for stochastic processes.

Chapter 5

Kramer Transition Rate Theory

Kramer theory gives us the ingredients to estimate the form of the potential for our model of Gene Regulatory Networks. The concept is an ensemble of particles in a thermal bath which can lay in two different stationary states. The idea is to find the shape of the potential of the system measuring the transition rates of the particles from one stationary state to the other. Let's make a briefly introduction on this theory.

5.1 Transition rates

Let's consider an ensemble of particles in a thermal bath:
We can describe the system with the Smoluchowski equation:

$$dx = -V'(x)dt + \sqrt{2T}dw_t \tag{5.1}$$

where $V(x)$ is a double well potential (see figure 5.1) with x_a and x_c local minima separated by a saddle point x_b: To compute the transition probability from the left well to the right well we consider the stationary solution of the Fokker-Planck equation:

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}V'(x)\rho + T\frac{\partial^2}{\partial x^2}\rho \tag{5.2}$$

with the boundary condition that we have a source in the point $x_- < x_a$ and an absorbing boundary at the point $x_+ > x_c$. We assume that the temperature T is much

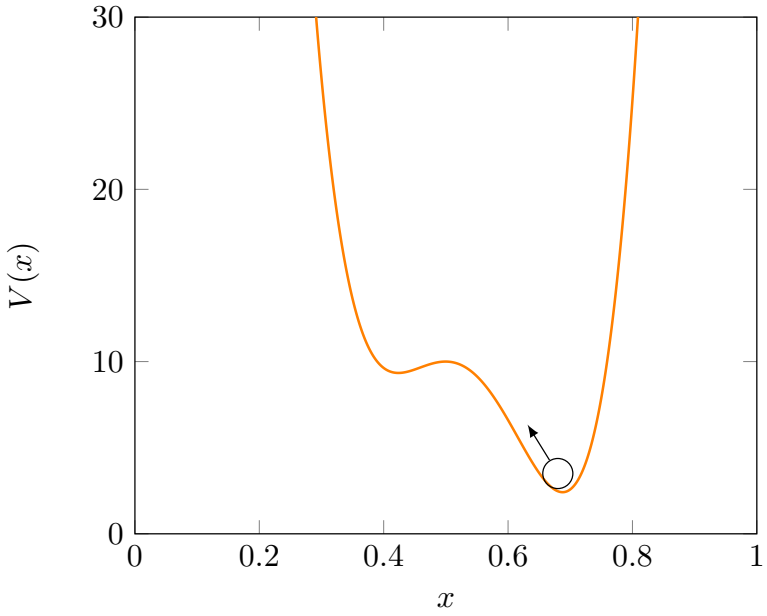

Figure 5.1: *Example of double-well potential.*

les of the potential barrier $V_b - V_a$ and we look for a solution that reduces to the form:

$$\rho \sim e^{-\frac{V(x)}{T}}$$

in the vicinity of the point x_a that vanishes at the absorbing barrier and gives a constant current J between the two wells. Let

$$\rho(x) = C(x)e^{-\frac{V(x)}{T}} \tag{5.3}$$

and the particle density current $-J = V'(x)\rho + \frac{\partial \rho}{\partial x}T$ reads

$$V'(x)C(x)e^{-\frac{V}{T}} + C'(x)Te^{-\frac{V}{T}} - V'(x)C(x)e^{-\frac{V}{T}} = -J$$

so that:

$$C'(x) = -\frac{J}{T}e^{\frac{V}{T}}$$

and we integrate with the condition $C(x_+) = 0$:

$$C(x) = \frac{J}{T}\int_x^{x_+} e^{\frac{V(y)}{T}}dy$$

The distribution reads:

$$\rho(x) = \frac{J}{T} e^{-\frac{V(x)}{T}} \int_x^{x_+} e^{\frac{V(y)}{T}} dy.$$

When$x \simeq x_a$ the integral

$$\int_x^{x_+} e^{\frac{V(y)}{T}} dy$$

is stationary since $V'(x_a) = 0$ and $\rho(x)$ has the behavior of $e^{-\frac{V(x)}{T}}$. Then we have to compute the number of particles n_a in the left well

$$n_a = \frac{J}{T} \int_{-\infty}^{x_b} e^{-\frac{V(x)}{T}} \int_x^{x_+} e^{\frac{V(y)}{T}} dy dx$$

We approximate the integrals using the saddle point method:

$$e^{\frac{V(y)}{T}} \simeq e^{\frac{V_b}{T} - \frac{\omega_b^2}{2T}(y-x_b)^2}$$

$$e^{\frac{V(x)}{T}} \simeq e^{\frac{V_a}{T} - \frac{\omega_a^2}{2T}(x-x_a)^2}$$

so we compute:

$$\frac{n_a}{J} = \frac{1}{T} e^{\frac{V_b - V_a}{T}} \int_{-\infty}^{x_b} dx e^{-\frac{\omega_a^2}{2T}(x-x_a)^2} \int_x^{x_+} e^{-\frac{\omega_b^2}{2T}(y-x_b)^2} dy$$

Then we can extend both the integrals between $-\infty$ and $+\infty$:

$$\int_{-\infty}^{+\infty} dx e^{-\frac{\omega_a^2}{2}(x-x_a)^2} = \frac{\sqrt{2\pi T}}{\omega_a} \frac{n_a}{J} = \frac{2\pi}{\omega_a \omega_b} e^{\frac{V_b - V_a}{T}}$$

And we find the transition probability rate:

$$k_{a\to c} = \frac{J}{n_a} \simeq \frac{\omega_a \omega_b}{2\pi} e^{\frac{V_b - V_a}{T}}$$

The fact that k_c does not depend on x is due to the fast relaxation scale time inside the potential well with respect to the escape time scale from the potential well and to the quasi-stationary distribution that concentrates the particles near the local minimal point where the effect of the potential can be approximated by a parabolic potential. In this systems the transition probability k_c can be used to estimate the potential.

For our model, Kramer theory will be useful to measure a possible potential for Gene Regulatory Networks. We will consider an ensemble of networks made by two different clusters which inhibit each other, in the sense that the activity of one cluster causes a decrease of the activity of the second one.

5.2 Our model related to Kramer Theory

Since the random nature of these networks, we use an ensemble approach for the analysis of the model. We consider an ensemble of networks made by two different inhibitory clusters. Given two clusters per network, in which the clusters have different size and different average number links per node, we expect that the average activity will have two possible stationary states, and so the system will undergo a bistable situation. As explained in Chapter 5, a bistable system can be represented by a stochastic process in which the stationary probability distribution satisfies the Fokker-Plank equation. In our case, in which we have networks made by two clusters, we can consider as stochastic variable the total activity of the networks:

$$I(t) = \nu_2(t) - \nu_1(t) \tag{5.4}$$

with $\nu_k(t) \in [0,1]$ and $I(t) \in [-1,1]$. So the Fokker-Planck equation associated to the system will be:

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial I} V'(I)\rho + T\frac{\partial^2}{\partial I^2}\rho \tag{5.5}$$

where we assume that the temperature T is much less of the potential barrier:

$$T \ll V_b - V_a$$

We expect that the network activity, and so the potential, should depend on the size of the clusters (the number of nodes of the clusters N) and on the average number of links in the clusters K:

$$V = V(N, K).$$

To estimate the form of the potential, we can measure the transition rates of the system, in which during the evolution the network activity can move from a cluster to the other. In this way, we can see the distribution of the activities of the networks during the discrete evolution. If we choose the two clusters with different size and different number of incoming links K, we expect to see that starting from the initial condition in which only the second cluster is active the system relaxes on the cluster with bigger size and links. Transition rates of activity from one cluster to the other can be registered during the simulationz and after the relaxation we expect to have a distribution of the logarithm of the rates which have the shape of a double well potential.

Chapter 6

Model Analysis and Numerical Simulations

In this Chapter we explain the starting implementation and analysis of the model following biological considerations. We will start with some numerical analysis of the model for a single cluster, and then we will analyse the evolution of an ensemble of networks made by two clusters, which hinibit each others. The theoretical model has been analyzed making computer simulations using Python, using the well known libraries for Data Science: *Numpy, Pandas,Scipy* and *NetworkX*, the most famous libraries for network analysis. The repository is available on GitHub.

6.1 Single Cluster - Implementation

For this work, we decided to create a class *Random Network* to have a random boolean graph as an object, with its nodes and links, represented by a numpy array and a numpy matrix, respectively.

Every random network is a directed graph and is built to avoid self-loops, this means to create a random, boolean, and non-symmetric adjacency matrix with null trace.

For example, let's consider the network in Figure 6.1: the adjacency matrix A of

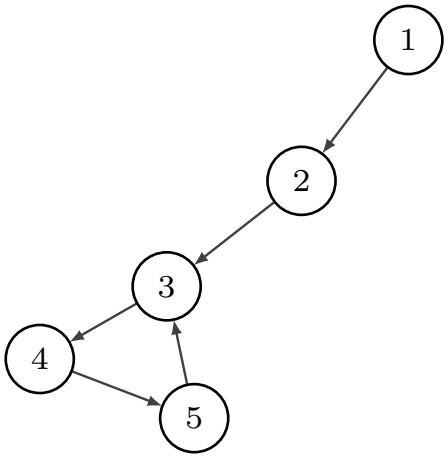

Figure 6.1: *Example of a simple random network constructed by our Python code.*

this network will be constructed as follows:

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Now, we have to decide how many links per node and how many nodes have to be used for this type of network. As shown in Chapter 4, the discrete time evolution of the network is given by the equation:

$$\sigma_i(t+1) = \Theta\left(\sum_j A_{ij}\sigma_j(t)\right)$$

where A is the connectivity matrix of the network.

So this means that each node which has at least one incoming link with a node which is active, in the next step this node will be active. At each time step we can measure the average activity of the network, which is the average number of nodes with the value $\sigma_i = 1$.

6.1.1 Number of links

To choose the number of links, we considered the average activity of the network, after many iterations of the discrete evolution. In Figure 6.2 we can see that the average discrete evolution of 100 different realizations of RBNs, with increasing size. In

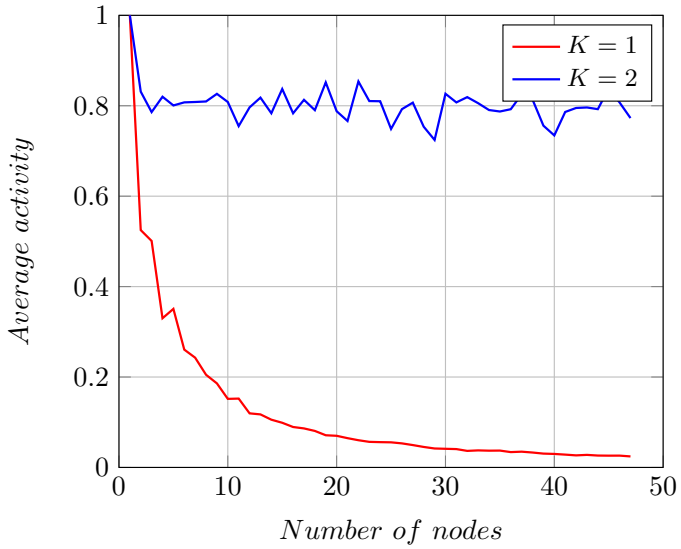

Figure 6.2: *Plot of the average activity of the nodes with network of increasing size. In the case of* $K = 1$ *(i.e. the average number of incoming link for each network is one), the average activity decreases exponentially with the size of the network; in the case of* $K = 2$ *instead, the average activity of the nodes remains stable with the network size.*

the case of $K = 1$, i.e. the average number of incoming links for each network is one, the average activity decreases exponentially with the size of the network; this is due to the fact that with only one link per node, the network will be composed by a big loop, and so during the evolution there will be in average only one active node and all the other nodes remains inactive. In the case of $K = 2$ instead, the average activity of the nodes remains stable with networks of increasing size. This is due to the fact that (differently from the case of $K = 1$ networks) each node of the network can be potentially active during the discrete evolution of the system.

Moreover, we can analyze the average number of outgoing links depending on the network size: In Figure 6.3, we can see that the average number of outgoing links

tends to the parameter K.

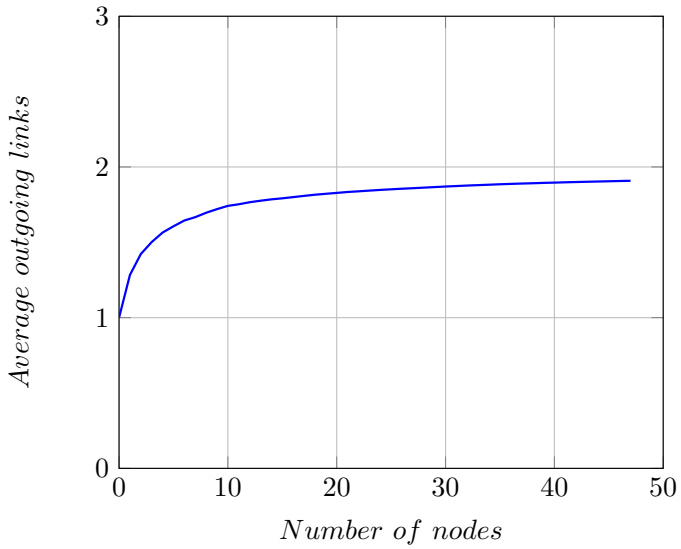

Figure 6.3: *Plot of the number of the outgoing links depending on the network size. We made an average of 100 networks with* $N = 10$ *and* $K = 2$*. We can see that the average tends to the parameter K.*

6.1.2 Control nodes

In simple random networks one can always find the number of loops. Loops determine the complexity of the networks, and are important for the construction of the model. In fact from loops one can find the *control nodes* of the network, which are the nodes that their state are able to force the state of the whole network, and are the nodes with maximum connectivity. Control nodes determine the whole activity and stability of the network. In figure 6.4 we can see that control nodes, independetely from the network size, control the 80% of the whole loops in the network.

The fact that each network is constructed in order ot have each node with a fixed average number of incoming links equal to 2, means that each network constructed is connected, so it is possible find always a node which is inside the maximum number of loops in the network. We remark that control nodes are different from the nodes with maximum betweenness centrality: in fact betweenness centrality measures the

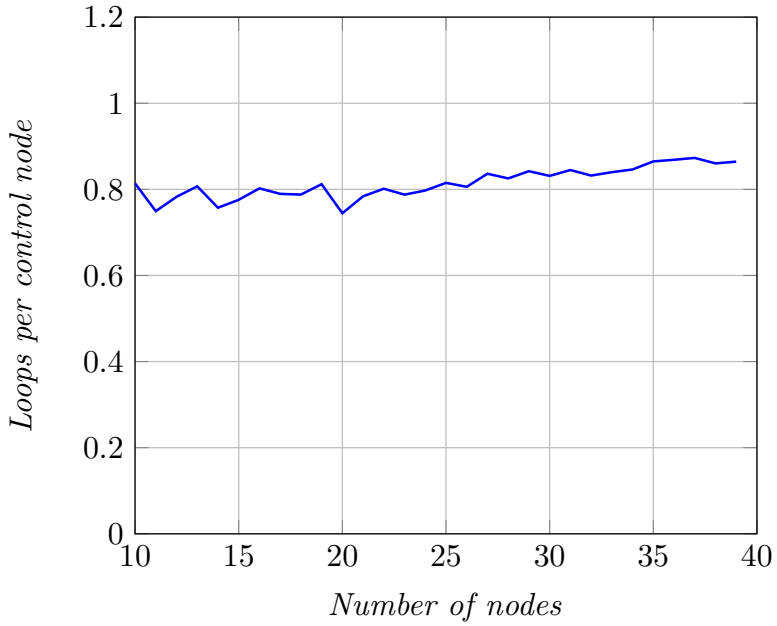

Figure 6.4: *Plot of the loops which belong to the control node, depending on the network size. Average values from* 100 *generated networks. Number of incoming links* $K = 2$.

number of the shortest paths that pass through the nodes; control nodes instead, are the nodes which are contained in the maximum number of loops in the network, not only the independents loops, and this means that they are not the nodes with maximum betweennees centrality.

Control nodes are a key point for networks with multiple clusters, beacuse since they have the capability to control the whole activity of the cluster, multple cluster networks will be connected through their control nodes. Moreover, these links will be negative, in order to have inhibition between two different clusters.

6.2 Noise

The second thing to evaluate is the effect of the noise on the evolution of the network and the difference between noise and parametric noise, where parametric noise refers to the noise which infers in the links and not on the nodes. To add noise to the system, during the discrete evolution of the network, at each time step there is a

probability p for the node or for the link to be turned off. In Figure 6.5 we can see the behavior of the average activity of the network depending on the amount of noise added. The plot is similar to a sigmoidal functions in both of the cases. In the case of noise added to the links the average activity results to be bigger than the case of noise added to the nodes. This is reasonable in the sense that since the number of links is less than the number of nodes so the effect of the noise among the links is less. In the next section we will consider only the noise of the system, keeping fixed the links.

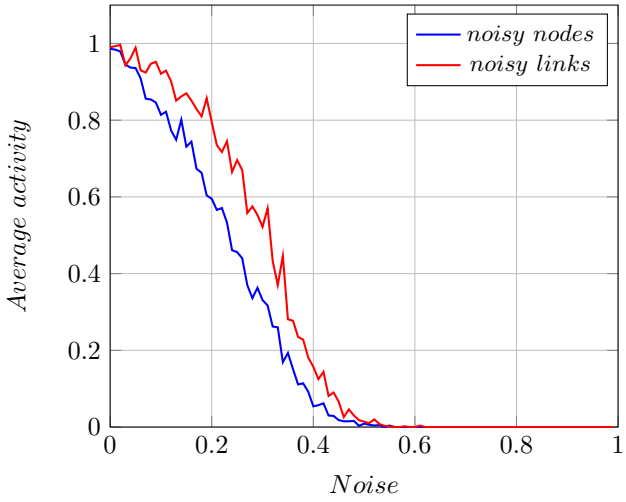

Figure 6.5: *Plot of the effect of the noise on the average activity on the network. In blue the noise works on the nodes on the network, while in red the noise works on the links. Number of nodes for each network: 10; Number of realizations for each value of noise: 100;*

In 6.6 we can see the evolution of single-cluster networks subjected to noise. The noise on the network is much more effective on the activity if we set to zero the links belonging to the control nodes. So we can say that the average activity depends on the activity of the control node. Also, we see that the network is very sensitive to the value of noise.

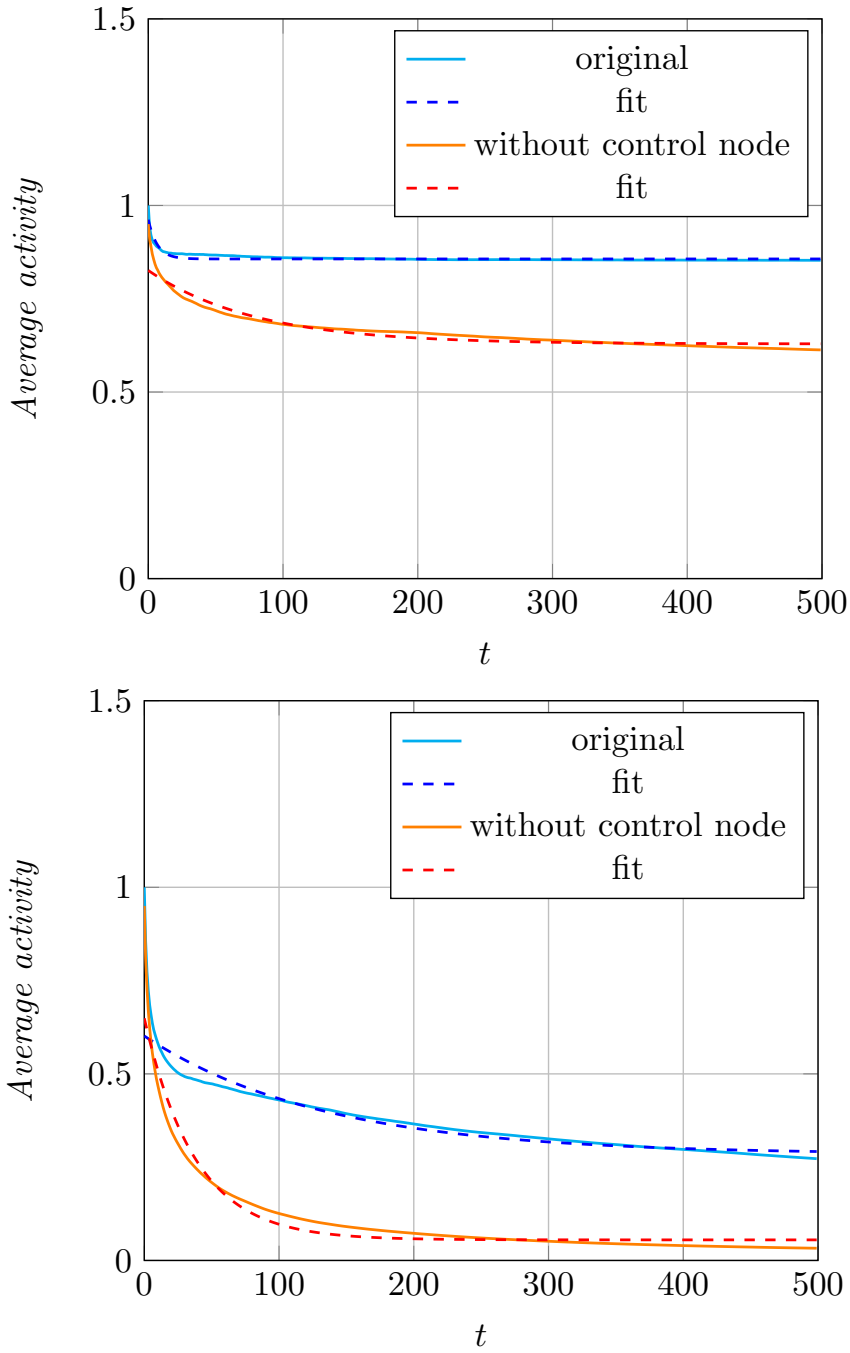

Figure 6.6: *Plot of the effect of the noise on the average activity on the network with and without control nodes. In the first plot(top) the noise is set to* 0.1. *In the last plot (bottom) the noise is set to* 0.3. *We can see that in both cases, the average activity of the network decreases exponentially if we set to zero all the links of the control nodes: The average activity depends on the activity of the control node. Fitting function:* $f(x) = ae^{-bx}$.

6.3 Double Cluster Network

Now we are ready to analyse the dynamics of a network composed by two different clusters. As explained in Chapter 4, two clusters will be connected by negative links with value -1, so we will have two clusters which hinibit the activity of the opposite cluster. The construction of this double-cluster network is made simply by the construction of two different random networks which have each control node which connected negatively to the control node of the other cluster. In Figure 6.7 we can see an example of this type of networks.

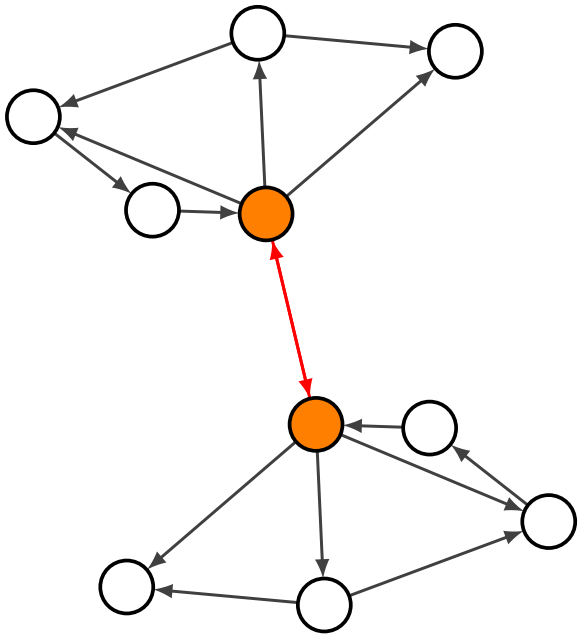

Figure 6.7: *Example of double-cluster random network. Control nodes of the two clusters are connected negatively (red).*

From double-cluster networks, what we want to find is a dynamics similar to Kramer's theory explained in Chapter 5. We expect to see that the activity of one cluster hinibits the activity of the other and viceversa: Adding noise to the system there will be a bistable dynamics. The number of nodes, and the number of links of the cluster are the parameters which determine the robustness of the network. If the number of nodes of one cluster is much higher than the number of nodes of the sec-

ond cluster, it will be difficult for the second cluster to hinibit the first one. In this sense, we will have that the first cluster will remain stable even if it's disturbed by noise and by the other cluster. In this system we add an other type of noise, what we called *envinromental noise*. envinromental noise represent the probability for a node to be turned "on" at a certain t. In this way, during the evolution of the system a cluster which is completely off can be actived by the envinromental noise. The nodes subjected by envinromental noise are the second control nodes of the clusters, i.e., the second node of the cluster which belongs to the maximum number of loops.

6.4 Discrete evolution

The discrete time evolution of the system for a double-cluster network is the same of the single-cluster network evolution, with the difference that the two clusters hinibit each other through the negative links of the control nodes. So the system dynamics follows equation (4.1), but we recall equation (4.5) in which we consider the average activity of the network $\nu_k(t)$ given by the equation for the metadynamics:

$$\nu_k(t+\Delta t) - \nu_k(t) = \phi(\nu_k(t)) - \gamma \left(H_{kj}\nu_j(t)\right)$$

In figure 6.8 we can see an example of the metadynamics of the system of a double-cluster network. We can see the transition of the activity from one cluster to the other during the time evolution.

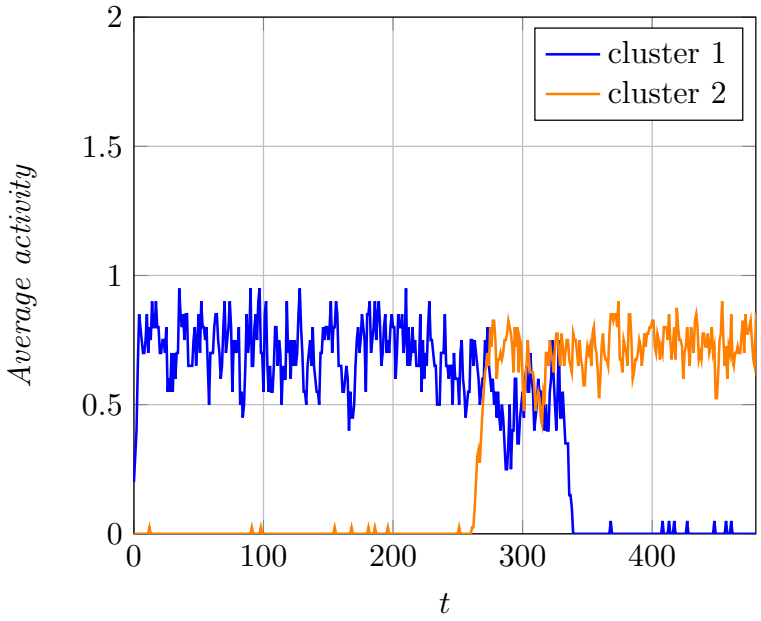

Figure 6.8: *Example of evolution of the activity of a single double-cluster network. We can see the transition of the activity due to noise and envinromental noise.*

6.5 Transition Rates

To estimate the form of the potential, we can measure the transition rates of the system using an ensemble approach, in which during the evolution of the system avery network activity can move from a cluster to the other. We made a MonteCarlo simulation of 1000 double-cluster networks. Each network undergoes a discrete evolution subject to noise and envinromental noise. To measure the activity transition, we can define the total activity of the network:

$$I(t) = \nu_2(t) - \nu_1(t) \tag{6.1}$$

with $\nu_k(t) \in [0,1]$ and $I(t) \in [-1,1]$. In this way, we can see the distribution of the activities of the networks during the discrete evolution. If we choose the two clusters with different size and different number of incoming links K, we can see that the system relaxes on the cluster with bigger size and links. In figure 6.9 we can see an example of the average activity of the system in the stationary state: We choose two

cluster of different size and linking: the first cluster is made $N = 20$ nodes and $K = 2$ incoming links per node. While the second cluster is made by $N = 10$ and $K = 1$. Starting from an initial condition in which only the second cluster is active:

$$\rho_0(I) = \delta(I - 1)$$

the system relaxes to the state in which only the first cluster is active and the second is totally inactive.

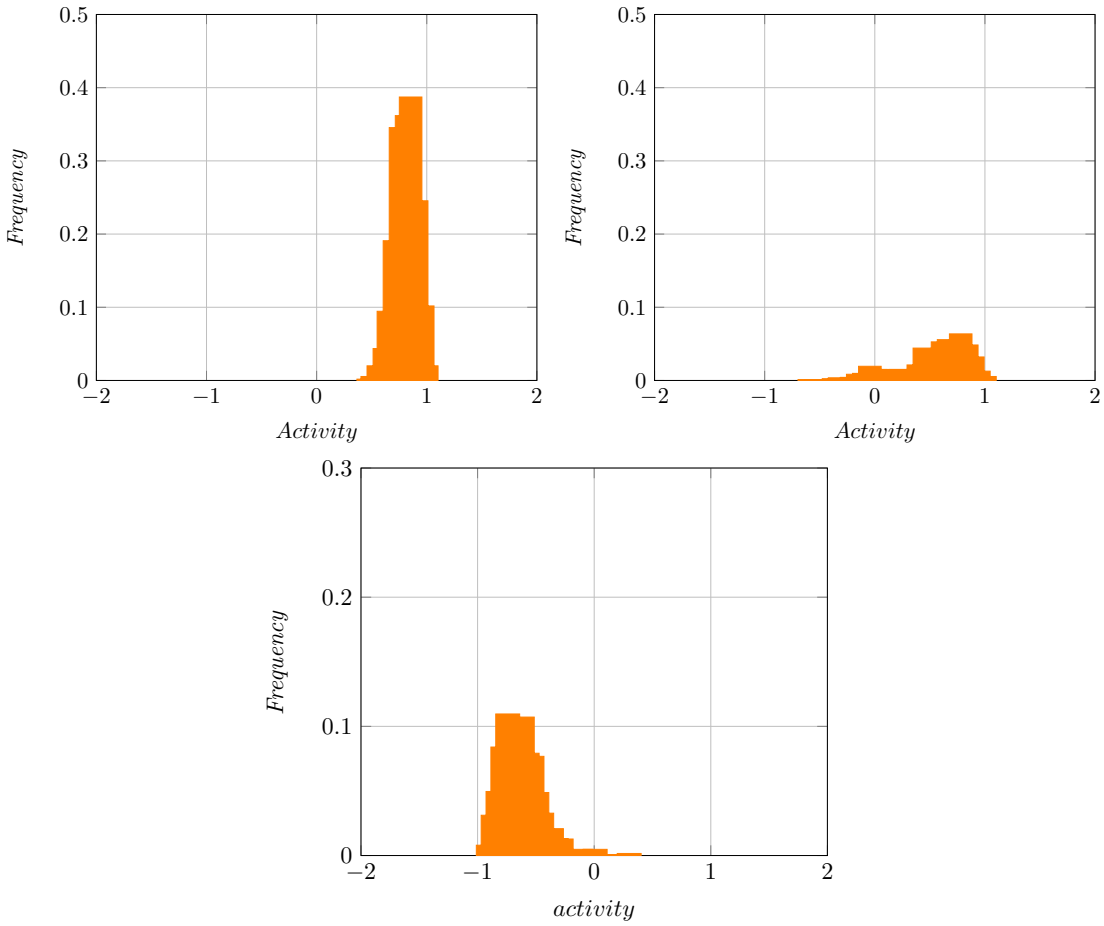

Figure 6.9: *Evolution of the histogram of activity of the network. The first cluster is constructed with $N = 20$ and $K = 2$, while the second cluster is constructed with $N = 10$ and $K = 1$. With initial conditions in which only the second cluster is active, during the evolution the system relaxes to the state in which only the first cluster is active and the second cluster is inactive.*

In figure 6.10, we can see the transition times of the system from one cluster to the other. According to Kramer theory, the transition rate k_c can estimate the form of the potential V. In this case, We can see the shape of the histogram which can estimate the form of the potential. We can clearly see that the histogram presents two different valleys which should correspond to the wells of the potential. The fit was made with a polynomial fitting of order 4, using the Python library *Numpy*.

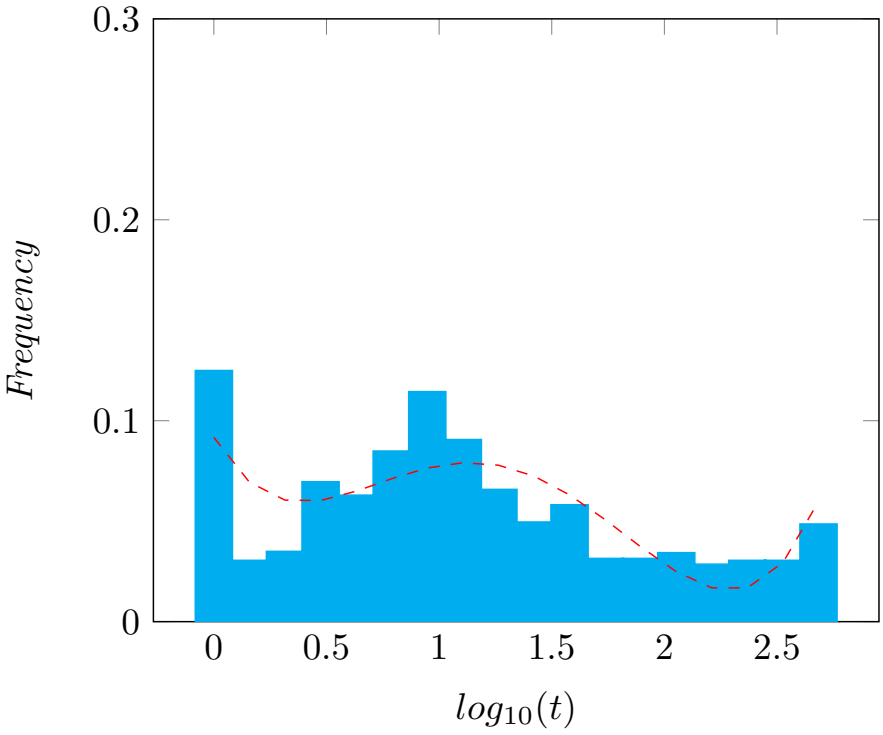

Figure 6.10: *Histogram of transition rates of activity from a cluster to one other. We can see the shape of the histogram which can estimate the form of the potential. We can clearly see that the histogram presents two different valleys which should correspond to the wells of the potential. Fitting function:* $f(x) = ax^4 + bx^3 + cx^2 + dx + e$.

The presence of a potential in this model can be useful to understand the properties of this models. It suggests that the proces of differentiation really can be governed by a dynamical pathway governed by different networks of different size and linkage. It may suggest that the production of progenitor cells, with respect to differentiated

cells are regulated by smaller networks, and going ahead to differentiation, regulatory networks becomes bigger and more linked.

The results found suggest that modeling Gene Regulatory Networks with Random Boolean Networks may be the right way towards a deeper understanding in the process of cell differentiation.

Conclusions

In this work we proposed a theoretical model for cell differentiation. Since this biological process is governed by Gene Regulatory Networks, these networks can be modelled by Random Boolean Networks, in which each gene can be represented by node which can be ”on” or ”off”. The process of differentiation is a multistable dynamical system, and involves different type of cells, but all of these start from one unique type of cell: the stem cells. The complexity of this process lays in the fact that cells ”can” decide if transforming in one type of cell with respect one other, and for this reason seems pheasible that if a network for a type of cell is active, it may inhibit the network for a different type of cell. The proposed model can give an estimate of a fitness potential for two different type of cells, according to the theory of Waddington. This process concerns also the birth of cancer cells: cancers cell can be governed by a specific type of regulatory network, which often is inactive, but due to some external stimuli it can be activated and gives an irreversible process of production of cancer cells. This can be seen as a local minima of the Waddington potential, which is impossible to escape. The role of noise in this model is crucial, but it is well known that biological process are indeed very dependent to external noise.

Today, the need arises for tools capable of unraveling the functionality of genes based on the analysis of microarray measurements. Modeling genetic interactions by means of genetic network models provides a methodology to infer functional relatioships between genes, for a deeper understanding on biological processes. Future studies on this model may focus on some aspects:

- control nodes, which are a key point for the dynamics of this model; knowing and governing control nodes may be a real improvement in understanding ge-

netic networks.

- average number of links K: the average degree of the network is crucial for the stability of itself.
- study the reversibility of the process: may be studied if these network can produce a reversible process or a completely irreversible process, where in the process of cell differentiation it is difficult to say.

Bibliography

[1] Waddington CH, *The strategy of the genes: a discussion of some aspects of theoretical biology*. London: Allen and Unwin, (1957)

[2] Cameron P. Gallivan, Honglei Ren and Elizabeth L. Read,*Analysis of Single-Cell Gene Pair Coexpression Landscapes by Stochastic Kinetic Modeling Reveals Gene-Pair Interactions in Development*,doi: 10.3389/fgene.2019.01387 ,(2019)

[3] Jifan Shi, Tiejun Li , Luonan Chen, Kazuyuki Aihara,*Quantifying pluripotency landscape of cell differentiation from scRNA-seq data by continuous birth-death process*,https://doi.org/10.1371/journal. pcbi.1007488 ,(2019)

[4] Jin Wang, Kun Zhang, Li Xu, and Erkang Wang ,*Quantifying the Waddington landscape and biological paths for development and differentiation*, https://doi.org/10.1073/pnas.1017017108 ,(2011)

[5] S. Huang, I. Ernberg, S. Kauffman,*Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective*, doi:10.1016/j.semcdb.2009.07.003, (2009)

[6] S. A. Kauffman, *Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets*, J. Theoret. Biol. (1969)

[7] Drossel B.,*Random Boolean Networks*,arXiv:0706.3351 ,(2008)

[8] Aldana et al., *Boolean Dynamics with Random Couplings*, arXiv:nlin/0204062v2 , (2002)

[9] R.Serra, M. Villani, A. Barbieri, S.A. Kauffman, A. Colacci,*On the dynamics of random Boolean networks subject to noise: Attractors, ergodic sets and cell types.*,J Theor Biol 265: 185–193, (2010)

[10] M. Villani, A. Barbieri, R. Serra,*A Dynamical Model of Genetic Networks for Cell Differentiation*, doi:10.1371/journal.pone.0017703.g001,(2011)

[11] Latchman D., *Inhibitory transcription factors*, doi:10.1016/1357-2725(96)00039-8, (1996)

[12] M. Ali Al-Radhawi , Nithin S. Kumar, Eduardo D. Sontag , Domitilla Del Vecchio, *Stochastic multistationarity in a model of the hematopoietic stem cell differentiation network*,doi:10.1109/cdc.2018.8619300, (2018)

[13] E. Davidson, M. Levine, *Gene Regulatory Networks*, doi:10.1073/pnas.0502024102, (2005)

[14] Chen L et al., *Biomolecular networks: methods and applications in systems biology*, Wiley, Hoboken ,(2009)

[15] Peccoud, J. and Ycart, B., *Markovian Modelling of Gene Products Synthesis.*, https://doi.org/10.1006/tpbi.1995.1027, (1995)

[16] T.B. Kepler and T.C. Elston, *Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations*, 10.1016/S0006-3495(01)75949-8 , (2001)

[17] J.M. Pedraza, J. Paulsson, *Effects of molecular memory and bursting on fluctuations in gene expression*, 10.1126/science.1144331, (2008).

[18] Y. Sasai *Cytosystems dynamics in self-organization of tissue architecture*, https://doi.org/10.1038/nature11859, (2013)

[19] B. Zhang and P. G. Wolynes, *Stem cell differentiation as a many-body problem*, https://doi.org/10.1073/pnas.1408561111, (2014)

[20] Zhou et al., *Fast Pyrolysis of Glucose-Based Carbohydrates withAdded NaCl Part 2: Validation and Evaluation of theMechanistic Model*,DOI 10.1002/aic.15107, (2016)

[21] A. Wuensche, *Genomic regulation modeled as a network with basins of attraction*, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, (1998)

[22] S. A. Kauffman, *Investigations*, Oxford University Press, (2000)

[23] MacArthur S. et al., *Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions*,doi:10.1186/gb-2009-10-7-r80 (2009)

[24] S. A. Kauffman: J. Theor. Biol., 44, Physica D, 10, 145 (1984)

[25] S. Kauffman, *A proposal for using the ensemble approach to understand genetic regulatory networks*,Journal of Theoretical Biology 230 (2004) 581–590 ,(2004)

[26] B. Derrida, *Random Networks of Automata: A Simple Annealed Approximat ion.*, (1985)

[27] B. Derrida and H.Flyvbjerg, *The random map model: a disordere model with deterministic dynamics*, J.Physique, (1987)

[28] B. Derrida, *Spin glasses, random boolean networks and simple models of evolution*

[29] R. V. Solè, B. Loque, *Phase transitions and antichaos in generalized Kauffman networks*,Physics Letters,(1994)

[30] C.W. Gardiner, *Handbook of Stochastic Methods*, Springer, (1985)

[31] Sui Huang , Ingemar Ernberg, and Stuart Kauffman,*Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective*, DOI:10.1016/j.semcdb.2009.07.003, (2009)

[32] Atefeh Taherian Fard and Mark A. Ragan, *Modeling the Attractor Landscape of Disease Progression: a Network-Based Approach*,DOI: 10.3389/fgene.2017.00048, (2017)

[33] V. Kampen, *Stochastic processes in phycics and chemistry*,ISBN 0444893490 ,(1992)

[34] M. Rybarsch and S. Bornholdt,*On the dangers of Boolean networks: Activity dependent criticality and threshold networks not faithful to biology*, arXiv:1012.3287v1. (2010)

[35] C. Gershenso,*Introduction to Random Boolean Networks*, arXiv:nlin/0408006,(2004)

[36] J. Park and M. E. J. Newman,*The statistical mechanics of networks*, DOI: 10.1103/PhysRevE.70.066117 (2004)

[37] A. Rèka and A-L. Barabàsi, *Statistical mechanics of complex networks*, Reviewes of modern physics, Volume 74,(2002)

[38] N. Masuda , M. A. Porter, R. Lambiotte ,*Random walks and diffusion on networks*,Physics Reports 716–717 1–58,(2017)

[39] Sui Huang, Yan-Ping Guo, Gillian May, Tariq Enver,*Bifurcation dynamics in lineage-commitment in bipotent progenitor cells*,Developmental Biology 305, (2007)

[40] Xin Kang, Chunhe Li,*Landscape inferred from gene expression data governs pluripotency in embryonic stem cells*,Computational and Structural Biotechnology Journal 18 (2020) 366–374,(2020)