COHERENT ORGANIZATION IN GENE REGULATORY NETWORKS

by

NEŞE ARAL

A Thesis Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Science

in

Physics

Koç University

SEPTEMBER, 2015

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral thesis by

NEŞE ARAL

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Asst. Prof. Dr. Alkan Kabakçıoğlu

_____

Prof. Dr. Özgür Müstecaplıoğlu

_____

Assoc. Prof. Dr. Deniz Yuret

_____

Prof. Dr. Nihat Berker

_____

Prof. Dr. Muhittin Mungan

Date: _____

توانا بود هر که دانا بود
ز دانش دل پیر برنا بود

فردوسی

Powerful is he who has knowledge
Young is an old heart that holds knowledge

Ferdowsi

ABSTRACT

With mounting information on details of biological networks, the interplay between structure and function has become a fertile field of research. For example, a power-law distribution of neighbor counts (or scale-freeness, an all-pervasive feature across biological and artificial networks) has been shown to promote robustness against internal or external influences, the abundance of certain substructures in intracellular networks can be linked with special tasks, the number of dynamical attractors (in correspondence with the number of cell types) can be predicted from the number of nodes and edges, etc.

In this thesis, we introduce a novel measure of coherence for dynamic networks which encompass both activatory and repressory interactions. We investigate the degree of coherence, i.e., the unity of purpose between the simultaneously active regulatory agents with a common target, in gene regulatory networks responsible for cell cycle and cell differentiation in several organisms. Upon comparison with carefully chosen null-model ensembles, we demonstrate that these biological networks are, in fact, biased towards high coherence. We argue that this skewness is selected for by evolutionary pressures (akin to Hebbian selection in neuronal connections), under a demand for economical consumption of materials and energy available to the cell. The dependence of coherence on various structural parameters, such as the mean number of inputs per node and the activatory/repressory interaction ratio, as well as on dynamically determined quantities, such as the basin size and the number of expressed genes, is also investigated.

ÖZETÇE

Biyolojik ağlar hakkındaki bilgiler arttıkça, ağların yapı ve fonksiyonları arasındaki etkileşim aktif bir araştırma alanı haline geldi. Örneğin, biyolojik ve yapay ağlardaki komşu sayısı dağılımının bir kuvvet fonksiyonu ile verilmesi (ölçeksizlik), bu ağların iç ve dış etkilere karşı dayanıklı olmasını sağlamaktadır. Ayrıca ağın içindeki bazı daha küçük altağlar belirli görevlerin yerine getirilmesinde etkilidir. Gen düzenleme ağlarında hücre tiplerine karşılık gelen çekicilerin sayısı da ağın düğüm ve kenar sayılarına bağlı olarak tahmin edilebilmektedir.

Bu tezde, dinamik ağlardaki uyum için hem aktive edici hem baskılayıcı etkileşimleri kapsayan yeni bir ölçü tanımlanmaktadır. Biyolojik ağlardaki uyum, aynı anda aktif olan ve ortak bir hedef paylaşan düğümlerin aynı şekilde hareket etmesi olarak tanımlanır. Çalışmada hücre döngüsü ve hücre başkalaşmasından sorumlu altı değişik gen düzenleme ağında uyumluluk derecesi incelenmektedir. Bu ağların rassal ağlarla karşılaştırılması sonucu biyolojik ağların yüksek uyumluluk sağlayacak biçimde şekillendiği gözlemlenmektedir. Bu özelliğin, sinir ağlarındaki Hebb seçilimine benzer bir işleyiş altında, hücrenin sahip olduğu malzeme ve enerjiyi en ekonomik biçimde kullanmasını sağlayan evrimsel bir süreç sonucu ortaya çıkmış olduğu düşünülmektedir. Uyumluluğun, bir düğümün ortalama komşu sayısı ve aktive edici ve baskılayıcı etkileşimlerin oranı gibi yapısal parametrelere bağlılığı araştırılmaktadır. Bunun dışında dinamik özelliklerin sonucu olan çekici havuzu ve çekicilerdeki aktif düğüm sayısı oranına bağlılık da incelenmektedir.

ACKNOWLEDGMENTS

I cannot overemphasize my deep gratitude to my advisor Assoc. Prof. Dr. Alkan Kabakçıoğlu. His excellent guidance, extensive understanding and tolerance are essential for the completion of my Ph.D. studies during which he cared about my work more than his. His encouragement and motivation in the periods of time when I was not as productive as I should have been, enabled me to embrace my scientific career with more enthusiasm. He has not only made a huge intellectual impact on me by sharing his invaluable time, intelligence, erudition and wisdom wholeheartedly, but he has also influenced me with his personality by showing that one can be hardworking, trustworthy, considerate, forgiving and modest at the same time. It is my fortune to have him as a life-long advisor.

Regarding their support and help throughout my studies I owe sincere thanks and gratitude to my thesis committee and jury members, Assoc. Prof. Dr. Deniz Yuret, Prof. Dr. Özgür Müstecaplıoğlu, Prof. Dr. Muhittin Mungan and especially to Prof. Dr. Nihat Berker who has been mentoring me for 14 years.

I was very lucky to share my office life with outstanding co-workers who provided me with a peaceful and enjoyable working environment. For their support, understanding and immeasurable friendship, it is a pleasure to me to send my heartfelt thanks to Neslihan Oflaz, Ahad Khaleghi, Seçil Gürkan Benli, Yasa Ekşioğlu Özok, Duygu Can, Sasan Hajizadeh, Murat Öztürk, Batuhan Kav, Emre Doruk Önal, Güneş Aydındoğan, Deniz Türkpençe and Nur Aslan.

I am indebted to my family for their support, trust and understanding throughout the long years I have spent for my education. I am gratefully aware that my parents Orhan and Ferihan Aral are working harder than me to provide the extraordinary opportunities I have. Big thanks go also to my sister Işıl Aral for always helping me

and cheering me up as I was writing the thesis and rehearsing its defense. I look forward to her doctoral graduation, certainly with a thesis superior to this one.

Most importantly, the biggest of the thanks are deserved by my fiancé A. Barış Sözener, who has always comforted me with his endless love. His support, help and understanding during my stressful, busy and tired times were remarkable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

GRN	Gene Regulatory Network
TF	Transcription Factor
PIN	Protein-protein Interaction Network
RG	Random Graph
RN	Random Network
YCC	Yeast Cell Cycle
RNA	Ribonucleic acid

Chapter 1

PRELIMINARIES

1.1 Introduction

Due to the discoveries and new techniques in genetics and molecular biology in the 20^{th} century, large scale data sets are gathered about the molecular basis of life. The amount of these experimental data is so abundant that mathematical and computational methods must be used to analyze them. The mathematical approach of network analysis has been found to be a successful tool to investigate data which contains biological entities along with their experimentally verified interactions between them, because the emerging collective behavior of these systems heavily depend on the interactions between the components rather than the components themselves. The most important applications of representing these so called "complex systems" in biology as networks are the food web, ecological, metabolic, signaling, neural, protein-protein interaction and gene regulatory networks.

Before the implementation in biology, network science uncovered many structural properties of networks describing real life phenomena, such as the small-world effect in social networks, enabling most nodes to be reached from others in few steps or the scale-free property of the Internet, indicating that its degree distribution follows a power law, increasing the robustness against failure. For biological networks similar design principles were observed that are not present in randomly generated networks such as local motif structures which enable the networks to perform specific tasks [12].

By uncovering these structural features we can understand the mechanisms underlying biological processes better because it has been shown that there is a relation

between the topological properties of a network and its dynamics. The unique processes that take place in a single cell or in the whole organism are only possible with specific network structures that join the components of the system together and determine how they interact with each other. Therefore investigating complex networks is an ongoing research in systems biology.

1.2 Motivation and Purpose of the Study

A big group of biological networks consists of nodes representing genes or their products, proteins. The directed edges between the nodes signify physical or genetic interactions which can be of different types such as transcriptional regulation, phosphorylation, ubiquitination etc. with varying strengths or time scales which can be modeled by assigning weights to the edges. But the key disparity lies in their end effect being activatory or inhibitory. Hence an edge in the network is labeled with a plus sign if the protein at the source node raises the expression level of the protein at the target node. And a minus sign for an edge is an indication that the source protein/gene has the effect of lowering the concentration of the target protein in the cell.

To quantify the architectures of biological networks and to differ them from the random networks various successful methods exist [4], the most important ones among them being degree distribution, clustering coefficient, robustness measurements [13, 14], modularity [15, 16] and motif detection [12]. Excluding motif detection, these quantities take only the direction of the interactions into account but not their signs. The motifs, small subnetworks of the original network, take cognizance of these features but they only give information about local structures instead of the network as a whole. There are also other techniques considering local signed subgraphs to explore the relation between the topology and the properties of the attractors, for example multistationarity [17]. But it has been pointed out that there is the necessity of constructing measures including these signs for the entire network and the concept of "regulation entropy" was an attempt to fill this void in the literature [18].

In this work we also try to make a contribution in this direction by proposing a new quantity called "coherent activity" in the biological networks indicating the unisonous behavior of agents sharing a common target. We show that biological networks which have to operate in an energy-efficient way tend to have higher values for coherence than the random networks having similar topology. We aim to find out the necessary structural features leading to this behavior that distinguish biological networks from the others.

1.3 Methods

The methods of analysis of networks can be grouped in two classes which are logical and continuous models. From the biological experiments scientists can collect continuous data such as reaction rates, concentration of molecules in the cell, cell cycle duration. A well grounded theory which can be compared with the experimental data in detail can only be established with a continuous model with parameters over a continuous time scale. Ordinary differential equations or regulated flux balance analysis are two main examples of this kind [19].

But when the focus is not a detailed quantitative description of the process and the actual time-scales are not important, where we want to acquire a basic understanding of the dynamics and functions of the network, discrete logical models which reduce the computational effort considerably can be used. Gaussian [20], Bayesian [21] and Boolean [8, 6]network models are successful methods which are used to construct and analyze genetic networks from micro-array data.

For our analysis we chose six different gene regulatory networks from literature, three cell cycle [6, 7, 8] and three differentiation networks [9, 10, 11] which had successfully implemented the Boolean approach introduced by Kauffman [22, 23]. These networks, where the nodes representing genes can have only the values 0 and 1 indicating that the corresponding gene is inactive or active respectively, enable us the investigate their structure easily rather than the dynamics of time-sequenced events. For these model networks we use the synchronous update scheme with the Boolean

functions as given in the original papers. To highlight their high levels of "coherent activity" we compare them with the ensembles of random networks with the same topology where the edges are shuffled but for each node the in- and outdegrees of the both positive and negative interactions are preserved separately. We also observe the behavior of the networks with the same number of nodes as the biological ones but with different number of edges and different ratios of positive and negative regulations.

Chapter 2

NETWORKS AND BIOLOGICAL SYSTEMS

2.1 Networks: Fundamental Concepts

Networks consist of nodes, representing entities in the investigated system. If there is an interaction between any two nodes they are connected by an edge, which is undirected if the interaction is symmetrical. The protein-protein interaction networks are in this class, since the binding of two proteins is a symmetrical process. But if the effects of these nodes on the other one are different from each other, there is a directed edge from the source node to the target node, where we can identify the source node as being a neighbor of the target node. This class contains among others the gene regulatory networks where one gene can affect another gene without being influenced from the other. The different types of interactions with nonidentical strengths can be represented by assigning a weight to the corresponding edge.

Mathematically the networks can be represented by an $n \times n$ adjacency matrix A where n is the number of nodes in the network which are labeled as $1, 2, \cdots, n$ and the nonzero elements in the matrix indicate the presence of an edge. In the simplest case the elements of the adjacency matrix can be given as

$$A_{ij} = \begin{cases} 1, & \text{if node } j \text{ has an effect on node } i \\ 0, & \text{if there is no interaction between nodes } i \text{ and } j. \end{cases} \tag{2.1}$$

If the interactions are symmetrical and the underlying graph is undirected the adjacency matrix is also symmetrical. In the gene regulatory networks, which are the main topic of this thesis, the interactions are not reciprocal so the adjacency matrix is not symmetric.

If the interactions between the nodes are not of the same type, we can assign weights to the edges which can be any real number. These weights can stand for the strength, frequency or speed of an interaction. The sign of an edge's weight may imply the behavior of the corresponding interaction. For example in the gene regulatory networks we distinguish mainly between two types of interactions, activating and inhibiting, represented by positive and negative weights respectively. Hence in general an adjacency matrix for a gene regulatory network with n genes consists of elements in the form

$$A_{ij} = \begin{cases} A_{ij} \in \mathbb{R}^+, & \text{if node } j \text{ activates node } i \\ A_{ij} \in \mathbb{R}^-, & \text{if node } j \text{ inhibits node } i \\ 0, & \text{if there is no interaction between nodes } i \text{ and } j. \end{cases} \tag{2.2}$$

We can also assign values for the nodes whose change with time can be determined by using a function, depending on the values of the neighboring nodes. When the nodes stand for genes or proteins the value for the corresponding node at time t can indicate the activity or concentration level at that time. The functions iterating node values in time can have many different forms, but biological experiments reveal important aspects of the possible structures and enable scientists to construct the correct class of functions yielding results compatible with investigated biological process.

Since the neighbors play an important role in the dynamics of the system the pattern of interactions is crucial for the network to function properly. To investigate these patterns and to uncover the underlying architecture of the networks we can use topological quantifiers which can refer to global or local structures in the network. The design principles depend on the desired task of the networks but it has been shown that many real life networks have some common topological features that are not seen in the random networks. As it is the case for many social and man-made technological networks, biological networks have their own architecture. Some of these topological quantifiers important for biological networks are explained below.

2.2 Networks: Topological Quantifiers

Here we list global and local network measures that characterize complex networks. The values in biological networks are different from the random networks [24]. They are responsible for the networks to perform specific tasks under changing conditions and environments.

2.2.1 Global Architectural Measures:

***Degree distribution**:*

The degree (k) is a quantity defined for the nodes and it gives the number of the edges connected to the node. If the network is directed we can distinguish between in-degree (k_{in}) and out-degree (k_{out}) of a node indicating the numbers of incoming and outgoing edges respectively. Once the adjacency matrix for a $n \times n$ network is known, for the node i they can be calculated as

$$k_{in}^{(i)} = \sum_{j=1}^{n} |\mathrm{sgn}(A_{ij})| \tag{2.3}$$

and

$$k_{in}^{(i)} = \sum_{j=1}^{n} \left|\mathrm{sgn}(A'_{ij})\right|, \tag{2.4}$$

where A' is the transposed matrix and $\mathrm{sgn}(x)$ is the sign function. The mean in- and outdegrees($\bar{k}_{in}, \bar{k}_{out}$) for the whole network can then be written as

$$\bar{k}_{in} = \frac{1}{n}\sum_{s=1}^{n} k_{in}^{(s)} = \frac{1}{n}\sum_{s=1}^{n} k_{out}^{(s)} = \bar{k}_{out}, \tag{2.5}$$

since the number of incoming edges must be equal to the outgoing edges in the network.

For an undirected network the degree distribution $P(k)$ gives the probability that a node, chosen randomly, has the degree k. This notion can be extended for a directed network having two separate in- and out-degree distributions $P_{in}(k), P_{in}(k)$ or we can

define the two dimensional distribution $P(j,k)$ of a node for the probability of having in-degree j and out-degree k. The in-degree and out-degree distributions might be correlated and their joint probability distribution will give the real degree distribution.

In the literature degree distributions are calculated mainly for undirected networks, namely graphs. In 1959 Erdős and Rényi showed that for randomly constructed graphs, where the probability to find an edge between two nodes is constant and independent from the presence of other edges, the degree distribution is binomial, becoming Poissonian in the limit of large n [25]. But it has been shown [26] that many real life phenomena are based on networks with non-Poissonian degree distributions. The two mostly encountered distributions leading to different topologies are exponential and power law distributions. These three main classes are summarized in Table 2.1.

Degree distribution in biological networks:

Many biological networks like metabolic, protein-protein interaction and gene regulatory networks have the property that their degree distributions follow a power-law in the form $P(k) = Ck^{-\gamma}$ where the value of γ lies typically between 2 and 3 [27, 28]. The examination of 43 cellular metabolic networks from different domains (archaea, bacteria, eukaryota) reveal their scale-free nature where most substrates take part only in one or two reaction chains as a small number of them is involved in numerous reactions [4, 29, 30]. The power-law distribution enables the network to have a big number of nodes having few edges and a small number of nodes having a lot of neighbors, which are called hubs. Such networks are also called scale-free networks because like power-law functions they do not exhibit a typical scale or a typical node.

One important benefit of the scale-free degree distribution is the robustness that makes the network resistant to failures in the systems. When some nodes are canceled from the network, for example as a result of genetic mutation, the system can perform its function properly. This is because the likelihood of deleting a hub is very small since there are not many in the network so it does not get separated when some nodes

Topology	$P(k)$	Example
Poisson	$e^{-2m/n}\frac{(2m/n)^k}{k!}$	Erdős and Rényi random graphs [25]
Exponential	$Ce^{-\alpha k}$	North American Power Grid [32]
Power-law	$Ck^{-\gamma}$	Metabolic networks [29], Internet [24]

Table 2.1: Three main classes of degree distributions $P(k)$ in random and real world undirected networks. Here n and m are node and edge numbers respectively. C, α and γ are network dependent constant parameters.

are removed. But a deletion of a hub can have fatal effect for the cell.

The emergence of scale-free topology is related to the fact that in the evolutionary process new nodes are added, for example as a result of gene duplication, to the system over a long time scale and these new nodes are frequently joined to the nodes with already high degrees, a phenomenon called "preferential attachment".

Although there are many examples of scale-free networks in biology there are counter-examples such as the transcriptional network of the budding yeast *S.cerevisiae* exhibiting exponential degree distribution [31].

Clustering coefficient:

The local clustering coefficient defined for the node i measures the tendency of its neighbors to be also connected with each other, such that they build a complete graph between themselves. Denoting the degree of the node i as k_i the clustering coefficient C_i can be written as the ratio of the number of existing edges between the node and its neighbors E_i to the number of maximum possible edges [3]. Hence we have

$$C_i = \frac{AE_i}{k_i(k_i - 1)}, \tag{2.6}$$

where A is 1 or 2 if the network is directed or undirected respectively. Nodes having degree less than two are defined to have $C_i = 0$. The global clustering coefficient is

then the average over all the $n$nodes in the network which is given as

$$C = \frac{1}{n} \sum_{i}^{n} C_i, \tag{2.7}$$

which can be interpreted as the probability a pair of neighbors of a randomly selected node are mutually neighbors. Networks with a high degree of cluster coefficient tend to build big clusters where the nodes are connected very densely. These highly connected clusters are also called modules.

Clustering coefficient in biological networks:

When compared to random networks biological networks tend to have higher values for clustering coefficient arising from their modular nature [15]. In the cell a specific processes is accomplished by a group of highly interacting molecules building the module which is partially independent from the other modules (subgraphs) in the networks. Assembled together, these functional modules governing different tasks create the so called "network of networks". This modularity enables us to investigate each functional module separately as it is the case in this study where small subgraphs responsible only for the cell cycle or differentiation tasks in the cell are analyzed.

Path length:

The geodesic distance l_{ij} is defined as the minimum number of edges connecting node i to node j, where $i \neq j$, along with the properties that $l_{ij} \neq l_{ji}$ and $l_{ij} = l_{ji}$ for directed and undirected networks respectively. Considering a network with size n and consisting only of one cluster we can write the mean geodesic path as the average over all the node pairs as

$$L = \frac{2 \sum_{i \neq j} \sum_{j} l_{ij}}{n(n-1)}. \tag{2.8}$$

Even if there is only one cluster in a directed network it is possible that there is no path connecting node i to node j. In this case the path is defined as having the length infinity and is not included when taking the average.

In general, regular networks, in which all the nodes have the same degree, have relatively high and randomized networks have relatively low mean geodesic path length (L). Another class consisting of irregular but non-random networks are observed to have also low L (between regular and randomized) scaling with the logarithm of the network size. Since the logarithm is a slowly increasing function of its argument it is possible to reach from one node to another just in few steps in spite of a big network size. This phenomenon is called the "*small world effect*" [26].

Path length in biological networks:

For the biological networks it is observed that many of them have even smaller mean geodesic path than those of the randomized networks and they are called "*ultra-small networks*". The metabolic networks including hundreds of nodes have $L \approx 4$ requiring only 3 or 4 reactions to convert one metabolite into another. The same property is observed both in a bacterium and in a multicellular organism [29].

2.2.2 Local Architectural Measures:

Another way of analyzing networks is to look at smaller local structures instead of the whole network. Subgraphs, which consist of a subset of nodes and the interactions between them, are widely studied local architectures. Here we list important examples for such analyses.

Functional Motifs

The subgraphs, whose numbers of presence in the real networks are exceptionally higher than those in randomized networks of the same type, are called as network motifs [1]. They serve as basic components regulating specific functions in the network. Table 2.2 shows that in gene regulatory networks the feed forward loop, which serves as a delay element and helps to construct the modular nature [33], is overrepresented where in the food webs the cascade motif is highly encountered. Technological networks with different functions also have different abundant motifs.

Some examples for the 3-node motifs along with their amounts in real and randomized ensembles are given in Table 2.2. Analyses for motifs containing higher number of nodes are also available.

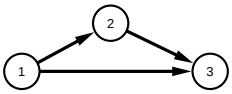
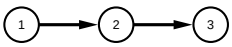
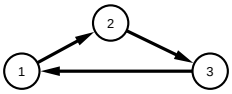
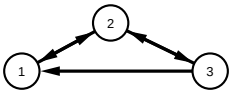
Network	Nodes	Edges	Motif	N_{real}	$N_{rand} \pm SD$
Gene regulation *S. cerevisiae*	685	1052	Feed forward loop 	70	11 ± 4
Food web B. Brook Lake	25	104	Cascade 	181	130 ± 7
Electronic Dig. frac. mult.	512	819	Feedback loop 	40	1 ± 1
World Wide Web nd.edu	3e5	1.5e6	Feedback with 2 dyads 	10^5	2e3 $\pm$ 1e2

Table 2.2: Examples for 3-node network motifs seen in biological and technological networks [1]. N_{real} and N_{rand} are the occurrences in real and randomized networks respectively.

Multistationarity and Cycles:

Multistationarity is the property of a network to have multiple steady states. Much effort has been devoted to local structures and their role in promoting multistationarity, stability and robustness in a background of fluctuating environmental conditions [12, 34, 35, 36]. In gene regulation multistationarity is important for differentiation networks where each different type of cell corresponds to a different point attractor in the system, whereas a cyclic attractor can be crucial for networks modeling the cell cycle process.

The necessary and sufficient conditions for multistationarity and cyclic behavior was investigated firstly by Thomas et al. and then extended further by many others [37, 17, 38, 39, 40, 41]. It has been shown that feedback circuits, where a

node's influence of another one can result in its own regulation, play an important role in the dynamics of complex systems. A circuit (loop) in a network in the form $i \to j \to k \to \cdots y \to i$ is said to be positive if the number of negative edges is even and negative otherwise. The presence of positive circuits in the network is necessary for multistationarity where as negative feedback loops are associated with cyclic or homeostatic behaviors [42].

2.3 Biological networks and their modeling

At first the biological networks were thought to be completely regular or completely random. But the homogeneous degree distributions of Erdős-Rényi (ER) model sets an obstacle to compare them with real networks, since they cannot account for the presence of hubs in many observed networks. Additionally they are also not capable of reflecting the local clustering although they provide the small world property. To overcome the absence of these properties Watts and Strogatz imposed a new model (WS) by starting from a regular circular lattice with a given size and mean degree and increasing the randomness by taking every edge for each nodes consecutively and changing their target nodes with a predefined probability β, resulting in a rewired graph [3], where by high randomness ($\beta = 1$) the ER model is recaptured. A further improvement came from Barabási and Albert [43] based on growth of networks considering the preferential attachment procedure (BA model). To an existing network with N_o nodes a new node is added at one time and the probability of the new node to attach itself to one of the existing nodes i is proportional to the degree of this specific node (k_i). Hence the nodes with higher degrees will be preferred by a newly added node enabling the emergence of hubs. As growth continues and network size increases the BA model displays additionally the scale-free nature. Finally it was shown that hierarchical models can uncover the reality more effectively by also including the modularity of these networks [15]. Further refinements can be found in Table III of Ref. [44]. Fig. 2.1 shows typical examples for the above mentioned model networks for comparison.

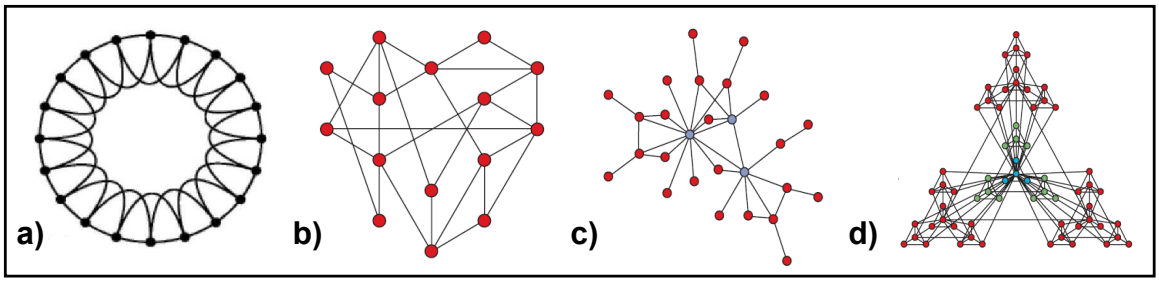


Figure 2.1: Common network models: a) Regular [3] b) Random c) Scale-free d) Hierarchical [4].

It is also worth noting that most biological networks are very sparse indicating that only a small fraction of all possible edges are actually present in the network. For example the transcriptional regulatory network of *Escherichia coli* has 424 nodes but only 519 edges which implies only 0.2% of possible edges are encountered. One of the reasons for this feature is that the interactions are selected throughout the evolution which eliminates edges abolished by mutations and keeps only the useful ones [12].

2.3.1 Common biological networks

In biological systems the level of detail can vary from microscopic (DNA, RNA, proteins, metabolites) and mesoscopic (tissues, organs) to macroscopic (organisms, species) scales but the interactions between the constituents, which also differ in their time-scales, can be modeled by the same formalism using the theory of networks. Especially for the molecular networks the task of finding the involved components was a challenge lasted several decades after the discovery of the DNA in 1953 and only in recent years it was possible to uncover the interactions between them with the help of the technological and computational advances. As it is the case for the genome projects the ultimate aim is to complete the big picture with all of the inherent elements and interactions giving rise to new terms as proteome, interactome, metabolome, regulome *etc.* Although the available data is very huge, for the large part of investigated networks they are incomplete and prone to experimental errors.

Below we list the main biological networks which are exhaustively analyzed and have been shown to possess certain topological features mentioned in 2.2. Considering the subject of the thesis we will concentrate ourselves on molecular intracellular interactions (biochemical networks) with the emphasis being on the gene regulatory networks. For additional information on these and here not mentioned intercellular (vascular or brain networks) or macroscopic networks such as ecological networks (food webs) and others one can refer to [24, 45, 46].

Protein-protein interaction networks (PIN):

The proteins are responsible for the majority of essential functions in the cell. The enzymes catalyze the reactions, transport proteins bind to substrates and move them in or between the cells, structural proteins give the cell its shape yet usually they cannot fulfill these tasks individually and have to interact and cooperate with each other. This interaction can be chemical in which a subgroup of a protein is transferred to another one (e.g. phosphorylation) but the common interaction method is physical where their three dimensional structures enable them to join together to form a protein complex.

The set of all the proteins in an organism is called the "proteome" and the "interactome" is the protein-protein interaction network that comprises all the interactions between the proteins of the whole proteome or a smaller subset of it under consideration such as those only in control of growth, motility, signal transduction etc. Here the nodes represent the proteins and the undirected edges between two proteins indicate that they physically bind to each other. In most cases these networks are unweighted but since the interaction data can be unreliable due to the possible false-positive or false-negative results encountered in experimental procedures (e.g. *two-hybrid screen* [47]) [24], the edges can also have a weight indicating the probability that the corresponding interaction is true-positive [48]. Among others, genome wide protein interaction networks are available for *Saccharomyces cerevisiae* [49], *Drosophila melanogaster* [50], *Caenorhabditis elegans* [51] and *Homo sapiens* [52].

In literature there are very diverse PINs but it has been shown that most of them exhibit the scale-free property, albeit with different exponent in the power law of degree distribution, while it is also observed that the proteins which are essential for the vitality of the cell have higher degrees of connectivity [27]. These sparse networks are actually "small-worlds since the average path length is proportional to the logarithm of the network size [45]. But looking at local structures containing nodes interacting densely with each other but almost separate from the rest of the network it is also possible to find functional modules performing special tasks in the cell hence the functions of proteins can be deduced from the structure of the network [45].

Metabolic networks:

The metabolism of an organism consists of two chemical processes where by *catabolic metabolism* the food is broken down to its simple building blocks and by *anabolic metabolism* the cells these building blocks are joined together to build larger molecules required by the cell for different functions. The metabolic network then consists of all the reactions catalyzed by enzymes and are performed sequentially to convert an initial substrate into an end product. The metabolites, whose complete set creates *the metabolome*, are represented by the nodes and there is a directed from the consumed chemical targeted to the produced molecule. The edges can be weighted by the strength of the biochemical reaction deduced from metabolic flux data [53].

By analyzing metabolic networks from 43 different organisms belonging to separate domains, Jeong et al. [29] were able to show that they all exhibit the scale-free property. As a possible reason it was suggested that this shared global feature is due to the natural selection in evolution which favors robust networks with an inherent tolerance against errors. The metabolic networks are also shown to have hierarchical structure by which certain motifs are overrepresented [54].

Gene regulatory networks:

To understand how the cells function it is crucial to know how the concentration levels of proteins change in time. The expression patterns of proteins are regulated by the interplay of genes, transcriptions factors and small molecules. The interactions between them are responsible for the timely occurrence of events in the cell. Gene regulatory networks (GRNs) are able to model this phenomena, where the nodes represent the genes or their products and the directed edges designate the effect of the influence from the source molecule to the target molecule. The influence can be positive or negative depending on the interaction being activatory or inhibitory respectively. Hence besides having directions the edges are also signed by plus or minus to indicate an increase or decrease of the concentration level of the target protein. If different types of regulation mechanisms are involved the edges can be weighted relative to the strength or speed of the corresponding interaction. The biological mechanism underlying the gene regulation and the way they shape the GRNs are explained in more detail in Section 2.4

The GRNs are sparse networks, where it is estimated that average number of regulator per gene is less than two, a property that is responsible for the robustness of the network against mutations [55]. They are usually scale-free with power-law degree distributions for in- and out degrees but it has been also shown that exponential or mixed behaviors are also possible [31]. The studies concerning the presence of the motifs reveal that depending on the function of the considered cell, different types of motifs are overrepresented. The cells requiring multistability (more than one attractor) like differentiation networks have an excess number of double negative feedback loops with autoregulations, where periodic expression levels for cell cycle networks are enabled by the high numbers of diamond shaped four-node loops of the incoherent type are needed [56].

2.4 Gene Regulation & GRNs

2.4.1 Historical Remarks:

After the discovery of the DNA's molecular structure in 1953 [57], the biochemical mechanisms responsible for the transfer of heredity information and for the production of proteins by means of transcription and translation processes could be explained in detail. It took over 40 years until the whole genome of a free-living organism, the bacterium *Haemophilus influenzae*, could be deciphered in 1995 [58]. In 1996 the budding yeast *Saccharomyces cerevisiae* was the first eukaryote with the whole genome sequence and in the following years the collection of complete genome sequences grew rapidly with the help of advances in experimental and computational disciplines, embracing the human genome in 2003 [59].

The first estimate for the number of genes in human cells had come from theoretical works by Kauffman in 1969, whose formalism on gene regulatory networks we build our study on, where he suggested that the cell replication time correlates with the total number of genes, resulting in 2,000,000 genes for humans [22]. Later experimental studies predicted the actual number as 100,000 in the late 70's. Using this outcome Kauffman proposed that there are 317 different cell types in the human body relying on his theory that the number of different cell types in organisms is proportional to the square-root of the number of genes [60]. At that time this result was compatible with the idea that humans possessed 256 distinguishable types of cells and estimated gene numbers of several organisms were in concordance with their cell types, where cell types are defined as cells sharing the same genotype but having different phenotypes and functions due to the non-identical proteins expressed in each of them.

The number of distinct cell types, which are specialized for specific tasks in the organism, is important because it is a measure for the organism's complexity. But the outcome of the Human Genome Project was a surprise for the scientists, reporting that the number of genes in the human genome was lower than anticipated and similar to different low-level organisms [61]. The fact that the number of genes in human genome

(20,000-25,000) is not significantly higher than that of the roundworm *Caenorhabditis elegans*, which was the first multicellular organism with the complete genome when in 1998 its 20,470 genes were sequenced [62], pointed to the lack of correlation between the number of genes and the number of cell types, considering that *C. elegans* has only 29 where humans are made up of 411 different types of cells [63], a phenomenon called the "C-value paradox".

Soon it was realized that the genome itself cannot be accounted for all the complexity in the cells or organisms. The post-genomic era began and new experimental techniques enabled scientists to completely identify proteomes, transcriptomes and interactomes of organisms. In a large number of organisms the size of the proteome (number of proteins) was larger than the number of genes, which could be explained by post-translational modifications of proteins and *alternative splicing* allowing the production of several distinct proteins from one single gene [64]. And recently it is observed that these explanations are not sufficient and the discrepancies in complexity depend mainly on the timing, location and expression levels of almost the same set of genes. These complex regulatory processes can be analyzed by the wisdom that it is possible to represent them by networks with the nodes being the molecules in the system and the edges signifying the interactions, which is a powerful abstraction of the real world [65]. The investigations of the interactomes, belonging to different organisms, displayed the interdependence of biological complexity and PINs [66] and showed that the human interactome is 3 times bigger than that of the *C. elagans* [61]. So it became obvious that the interactions are more important than the constituents themselves.

2.4.2 Biological Mechanism of Gene Regulation:

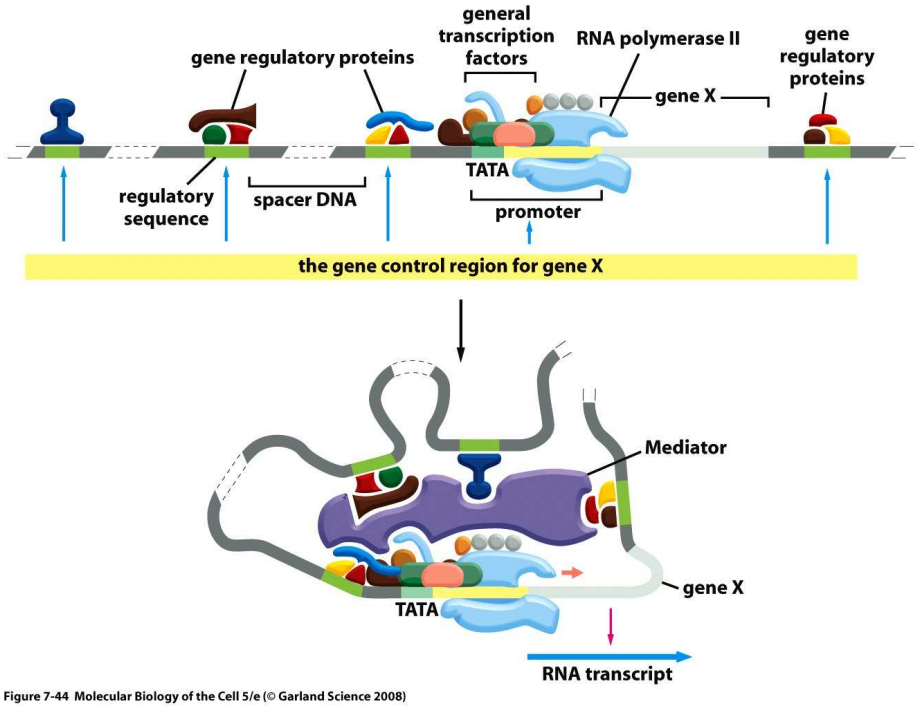
The gene regulation process is the mechanism which enables the cells to differentiate into different cell types by synthesizing distinct proteins in each of them or to produce only the required ones at a particular time depending on the function that must be fulfilled according to the needs of the cell. It controls the transcription rates of genes

at particular time and places. Below we give a simple overview of this complex process in eukaryote for a single cell [5].

The genes are nucleotide sequences on the DNA which code for proteins. The expression of a protein from the gene begins with the *transcription* where the double stranded DNA is unzipped and one strand serves as a template to assemble the complementary RNA (ribonucleic acid) such that the monomers are lined up to represent a portion of genetic information. In the ribosomes (large molecular machines) the information of RNA is translated into a sequence of aminoacids creating the protein, a process called the *translation*. The cell can regulate the rate of transcription and translation along with the activity level of the produced protein.

Besides genes, coding for the proteins, the DNA molecule contains noncoding nucleotid sequences. A small portion of these noncoding sequences are thought to be useless as being residues of evolution and they called *junk* DNA. But a big part is responsible for the regulation of expressions of the genes. The most important noncoding nucleotide sequences are next to the coding gene sequences and these are called the *promoter* regions. The enzyme *RNA polymerase II* is bound to this region to unzip the DNA for transcription. The region at the end of the gene sequence is the *terminator* region and when it is encountered the transcription stops. But the enzyme RNA polymerase II cannot fulfill its task without binding to other proteins, called the *general transcription factors*, and building a complex. These general transcription factors help the enzyme to correctly bind to the promoter region. Since the DNA in eukaryotes is condensed in small volumes as nucleosomes and chromatin structures the initiation of transcription requires more proteins. *Transcriptional activators* are gene regulatory proteins which bind to specific regions in DNA (mostly located adjacent to promoter) to attract RNA polymerase II to the promoter region. There is also the need for a *mediator* protein that helps the activator proteins to accurately interact with RNA polymerase II and the general transcription factors, building a huge protein complex, that enables the chromatin structure of the regulatory region to change so that the transcription begins. The mediator and the general transcription factors are

the same for the genes transcribed by RNA polymerase II, but the gene regulatory proteins and their binding regions on DNA is different for each gene. There are also *gene repressor proteins* regulating the transcription rate of a gene by preventing the activators from binding to the mediator, hence blocking the transcription. Actually the gene regulatory proteins themselves are not mostly considered to be activators or repressors because they don't act individually but as a part of a much bigger complex, whose function can be different (activatory or inhibitory) for each gene depending on the total set of its components. Hence an activator for a specific gene can act as repressor for another one. A schematic view of these complexes are seen in Fig. 2.2.



.

Figure 2.2: Assembly of gene regulatory proteins with RNA polymerase II via the mediator protein to initiate transcription [5].

The gene regulatory proteins are also products of other genes. So the presence or absence of a protein x, which is the product of gene X, can influence the transcription rate of the gene Y, resulting in increase or decrease of its protein y, and we say

that gene X regulates gene Y. This type of regulation is called the *transcriptional regulation*.

Once the proteins are synthesized their activities in the cytoplasm can also be controlled by post-translational regulations induced by enzymatic modifications to the proteins. *Ubiquitination* is the process where a protein is labeled for degradation by a covalently bound protein called *ubiquitin*. Another example is the change of the three dimensional structure of the proteins causing them to be activated or deactivated. Phosphorylation and dephosphorylation are reversible procedures where a phosphate molecule is attached to or removed from the protein causing a conformational change. There are over 40 types post-translational modifications controlling the activity of a protein and detailed explanations can be found in [5].

Gene regulation responsible for specific functions in the cell can be analyzed separately. In our analyses we use two types of gene regulation processes being cell cycle and cell differentiation. Below we briefly explain their biological mechanisms.

Cell cycle:

The cell cycle consists of an orderly series of events in the cell that enables it to duplicate its DNA and segregate its chromosomes equally into two new daughter cells via division of the cytoplasm. Although the mechanism vary in different organisms an oversimplified view for eukaryotes can be given as follows.

Mainly three periods are observed during the cell cycle, the *interphase*, the *mitotic phase* (M) and the *cytokinesis*. The interphase can be divided into 3 phases. In the G_1 (Gap 1) phase the cell grows in size and the number of proteins and organelles increases. If the internal and external conditions are suitable, the *synthesis* (S) phase is entered where the whole genome of the cell is doubled so that all the chromosomes are replicated. A second gap phase G_2 follows in which checkpoint mechanisms are triggered to control if the duplication was successful. At the end of the G_2 the mitosis begins, having its own subphases. In the *prophase* the chromosomes are condensed in pairs of compact rods called sister *chromatids*. This is followed by *metaphase* where

the nuclear envelope disassembles and the *mitotic spindle* takes its shape. All the sister chromatids are aligned at the spindle equator. In the *anaphase* they are they are pulled to the opposite of the spindle. The mitosis is completed with the *telophase* where the spindle is destroyed and the chromosomes are separated into two nuclei. The cell cycle is completed with the last step *cytokinesis* enabling the cytoplasm to divide into two cells in G_1 phase again, each having one copy of the genome and also one set of the duplicated organelles. These phases can be broadly summarized by four main phases being G_1, S, G_2 and M. In Fig. 2.3 schematic view of the cell cycle is shown.

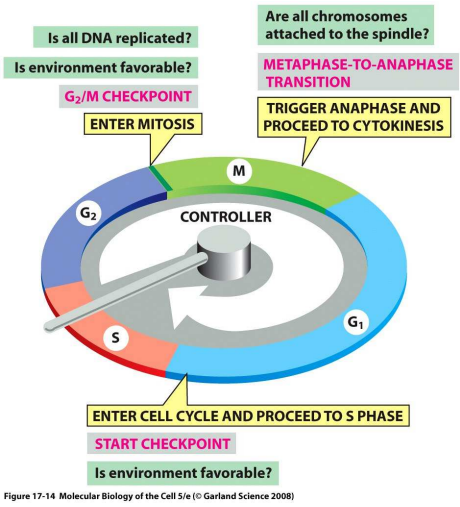


Figure 2.3: The cell cycle process with 4 main phases [5].

Regulation of the cell cycle:

In the eukaryotes the regulation of the cell cycle and the proteins responsible for this task are highly conserved throughout the evolution [67]. The mechanism of regulation and coordination of the orderly events during cell division is enormously complex since an error encountered in the process is can have fatal consequences for the cell. For example aneuploidy, presence of wrong number of chromosomes in the cell, occurs if segregation of chromosomes takes place before the duplication of the

DNA or cytokinesis precedes chromosome segregation. Additionally each event must be restricted to happen only once per cycle. For instance, duplication of chromosomes more than once leads to polyploidy.

Although the cell cycle process is highly conserved there are differences between the organisms in their proteins and details of the process. But an oversimplified scheme can be given for the single celled eukaryote *Saccharomyces cerevisiae*, the budding yeast, which helps to understand the notion of regulation through cell cycle. Our study also includes this cell cycle regulatory network explained below in detail.

Out of the nearly 6300 genes in *S. cerevisiae* 800 are taking part in the cell cycle, changing their expression levels throughout the process [68]. In their work Li et al. [8] identified a small set of 15 key elements whose intricate interactions regulate and coordinate the division. These consist of 7 cyclins (Cln1,2,3 and Clb1,2,5,6) which have to bind to the cyclin dependent kinase CDC28 and form a complex to operate, 4 molecules acting as inhibitors, degraders or competitors of the CDC28/cyclin complex (Sic1, Cdh1, Cdc14,20) and 4 transcription factors (SBF, MBF, Swi5, Mcm1/SFF). The concentrations of cyclins, as the name suggests, varies cyclically according to the phases of the cycle. Cln3 is specific to G_1 phase, Cln1,2, Clb5,6 are S-cyclins where Clb1,2 are mitotic cyclins. Hence the systematic progress of the cycle depends mainly on gene regulation mechanism leading to periodically expression of cyclins.

When the cell is in the G_1 state only Sic1 and Cdh1 are present. Sic1 inhibits the Clb-Cdc28 complexes to prevent premature DNA duplication and blocks the entry to the S phase. When the cell size reaches a critical value under favorable internal and external conditions the cyclin Cln3 becomes activated and binds to Cdc28. This complex phosphorylates Whi5, an inhibitor of SBF and MBF, to deactivate it, so that SBF and MBF can be released. In turn they act as transcriptional activators for the genes Cln1,2 and Clb5,6 respectively. Once the proteins of the gene Cln1,2 increase in the cell they accelerate the degradation of Sic1 so that DNA replication can begin and the cell enters the S phase where Clb5,6 are crucial for the initiation of DNA replication. In the late G_2 phase Clb5,6 activate Mcm1/SFF which are tran-

scription factors for Clb1,2, proteins leading to the transition to M phase. When the DNA is completely replicated and the chromosomes are aligned Mcm1/SFF activates Cdc14,20 to promote the entry to the anaphase of the mitosis. Cdc14,20, whose transcription rates are also enhanced by Clb1,2, activates Cdh1, which was inhibited by Cln1,2 via phosphorylation previously, and through ubiquitination Cdh1 causes degradation of the mitotic cyclins Clb1,2. The decrease of Clb1,2 that was inhibiting Swi5 to enter the nucleus, leads to the activation of Swi5 which, once in the nucleus stimulates the expression of Sic1 and takes the cell back to the G_1 phase. It also acts as a transcription factor for the genes expressed at the M/G_1 boundary.

The cell cycle GRN:

Li et al. constructed a successful model for the GRN of the cell cycle process in *S. cerevisiae* [8]. It is generated by representing the N key elements as nodes, where each node number i corresponds to a specific element, and their interactions as directed edges from node i to j. An inhibitory action, such as the ubiquitination of Clb1,2 by Cdh1 or the phosphorylation of Cdh1 by Cln1,2, is drawn red whereas an activatory interaction, such as the transcriptional regulation of Cln1,2 by SBF is drawn red. In the $N \times N$ adjacency matrix A we have the condition

$$a_{ij} = \begin{cases} 1 & \text{if node } j \text{ activates node } i \\ -1, & \text{if node } j \text{ inhibits node } i \\ 0, & \text{no interaction,} \end{cases} \tag{2.9}$$

for the case where distinctions of the strengths and speeds for the different interaction mechanisms are not taken into account, such that the edges all have the weight 1. The emergent network with corresponding adjacency matrix can be seen in Table 2.3.

Cell differentiation:

Cell differentiation is the process where a cell changes its type to specialize at a specific function, where its size and metabolic activity is altered. In multicellular organisms

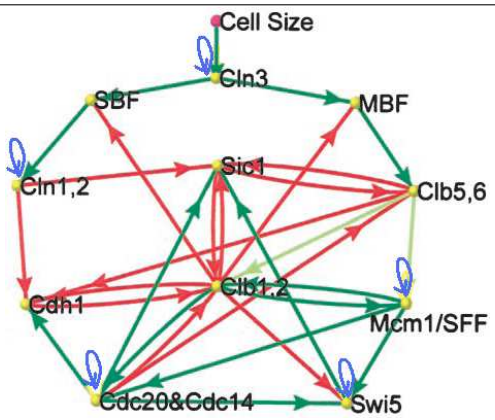
Network [8]	Adjacency matrix A
	$\begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -1 \end{bmatrix}$
Node numbers: (1-Cln3), (2-MBF), (3-SBF), (4-Cln1,2), (5,Cdh1), (6,Swi5) (7-Cdc14,20), (8-Clb5,6), (9-Sic1), (10-Clb1,2), (11-Mcm1/SFF)	

Table 2.3: The cell cycle GRN for the budding yeast along with its adjacency matrix. $a_{2,1} = 1$ indicates that Cln3 activates MBF and $a_{5,4} = -1$ indicates that Cln1,2 inhibit Cdh1.

differentiation occurs multiple times as the organism grows from a zygote to a complex of several tissues each containing cells with different functions. Differentiation is the core process to understand the morphogenesis in organisms.

Although all the cells differentiated from a common cell possess the same genome in their nucleus they express only a subset of their genes, a process made possible through gene regulation. Hence each cell type is defined by its protein expression profile. In the GRN for differentiation, the nodes are regulatory genes and its cis-regulatory modules. The main determinant of regulation mechanism is the cell-signaling, where signaling molecules, called growth factors, transfer information from one cell to another. This is generally enabled by a ligand, produced by one of the cells, that binds to a receptor in the extracellular region of another one where it conformationally changes a receptor. This receptor is now able to initiate a chain of phosphorylation events of protein which eventually activates a transcription factor responsible for the differentiation. Different signals lead to different target cells which in network language can be thought as different initial conditions leading to different attractors.

Asymmetric cell division, where the cytoplasm is unevenly distributed during cytokinesis, can lead to differentiated cell types due to unequal regulatory molecules in each of the daughter cells.

GRNs of hematopoietic stem cells, which give rise myeloid and lymphoid lineages of blood cells, are used as model networks in this study.

2.5 Boolean formalism for modeling

Boolean formalism was introduced in biology firstly by Kauffman as random models to investigate dynamics of gene regulatory networks when their structures were not known in 1969 [22]. The random ensembles of these networks, which were able to perform some functions of biological networks, shed light on very essential principles of living systems in all scales, like evolvability, robustness, information storage and relationship between structure and dynamics [69]. This method was not only import for studies in biology but also in mathematics and physics where the main focus is rather on the emergent properties of a system as a whole than on the single elements comprising it. As an example we can give the classical random Kauffman networks with uniform in-degree 2, which exhibit a phase transition between an ordered and a chaotic phase with respect to propagation of perturbations. Here we will summarize only the basic notions required to understand GRNs.

In the study of GRNs the nodes of a network represent the genes, proteins or RNAs and the directed edges stand for the influences of genes on the others, generally through their products. But these interactions are not incorporated in the structure of the network but are represented as Boolean functions, depending on the activity levels of the nodes which can take only two values, 0 or 1, indicating ON or OFF states respectively. More formally we can say that a Boolean network of size n, as introduced by Kauffman [22], is determined by the set of nodes, representing genes, $\{n_i : 1 \leq i \leq n\}$ and a set of Boolean functions for each node $\{f_i : 1 \leq n\}$. Hence each node n_i has a binary value $\sigma_i \in \{0, 1\}$ whose change in time is given by the

function $f_i : \{0,1\}^m \to \{0,1\}$ in the form

$$\sigma_i(t+1) = f_i(\sigma_{k_1}(t), \sigma_{k_2}(t), \cdots, \sigma_{k_m}(t)),. \tag{2.10}$$

where the set $k_j, (1 \leq j \leq m)$ denotes the neighbors of node n_i. The behavior of the functions f_i are determined according to experimental observations. The values of the nodes represent the expression states of the corresponding genes such that $\sigma_i = 1$ and $\sigma_i = 0$ indicate that the gene is expressed and not expressed respectively. The state of the whole network then can be written as

$$\sigma(t) = \{\sigma_1(t), \cdots, \sigma_n(t)\}, \tag{2.11}$$

with the attractors of the dynamics having the property that

$$\sigma(t+q) = \sigma(t), \tag{2.12}$$

where $q = 1$ identifies a fixed-point (point attractor) of the system.

This deterministic behavior is experimentally justified with the observation that many organisms exhibit a fixed gene activity under molecular stochasticity. Additionally, the switch-like behavior can be attributed to the fact that although the concentration level of protein span several orders of magnitude, many molecular reactions take only place either in a saturated or in very small concentrations, where the transition between the two regimes is very fast [70].

The order of the updates of the nodes can be synchronous, asynchronous or a mixed scheme of both. Although the synchronous scheme is not biologically realistic due to the fact that biochemical reactions operate at different time scales, it simplifies the computation without loosing qualitative properties of the network. The more realistic asynchronous updating can have the consequence that attractors of the synchronous scheme become unstable and decompose into stable cycles. Recent studies suggest that molecular networks are designed so robustly that the timing of update is not a

crucial factor in the dynamics [71].

According to Kauffman the cell types of an organisms corresponded to the attractors of the underlying network. This idea was based on the fact that in spite of containing the same genome with the same network structure the cells could be differed from each other by their activity levels of the individual genes. As he investigated the regulatory networks in 1970's the available biological knowledge suggested that the number of different cell types of an organism was linearly dependent to the square root of the gene number in the genome. His results on the observations with computer simulations possible at that time also showed that the number of attractors of random networks of size n with exactly two inputs for each node was also scaling with the $\sqrt{n}$. This was a success for the simple Boolean networks to represent more complex real phenomena. Although we know today that biological complexity is not proportional to the size of the genome and that the square root law for number of attractors is not valid, the success of Boolean models survives for an important of biological networks. This is due to the fact that the structure of the networks plays a more crucial role than the details of the kinetic interactions. Boolean networks also help scientists to understand or model large systems where the quantitative details of the whole system are poorly known making a continuous approach impossible. Hence Boolean networks serve as a successful model complementary to continuous models, depending on the information available, size of the systems and questions to be investigated.

Chapter 3

COHERENT REGULATION IN GENE REGULATORY NETWORKS

The global and structural features mentioned in Section 2.2 are very powerful methods to explain evolutionary design principles among the biological networks including gene regulatory networks. They help to understand how the structure is determined and constrained by the requirement that the regulatory dynamics delivers a timely production of necessary proteins. However, it was pointed out by Wu et al. [18] that despite their success they are incomplete because an important feature, the signs of the edges arising from their functions, are overlooked. Even though functional motifs are taking this property into account, they are restricted to local subgraphs. The quantity *regulation entropy* is proposed as a global property for the networks taking the signs of the edges into account, which is a measure for the consistency of regulations along different pathways [18]. It is shown that the cell cycle networks of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, which we also consider in our study, are associated with low levels of regulation entropy when compared with similar random networks. Our present investigation of *coherent regulation* in biological networks is a progress in this direction.

3.1 Definition: Coherent Regulation

It has been shown that a differentiating mammalian cell consumes one-third of raw material and energy to produce the proteins it needs whereas for a growing bacterial cell this ration increases to one-half [72, 73, 74]. How these economic pressures have been a determinant of functional organizations in all living things in the history of evolution is widely studied [75, 76, 77]. Since proteins reach high expression levels to

regulate and coordinate the gene activity in the cell, it is plausible to ask if there is an optimization in GRNs in this aspect, where the networks are wired such that the desired functionality is attained with minimal use of the material available to the cell. Previous studies about the idea that the evolutionary pressure for economy may have shaped the regulatory interactions, identified the class of Boolean functions that are better at modeling regulatory dynamics [78] and investigated the frequency of gene duplications, creating new interactions, in microbes [79].

Our point of view is that to establish energy-optimal functionality in GRNs they should be wired such that simultaneous expression of contradictory regulators for the same gene is avoided because this would lead to a situation where one or more proteins are produced redundantly which are unable to fulfill their tasks because of the opposition from the others, leading to waste of material and energy. Additionally, in a steady state, a gene may receive contradictory inputs in the presence of two distinct transcriptions factors trying to activate and inhibit the gene respectively. The opposition between them may not cause a change in the current state because of the regulatory mechanism involved but the silence of one of them due to a mutation or error could influence the fate of the cell by changing first the state of the target gene and then the whole steady state of the cell. Therefore it would be advantageous that in the stable state the protein levels of an activator and of an inhibitor of the same gene are not high simultaneously. Circadian cycles [80, 81] can yield exceptions with brief transition periods (caused by internal or external conditions) between steady states or between the states in cyclic attractors, where changes in the expression levels are due to role switches between such antagonistic regulators. We coin the term "coherent regulation" to denote this unity of purpose among simultaneously expressed genes sharing a common regulatory target when the cell is in the steady state. It measures the degree to which a GRN is refined to balance its structure and dynamics towards minimum waste of material or energy. To be specific, we say that a gene is coherently regulated if its expression state is determined unanimously by its active regulatory partners, as described in Fig. 3.1. The details of how we quantify coherence will be

given in Section 3.3. Hence starting from the fact that the structure shapes regulatory dynamics in a phenotype, the question we ask here is whether evolutionary processes may have provided the means to feed information from the dynamics back to the structure, promoting modifications that maximize coherent regulatory action.

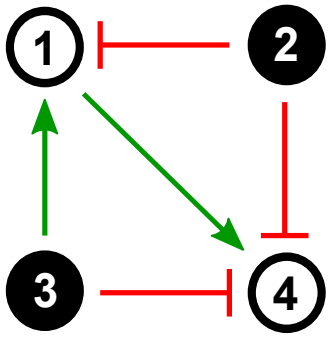


Figure 3.1: An example network with coherently and incoherently regulated genes. Black and white circles indicate active and inactive nodes, respectively. Node 4 receives coherent input (only repressive) since node 1 is inactive. But node 1 itself is incoherently regulated since it receives conflicting inputs from nodes 2&3.

We have to mention here that in literature there other definitions for "coherence" in biological networks in the concept of robustness analysis [13]. Yet another use of this term can be found in the categorization of network motifs [12], where it refers to the compatibility of alternative directed paths connecting two nodes, a similar terminology appearing in the definition of regulation entropy [18]. In contrast, besides being a function of the interactions (edges), the *coherence* in our definition depends also on the expression states of genes (nodes).

Our main claim is that evolutionary processes may have provided the means to feed information from the dynamics back to the structure, promoting modifications that maximize coherent regulatory action in biological networks.

3.2 Time Evolution

Dynamical aspects of genetic regulation can be analyzed by Boolean models [6, 7, 82, 83], Petri nets [84, 85, 86], and differential equation based continuum models [87, 88], depending on the desired resolution. Since in our study we concentrate our attention to only on the steady states (attractors) of the system, the details of the dynamical steps are not crucial. Therefore the application of Boolean formalism (see Section 2.5) with coarse-grained characterization of stationary states is suitable for our purpose having the advantage of being simple and accessible. The power of this formalism has been demonstrated in modeling the regulation of the cell cycle [6, 89, 90], cell differentiation [9, 91, 92], circadian clocks [93, 81], etc. Yet our hypothesis on Boolean systems can be generalized in a straightforward manner to continuum models.

A (global) state in a GRN with size n can be described by the expression levels of each of the n genes (local states) separately. Following the Boolean formalism, the local state $\sigma_i(t)$ of the node i $(1 \leq i \leq n)$, can be modeled as having a switch-like behavior where the binary value 1 (or 0) for a node indicates that either the corresponding gene is active (or inactive) or the expressed protein level is above (or below) a given threshold value. Hence a state at time t can be given as binary vector $\boldsymbol{\sigma}(t)$ having n local states as its elements. The adjacency matrix $\boldsymbol{A}$ has the signed elements A_{ij} with values in an interval of integers. Then, the evolution of the states in time can be analyzed by defining a time-evolution operator T, such that given an initial state $\boldsymbol{\sigma}(t)$, the following state $\boldsymbol{\sigma}(t+1)$ can be written as

$$\boldsymbol{\sigma}(t+1) = T\left[\boldsymbol{\sigma}(t), \boldsymbol{A}\right]. \tag{3.1}$$

For the cases where the strengths or speeds of the interactions are thought to be equal, the elements of the adjacency matrix would differ only by their signs. If suitable, using the negation operator, the negative effects can be incorporated into a logical function, which depends only on the previous state of the network without making explicit use

of the adjacency matrix. In this respect one can use

$$\boldsymbol{\sigma}(t+1) = T\left[\boldsymbol{\sigma}(t)\right]. \tag{3.2}$$

In our models for the lymphocyte and myeloid differentiation along with the mammalian cell cycle, this kind of logical functions are used.

The synchronous updating scheme is used for all the considered networks in the study, such that the time-evolution equations of the systems are deterministic with all the initial states having a definite attractor (steady state), where the cell spends most of its time. The attractors of the system are cycles denoted by $\boldsymbol{\sigma}_s$ of length q_s, with the property that a member $\boldsymbol{\sigma}_s^{(k)}$ $(1 \leq k \leq q_s)$ of the cycle satisfies the relation

$$\boldsymbol{\sigma}_s^{(k)} = T^q[\boldsymbol{\sigma}_s^{(k)}]. \tag{3.3}$$

When $q = 1$ the cycle reduces to a fixed point.

For each model network we will consider in this study [8, 6, 94, 7, 10, 2], the rules of dynamics shaping the operator T are different. Their details will be given in Section **??**.

3.3 Quantifying Coherence in GRNs

It is evident that an analysis of an optimization in protein level requires a deeper wisdom than we presently have about the mechanisms in the cells. Here, we will define coherence for efficiency in GRNs, by only concentrating on the choice of interactions between the genes which coordinate the well-timed expressions of proteins.

Without loss of generality we will work with adjacency matrices whose elements are given as $a_{ij} \in \{-1, 0, 1\}$. Here ± 1 stands for positive and negative regulatory influence from the node j to the node i, where 0 indicates the absence of any interaction. First the coherent regulation of a single gene will be described. Then the degree of coherence for a single steady state will be defined, which will be extended to the coherence coefficient for a GRN.

In a GRN of size n, denote the set of the neighbors belonging to node i ($1 \leq i \leq n$) as I_i, which can be given as

$$I_i = \{j : a_{ij} \neq 0\}. \tag{3.4}$$

We say that the node i is coherently regulated if and only if it receives same type of influence (all activatory or all inhibitory) from all of its active neighbors, i.e. when one of the conditions

$$a_{ij} = 1, \qquad \forall j \in I_i \wedge \sigma_j = 1 \tag{3.5}$$

or

$$a_{ij} = -1, \qquad \forall j \in I_i \wedge \sigma_j = 1 \tag{3.6}$$

is satisfied (see Fig. 3.1). This requirement can take the more compact form

$$\left| \sum_{j=1}^{n} a_{ij}\sigma_j \right| = \sum_{j=1}^{n} |a_{ij}|\, \sigma_j \tag{3.7}$$

since the non-neighbor nodes with $a_{ij} = 0$ do not make any contribution to the sum. For a state $\boldsymbol{\sigma}$ the degree of coherence is then the fraction of the coherently regulated nodes (for which Eq. 3.7 is satisfied) in the state which can be given as

$$\alpha_c[\boldsymbol{\sigma}] = \frac{1}{n} \sum_{i=1}^{n} \mathrm{int} \left(\left| \sum_{j} a_{ij}\sigma_j \right| \Big/ \sum_{j} |a_{ij}|\, \sigma_j \right), \tag{3.8}$$

where $\mathrm{int}(x)$ is the integer function taking the value 1 if Eq. 3.7 is satisfied or 0 otherwise. In the case when the nominator and denominator vanish in Eq. 3.8 due to the absence of any input to it, the node is interpreted to be coherently regulated. A similar circumstance where all the nodes in the state are silent ($\boldsymbol{\sigma} = 0$) is also defined to have $\alpha_c[\boldsymbol{\sigma} = 0] = 1$.

The coherence for the whole network is measured over all of its steady states (attractors) which are fixed points or limit cycles of the given dynamics satisfying Eq. 3.3. For a network having m_a attractors, denoting members of the cycles as $\boldsymbol{\sigma}_s^{(k)}$

$(1 \leq s \leq m_a)$, the global coherence coefficient for the network can be expressed as

$$\alpha_c = \sum_{s=1}^{m_a} \frac{w_s}{q_s} \sum_{k=1}^{q_s} \alpha_c[\boldsymbol{\sigma}_s^{(k)}], \tag{3.9}$$

where q_s is the length of the limit cycle s and w_s is a predefined weight for the attractor, usually taken as the relative basin size, which gives the ratio of initial states ending up at that attractor, with the condition that $\sum_s w_s = 1$.

Usually the cell cycle networks have one biologically relevant fixed point which is reached from the initial state corresponding to the biological G_1 phase and other possible steady states which are attractors of the dynamics but are not seen in the living cells. Differentiation networks can have several fixed-points representing the various observed cell types after the differentiation process.

Although we do not consider it in our study, the formalism above can be generalized to continuous models of gene regulation, where averages taken over states are replaced with averages over time. This will also require that the coherence measure is extended to a suitable range of continuous variables. As an example one can look at the case of a linearized model of the form

$$\frac{dN_i(t)}{dt} = \sum_j a_{ij} N_j(t), \tag{3.10}$$

where N_i denotes the protein or RNA concentrations. A feasible generalization of Eq. 3.8 could be

$$\alpha_c\left[\boldsymbol{\sigma}\right] = \frac{1}{n} \cdot \frac{1}{T} \sum_{i=1}^{n} \int_0^T \frac{\left|\sum_j a_{i_j} N_j(t)\right|}{\sum_j |a_{ij}| N_j(t)} dt, \tag{3.11}$$

where T is the period of the atrractor in consideration.

For the stochastic systems the proposed formalism can also be extended by substituting the operator T in Eq. 3.2 $\boldsymbol{A}$ with a stochastic transition matrix $\boldsymbol{\tau}$ and replacing the state vector $\boldsymbol{\sigma}$ with the probability distribution vector of the ensemble of states $\{\boldsymbol{\sigma}\}$.

3.4 Generation of Randomized Networks

To test the idea that biological networks have a bias towards high coherence values, we compared them with the ensembles of random networks. Two different sets of randomized networks are used where the first set strictly resembles the original biological networks under consideration (see Section 3.4.1) whereas the second set has more relaxed constraints but keeps some of the structural features observed in original ones (see Section 3.4.2). To create a reference ensemble for each type 10^4 networks are generated which are tested to be non-isomorphic to each other (by comparing the eigenvalue spectrum) and to possess a connected underlying undirected graph. Below we clarify these two classes of random networks.

3.4.1 Random Networks with Similar Topology

The first class of randomized networks is generated by reshuffling the edges of the original network. We follow the procedure used by Maslov et al. [95] to investigate linking properties of biochemical networks and by Alon [12] to determine frequent network motifs. Here a null-model ensemble consists of networks which are generated by rewiring the initial network by keeping the in- and out-degrees for both activatory and inhibitory edges of each node constant. To do this a pair of edges of the same type (+ or -), $i \to j$ and $m \to n$, is randomly selected. These edges are removed from the network and new edges, $i \to n$ and $m \to j$, are added, provided that neither of them was present previously. These steps are repeated sufficiently many times, for both activatory and inhibitory edges separately, such that the resultant network has statistically no correlation with the original network except than the local similarities imposed by the constraints above. Throughout the text a network generated from the initial directed graph G by this method will be referred to as a random network of the first type and denoted by $RN_1(G)$.

3.4.2 Random Networks with Same Structural Parameters

In the second class of randomized networks the constraint to resemble the original network is more relaxed. Here only the number of edges per node (k) and the ratio of positive edges to the total number of edges (p) are conserved, which are two obvious factors contributing to coherence (α_c). For a network of size n, the value of k is calculated by excluding the self-loops and hence it satisfies $1 \leq k \leq n-1$, where the minimum value 1 is necessary but not sufficient for connectivity. A random network having the same values for (n, k, p) set as the original network is constructed by initially generating a connected, undirected random graph with n nodes and $E = n \cdot k$ interactions, where any edge has the same probability to be present. Each edge is then given a direction and sign (+ or -) randomly, regarding the condition that a fraction p of them activate, and the rest inhibit their target nodes. Note that, a node may activate one of its targets while inhibiting another one. This behavior is consistent with the ones found in our model networks and recent studies suggest that the transcription factors acting in both ways are more common than previously believed [96, 97]. Finally the self-loops with the same quantity as in the original network are added to the randomly chosen nodes. In the text a random network with the same set of (n, k, p) values of an initial network G will be referred to as a random network of the second type and denoted by $RN_2(G)$. If the (n, k, p) set is given explicitly, without a reference to an existing network, we will simply write $RN_2(n, k, p)$ generating a network without self-loops.

Chapter 4

COHERENT REGULATION IN SIX GENE REGULATORY NETWORKS

4.1 Model Networks for Budding Yeast's Cell Cycle (YCC)

Among the eukaryotes, the unicellular budding yeast, *S. cerevisiae* is one of the most extensively studied organism [98, 99, 100, 101]. Its completely sequenced genome lets scientists estimate that its $\sim$ 6300 proteins, among which 307 are TFs, take part in $\sim 2 \cdot 10^5$ regulatory interactions [101]. Unfortunately it is very demanding to characterize all of these interaction so with the currently available data the regulatory network as whole is not realizable.

However, the small subnetwork controlling the cell cycle process is studied in detail [8, 102, 103, 104, 105]. And successful models for the regulatory dynamics have been shown to produce expression profiles in proper order which were experimentally observed [8, 104, 105]. Here, we will investigate two such models, one proposed by Li et al. [8] and one proposed by Hong et al. [105], sharing the key elements but having minor differences between their interaction and dynamics, to demonstrate their tendency towards high coherence. Following the Boolean formalism explained in Section 2.5 we will briefly describe their structure and dynamics.

4.1.1 Cell Cycle Model by Li et al. (M1)

In their work Li et al. successfully constructed a Boolean network model for the cell cycle regulation of the budding yeast *S. cerevisiae* where they classified a small subset of key elements and interactions between them as mentioned in Section 2.4.2 [8], which was used by others in their later studies. The model, that will be referred as the first

model ($M1$), can be seen in Fig. 4.1(a). Here the blue nodes represent the auto-repressor elements ($a_{ii} = -1$) and nodes with two colors (magenta/blue) indicate the protein complexes.

The interaction matrix with the corresponding numbering for the genes, which can also be deduced from the Fig. 4.1(a), is given as follows:

$$A = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -1 \end{bmatrix}, \text{with}$$

i, j	Protein/TF
1	Cln3
2	MBF
3	SBF
4	Cln1
5	Cdh1
6	Swi5
7	Cdc14,20
8	Clb5,6
9	Sic1
10	Clb1
11	Mcm1

(4.1)

Each state $\boldsymbol{\sigma}(t)$ is iterated using the deterministic and synchronous Boolean time evolution equations as seen in Eq. 4.2, which can be interpreted as a majority rule implying that a node is activated (inhibited) when it receives more activatory (inhibitory) inputs from its active neighbors and they are in form:

$$\sigma_i(t+1) = \begin{cases} 1, & \sum_{j \neq i} a_{ij}\sigma_j(t) > 0 \\ 0, & \sum_{j \neq i} a_{ij}\sigma_j(t) < 0 \\ & \text{or} \\ & \sum_{j \neq i} a_{ij}\sigma_j(t) = 0 \text{ and } a_{ii} = -1 \\ \sigma_i(t), & \text{otherwise.} \end{cases} \tag{4.2}$$

From Eq. 4.1 it is easy to read that the network of size 11, excluding the self-loops in the diagonal elements, has 29 edges among which 15 are activatory and 14 inhibitory, leading to structural parameter set (see Section 3.4.2)

$$(n, k, p)_{M1} = (11, 2.64, 0.52). \tag{4.3}$$

4.1.2 Cell Cycle Model by Hong et al. (M2)

The model $M1$ was later modified by Mangla et al. by inserting new interactions and allowing the edges having unequal weights [104]. The Boolean formalism was also extended such that the nodes may have three different values (0,1,2). Hong et al. later refined this work to enable it to capture checkpoint conditions which were omitted in M1 [105]. Their model that we will refer to as model $M2$ is seen in Fig 4.1(b) with the same coloring scheme and node numbering as in $M1$.

In spite of possessing the same nodes, the structure of the network in $M2$ differs from the one in $M1$ such that the set for the possible edge weights is extended to $a_{ij} \in \{0, \pm 1/3, \pm 1, \pm 3\}$, which are shown with different thicknesses in Fig. 4.1(b), leading to the adjacency matrix in Eq. 4.5. The nodes again have the binary Boolean values, with the exceptions being Swi5 and Clb2, which are allowed to have the additional value '2', where the transitions $0 \leftrightarrow 1$ and $1 \leftrightarrow 2$ are triggered, in contrast to $M1$, by the nonzero threshold values θ_1 and θ_2 respectively. The dynamical discrepancies can be noticed in the following time-evolution equation,

$$\sigma_i(t+1) = \begin{cases} \sigma_i(t) + 1, & \sum_j a_{ij}\sigma_j(t) \geq \theta_{\sigma_i(t)+1} \\ \sigma_i(t) - 1, & \sum_j a_{ij}\sigma_j(t) < \theta_{\sigma_i(t)} - 0.75 \\ \sigma_i(t), & \text{otherwise,} \end{cases} \tag{4.4}$$

where $\theta_1 = 0.5$ and $\theta_2 = 1.5$. Here θ_0 is not defined since the local state $\sigma_i(t) = 0$ cannot be lowered. The coefficients a_{ij} are the elements of the adjacency matrix

$$A = \begin{bmatrix} 0 & -1/3 & -1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 3 & 0 & 0 & -1/3 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1/3 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1/3 & 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1/3 & 1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 & 1 & -1 \end{bmatrix}, \tag{4.5}$$

which shows that the structural parameters for the model $M2$ are

$$(n, k, p)_{M2} = (11, 3.18, 0.43). \tag{4.6}$$

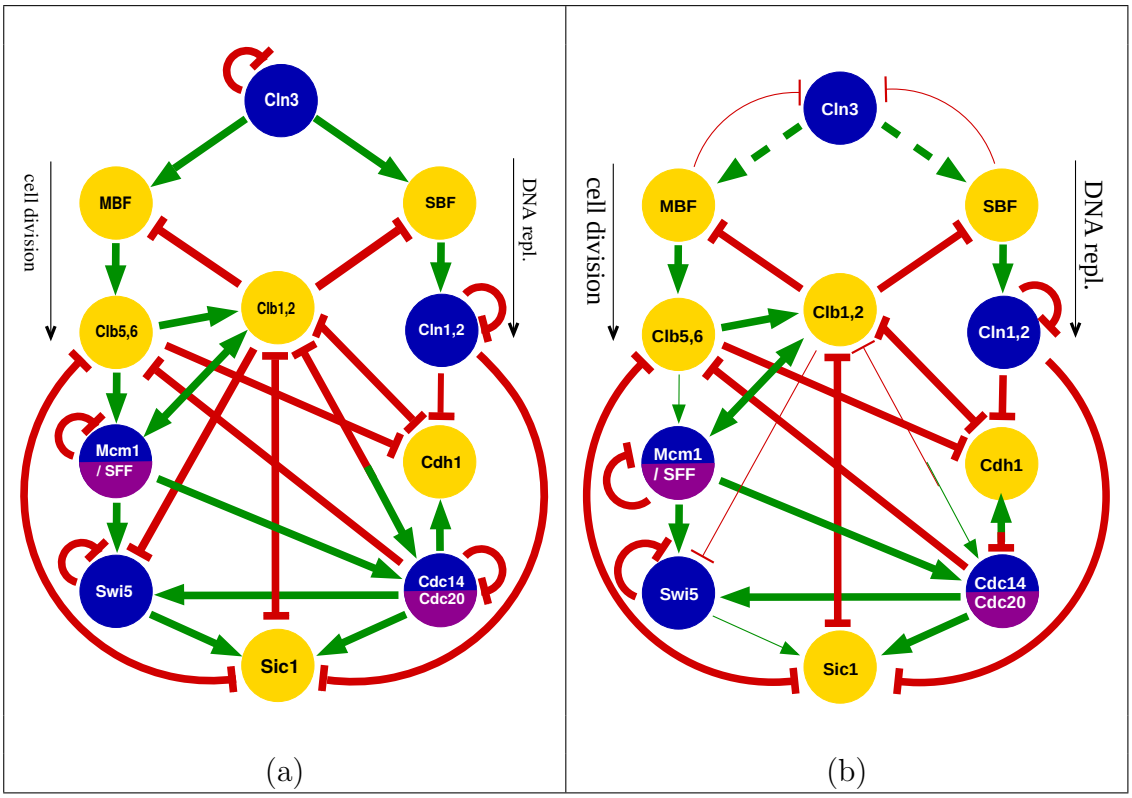


Figure 4.1: Two model systems for yeast cell cycle. (a) Model $M1$ by Li *et al.* (b) Model $M2$ by Hong *et al.* The weights for interactions in $M2$ are 1, 1/3 and 3 for normal, thin and dashed lines, respectively.

4.2 Coherence in Yeast's Cell Cycle Network

When the time-evolution equations for both $M1$ and $M2$ are implemented in a deterministic manner the experimentally observed expression states are visited in proper order ending up in the G_1 phase. But a stochastic application with an asynchronous updating scheme allows the possibility of different point attractors listed in Table 4.1. The biological interpretation of these low probability states is not clear [91, 8]. How-

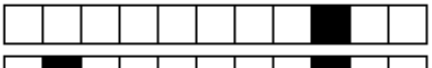
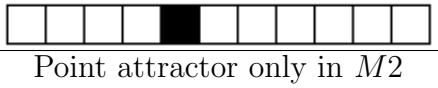
F	Cln3 MBF SBF Cln1,2 Cdh1 Swi5 Cdc20 Clb5 Sic1 Clb1,2 Mcm1	$\alpha_c[F]$	
	Point attractor in $M1$ & $M2$	$M1$	$M2$
F1		1	1
F2		1	0.91
F3		0.91	0.91
F4		1	1
F5		0.91	0.91
F6		1	1
	Point attractor only in $M2$		
F7		-	0.64
F8		-	0.73

Table 4.1: YCC attractors under the dynamics given in Eqs. 4.2 and 4.4 and their coherence, $\alpha_c[F]$. Black boxes indicate the active genes and the first row corresponds to the experimentally verified G_1 phase. The trivially coherent $\boldsymbol{\sigma} = 0$ fixed point is left out of the analysis.

ever, our coherence analysis is based on all of the attractors of a given network which promotes the comparison of the YCC network with randomized ones where the attractors do not have any biological relevance. Furthermore, we also present results obtained by averaging over basin-sizes which enable us to focus on the G_1 attractor with the biggest basin-size. Additionally, the trivial zero fixed point, which is biologically irrelevant, is not included in our analysis.

The coherence values for the attractors of the two models, which are calculated using Eq. 3.8, are also given in Table 4.1 where the F1 state corresponds to the G_1 phase of the cell cycle. The uniformly weighted average of these values gives the coherence of the whole network being $\alpha_c = 0.97$ ($M1$) and 0.89 ($M2$) whereas a basin size weighted average yields $\alpha_c = 0.99$ (M1) and 0.98 (M2). It is observed that all the common attractors of $M1$ and (M2) (including G_1) are either fully coherent or

almost so.

Having made this observation we can ask if the YCC network is biased towards high coherence. A valid interpretation of the numbers above is possible by comparing them with results obtained by a properly chosen random ensembles, a method which was implemented successfully to find biological network motifs, for example [12, 1]. Therefore we perform our analysis on two different ensembles of networks of the same size and similar characteristics.

4.2.1 Coherence: Yeast versus Random Networks

For the comparison of YCC networks with the randomized ones, four ensembles of 10^4 distinct networks, namely the Yeast-like $RN_1(M1)$, $RN_1(M2)$ and more relaxed $RN_2(M1)$, $RN_2(M2)$ networks with only fixed edge numbers, are generated (for details of randomization process see Sections 3.4.1 and 3.4.2). Their attractors were found by an exhaustive search of states using Eqs. 4.2 & 4.4. In contrast to the YCC network cyclic attractors were also occasionally encountered. The probability distributions for α_c, obtained by implementing Eq. 3.9, for each of the ensembles are shown in Figs. 4.2 & 4.3, separately for uniformly and basin-size weighted cases. For the numerical values of structural parameters of the original models one can refer to Eqs. 4.3 & 4.6.

Surprisingly, a closer look at Figs. 4.2 and 4.3 reveals that $RN_1(M1)$ (both for uniform and basin-size weighted averaging) and $RN_1(M2)$ (excluding basin-size averaging) yield a lower degree of coherence in comparison with the ensembles of $RN_2(M1)$ and $RN_2(M2)$. The corresponding mean values along with the standard deviations of $\overline{\alpha}_c$ are listed in Table 4.2. It is calculated that the coherence values of the actual networks lie in the top 4% in all of the random ensembles $RN_1(M1)$, $RN_1(M2)$, $RN_2(M1)$ and $RN_2(M2)$. These observations suggest that the exceptionally high level of coherent regulatory activity in the two YCC models is not solely due to by the particular local decoration of the networks' nodes.

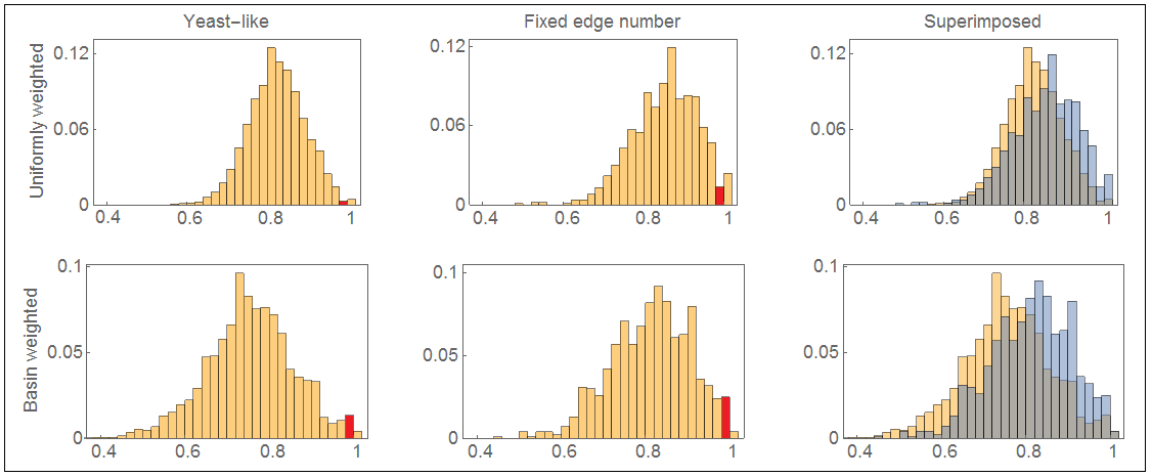


Figure 4.2: The histograms for α_c calculated for "YCC-like" ($RN_1(M1)$) networks (first column) and the random ($RN_2(M1)$) ensemble (second column) generated by using model $M1$ described in the text. The third column shows the superposition of the two histograms for comparison. α_c for the yeast's cell-cycle model is shown in red everywhere.

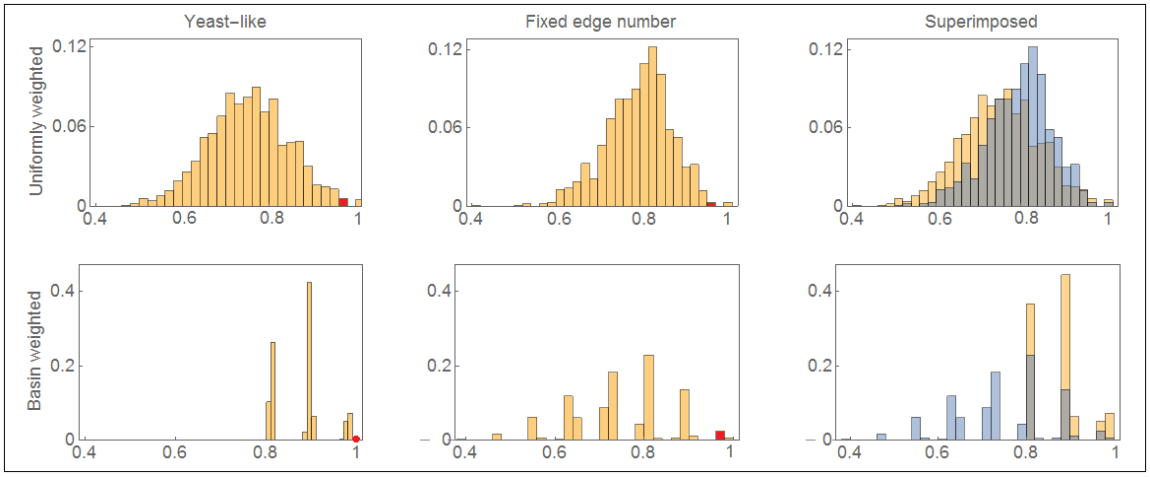


Figure 4.3: The histograms for α_c calculated for "YCC-like" ($RN_1(M2)$) networks (first column) and the random ($RN_2(M2)$) ensemble (second column) generated by using model $M2$ described in the text. The third column shows the superposition of the two histograms for comparison. α_c for the yeast's cell-cycle model is shown in red everywhere.

	Weight	$RN_1(M1)$	$RN_2(M1)$	M1	$RN_1(M2)$	$RN_2(M2)$	M2
$\bar{\alpha}_c \pm \sigma$	Uniform	0.82 ± 0.07	0.85 ± 0.08	0.97	0.75 ± 0.09	0.79 ± 0.07	0.89
	Basin-size	0.75 ± 0.10	0.81 ± 0.09	0.99	0.87 ± 0.06	0.75 ± 0.11	0.98

Table 4.2: Mean and standart deviation values for α_c of random ensembles along with the α_c values of model networks $M1$ and $M2$.

4.2.2 Coherence versus Structural Parameters

To investigate the dependence of mean degree of coherence on varying structural parameters n, k and p, we considered only the YCC model $M1$, for the simplicity of implementation. The dependence on k was analyzed by first generating an ensemble of 10^4 non-isomorphic random networks with size $n = 11$ (as in YCC) and $p = 0.5 \simeq p_{M1}$ for each possible $k \in [1.6, 10]$. Similarly, the dependence on p was analyzed with the same-sized ensemble consisting of networks of the type $RN_2(11, k_{M1} = 29/11, p \in [0, 1])$. To see how our results depend on the network size, much larger networks with $n = 110$ nodes (corresponding to a 10 fold increase in size) were generated with the same principle on (k, p) pairs. For $n = 11$ the attractors were again found with exhaustive search whereas for $n = 110$ a sampling of 2^{11} random states were used. Using Eq. 3.9, α_c values were calculated sequentially for each network over corresponding attractors with the weight being uniform or relative to basin-size separately. Finally the ensemble averages ($\overline{\alpha}_c$) are obtained for each k and p values in the range given above. The behavior of ensemble mean $\overline{\alpha}_c$ with respect to varying p and k is depicted in in Fig. 4.4(a,b) & 4.4(c,d) respectively.

As Fig. 4.4 indicates, for a given (k, p) pair $\overline{\alpha}_c$ changes only slightly with the number of nodes. A ten fold increase in network size makes a difference of $< 9.8\%$ in the worst case. Hence we can conclude that despite the smallness of the model system considered here, the results on random ensembles may be expected to serve as a reasonable null-hypothesis for coherence in the global regulatory network of yeast and other organisms.

The asymmetry with respect to $p = 1/2$ in the behavior of α_c is unexpected at

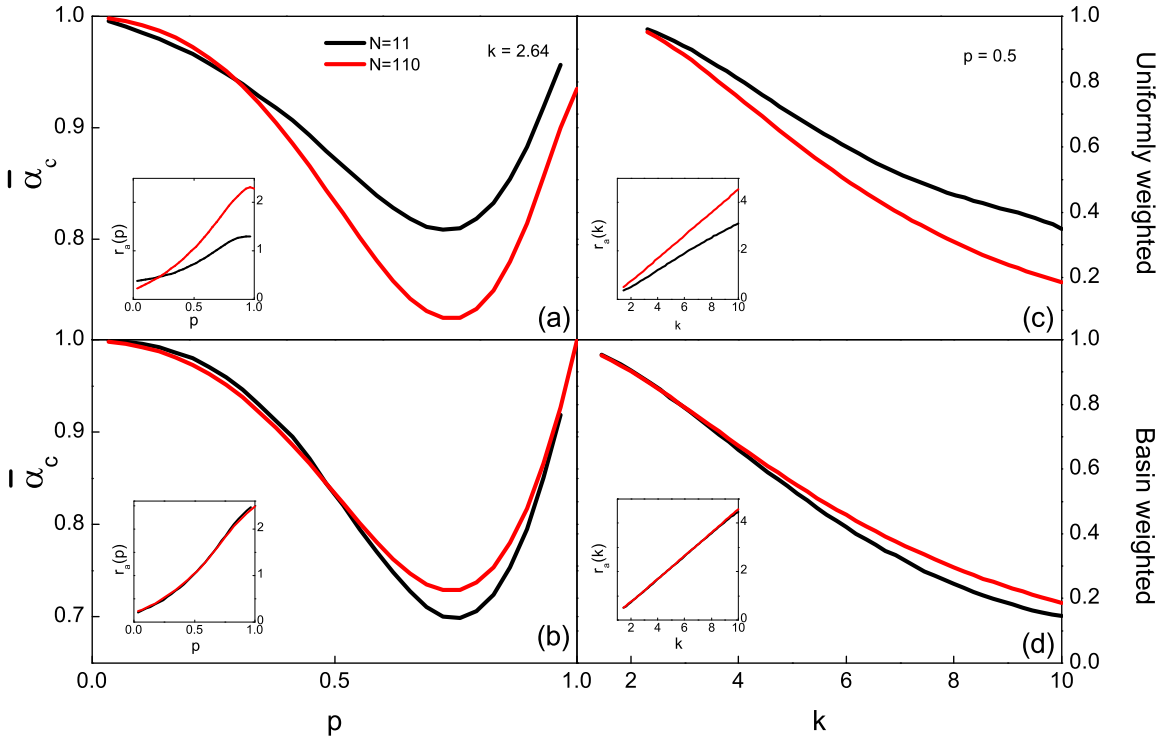


Figure 4.4: (a,b) Degree of coherence, $\bar{\alpha}_c$ *vs.* p for $k = k_{M1} = 29/11$ with and without consideration of the basin sizes. (c,d) Degree of coherence, $\bar{\alpha}_c$ *vs.* k for $p = 0.5 \simeq p_{M1}$ with and without consideration of the basin sizes, respectively. The insets show the average fraction of the regulating active nodes at the fixed points.

the first sight, because the number of nodes receiving both activating and inhibiting regulations is maximized precisely at this point. The origin of the asymmetry lies in the fact that the ratio of active nodes at the fixed points increases monotonically with p, which can be seen in the insets of Fig. 4.4. Close to $p = 0$ and $p = 1$, most genes receive same type of influence from its regulatory partners, which leads to an increase of coherence in these extreme regions. While the fraction of genes with conflicting incoming edges peaks at $p = 1/2$, the minimum of $\overline{\alpha}_c$ is shifted towards larger p values (more up-regulation), where the chances are higher for antagonistic regulatory partners of a gene to be simultaneously expressed.

The results above obtained by random sampling can be reproduced by a simple analytical model. Since we know that the results depend weakly on the network size,

we can consider the large-n limit without loss of generality. Now for fixed (k,p), we define a new quantity $r_a(k,p)$ which gives the average number of *active* neighbors per node. Recalling the empirical observation that when k is fixed, r_a increases uniformly with p (see Fig. 4.4(a,b) inset), we can fit the data for $n = 11$ and $n = 110$ with the analytical form

$$r_a(k,p) = r_{min} + p^{\gamma}(r_{max} - r_{min}), \tag{4.7}$$

where $r_{min,max}$ are $k-$dependent. Even in the case of a network with all inhibitory interactions ($p = 0$), a typical attractor will still have active nodes whose neighbors are inactive, yielding $r_{min} > 0$. A node is coherently regulated, if all its *active* regulating partners act unanimously. α_c measures essentially the probability for this event across the network's attractors. Treating the states of neighbors regulating a given node as independent random variables, this probability can be modeled as that of observing two extreme outcomes in r binomial trials where the binomial outcomes reflect the two possible assignments, up and down regulation with respective probabilities p and $(1-p)$, for each interaction between the target node and its r active regulators. Substituting for r its expected value r_a with a mean-field approach, we obtain:

$$\overline{\alpha}_c(p,k) = p^{r_a} + (1-p)^{r_a} \ . \tag{4.8}$$

Even though our treatment is empirical Eqs. 4.7 and 4.8 produce a good fit for both r_a and $\overline{\alpha}_c$ with the values for $(r_{min}, r_{max}, \gamma)$ listed in Table 4.3. The agreement of the

		for r_a			for $\bar{\alpha}_c$		
		r_{min}	r_{max}	γ	r_{min}	r_{max}	γ
n=11	no basin	0.36	1.46	1.49	1.03	1.66	1.91
	with basin	0.19	2.69	1.54	1.02	2.55	2.64
n=110	no basin	0.21	2.55	1.44	1.03	2.26	2.42
	with basin	0.22	2.60	1.49	1.03	2.23	2.34

Table 4.3: Fit values for the model.

analytical form with the graph depicting basin-size averaged $\overline{\alpha}_c$ versus p for $n = 110$ is

shown in Fig. 4.5 On the other hand, the parameter values giving the best fit for $\overline{\alpha}_c$ and

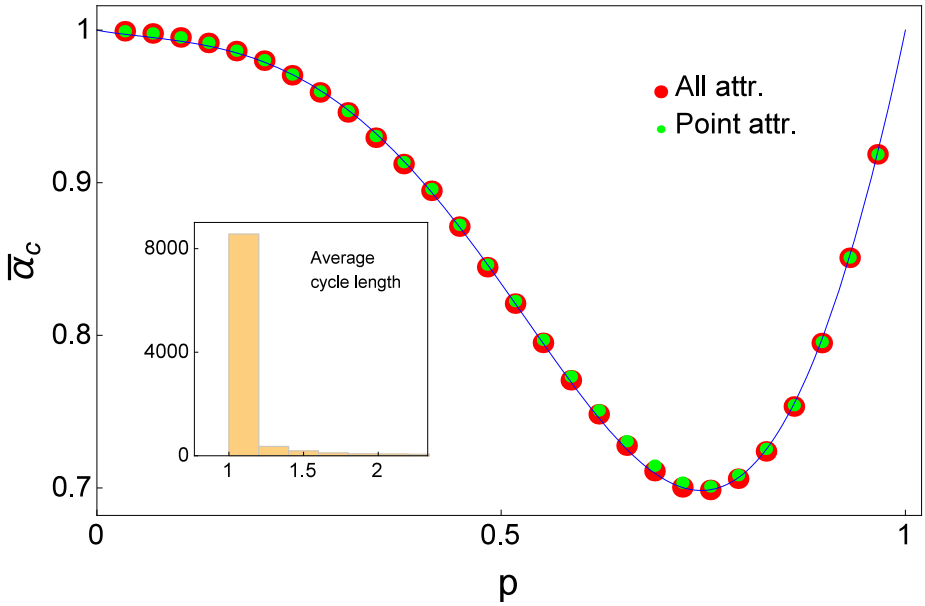


Figure 4.5: The contribution of cyclic attractors for $n = 110$ to basin-size weighted values of $\overline{\alpha}_c$. The dependence of $\overline{\alpha}_c$ on p is calculated over all attractors (red) and over point attractors only (green), for comparison. The blue curve is a fit calculated by using Eqs. 4.7 & 4.8. The average cycle length distribution for $p = 0.5$ given in the inset is typical of all the ensembles considered here, irrespective of the averaging method used.

r_a are different. As the sources of this discrepancy we can identify the nonuniformity of the in-degree distribution and the contribution of correlations, which cannot be captured by the simple treatment above.

It is not surprising that as number of neighbors per node, k, increases, $\overline{\alpha}_c$ is decreasing as shown in Fig. 4.4(c,d). That is an indication of the fact that the nodes with fewer neighbors are less likely to receive opposing inputs. The number of active neighbors per node when $p = 0.5 \simeq p_{M1}$ increases almost linearly with k, suggesting that the overall number of simultaneously active nodes remains a constant, essentially determined by p.

Our results are identical to the naked ey when non-point attractors (cycles with period > 1) are not taken into account in the calculations for α_c of random networks.

Because as it is evident from the inset of Fig. 4.5 such cycles appear infrequently under the given dynamical rules. Therefore we can conclude that the cause for the high coherence in YCC networks (M1, M2) is not the fact that they only possess point attractor.

4.3 Coherence in Various GRNs

4.3.1 Model Cell Cycle and Differentiation Networks

Having made the observation that the YCC networks operates in a highly coherent manner, we will investigate if other GRNs networks also exhibit this feature. For this aim we chose 5 other GRNs from literature, which again produce the experimentally observed cell phases in a Boolean formalism. These include the cell cycle networks in *Schizosaccharomyces pombe* (fission yeast) [6] and mammals [7], cell-differentiation networks of *Arabidopsis thaliana* whorls [9], Th lymphocyte [11] and myeloid progenitors [10]. We will present these new result along with the ones obtained by YCC (M1) for the sake of completeness. In the following section we briefly explain the investigated networks. Their structural and dynamical features are listed together in Table 4.7 at the end of this section.

Schizosaccharomyces pombe (Fission Yeast) Cell Cycle Network:

The Boolean model for the cell cycle process proposed by Davidich and Bornholdt [6] can be seen in Fig. 4.6(a). It successfully produces the experimentally observed biological dynamical sequence ending up in the steady state corresponding to G_1 phase. This state has the dominant basin of attraction in the system, as it was the case for *S. cerevisiae*. A high robustness is observed in the network in the sense that when a state in the biological pathway is perturbed, in 80% of the cases the trajectory return to G_1 phase. In addition, the implemented dynamics (Eq. 4.9) yields other 11 point attractors and one cycle of period three.

The temporal evolution is governed by threshold functions for each node. The

nodes, in the set self=$\{SK, SLP, PP\}$, which have an inhibitory self-loop and the others are treated differently such that

$$\sigma_i(t+1) = \begin{cases} 1, & \sum_j a_{ij}\sigma_j(t) > \theta_i \\ 0, & \sum_j a_{ij}\sigma_j(t) \leq 0 \wedge i \in \textbf{self} \\ & \text{or} \\ & \sum_j a_{ij}\sigma_j(t) < \theta_i \wedge i \notin \textbf{self} \\ \sigma_i(t), & \sum_j a_{ij}\sigma_j(t) = \theta_i \wedge i \notin \textbf{self,} \end{cases} \tag{4.9}$$

where $\theta_i = 0$ except for Cdc2 and Cdc* which have θ_i =1 and -1 respectively. The edges all have the same weight with $a_{ij} = \pm 1$ for a activatory or inhibitory influence from node j to node i respectively.

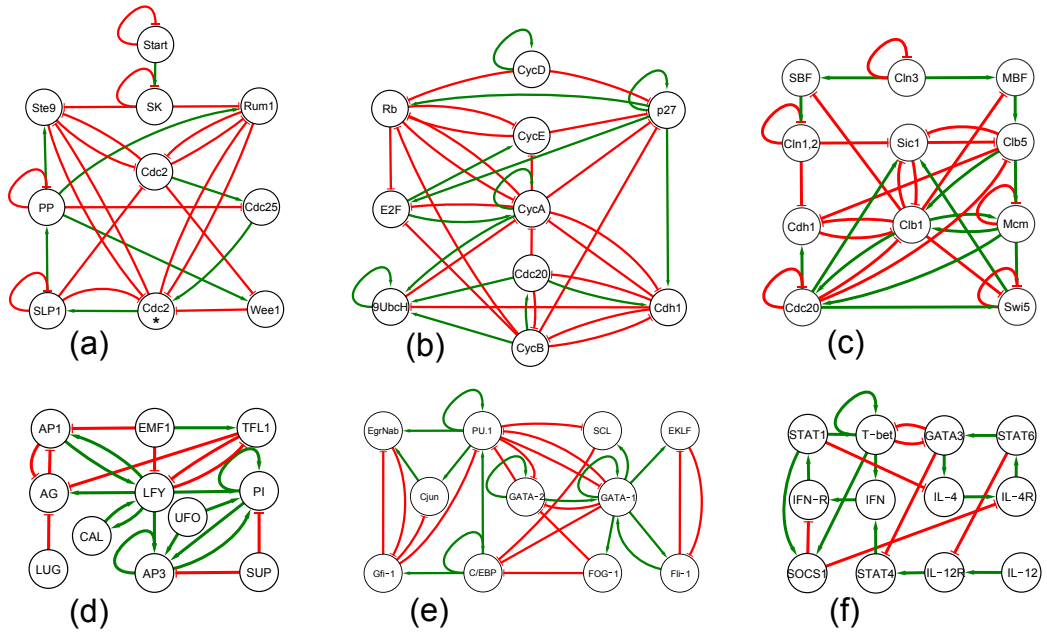


Figure 4.6: Gene regulatory networks for (a) Fission yeast cell cycle [6]; (b) Mammalian cell cycle [7]; (c) Budding yeast cell cycle [8]; (d) *Arabidopsis thaliana* whorl differentiation [9]; (e) Myeloid differentiation [10]; (f) Th-Lymphocyte differentiation [11]. The nodes represent transcription factors, proteins or protein complexes where green and red arrows stand for activation and inhibition caused by physical or genetic interactions.

Mammalian Cell Cycle Network:

To understand the advantages and limits of synchronous, asynchronous and mixed updating schemes Fauré et al. [7] constructed a logical model for the core network responsible for the mammalian cell cycle process, seen in Fig. 4.6(b). Here the logical rules for each node, listed in Table 4.4, were determined relying on experimental evidence. The resulting dynamical attractors are independent of the updating scheme and include a fixed point and a limit cycle, in agreement with the experimental expression data.

Gene	Boolean Update Functiont
CycD	$CycD$
Rb	$(\overline{CycD} \wedge \overline{CycE} \wedge \overline{CycA} \wedge \overline{CycB}) \vee (p27 \wedge \overline{CycD} \wedge \overline{CycB})$
E2F	$(\overline{Rb} \wedge \overline{CycA} \wedge \overline{CycB}) \vee (p27 \wedge \overline{Rb} \wedge \overline{CycB})$
CycE	$E2F \wedge \overline{Rb}$
CycA	$(E2F \wedge \overline{Rb} \wedge \overline{Cdc20} \wedge \overline{(Cdh1 \wedge Ubc)})$ $\vee (CycA \wedge \overline{Rb} \wedge \overline{Cdc20} \wedge \overline{(Cdh1 \wedge Ubc)})$
p27	$(\overline{CycD} \wedge \overline{CycE} \wedge \overline{CycA} \wedge \overline{CycB}) \vee (p27 \wedge \overline{(CycE \wedge CycA)} \wedge \overline{CycB} \wedge \overline{CycD})$
Cdc20_α	$CycB$
Cdh1	$(\overline{CycA} \wedge \overline{CycB}) \vee Cdc20 \vee (p27 \wedge \overline{CycB})$
UbcH10	$\overline{Cdh1} \vee (Cdh1 \wedge Ubc \wedge (Cdc20 \vee CycA \vee CycB))$
CycB	$\overline{Cdc20} \wedge \overline{Cdh1}$

Table 4.4: Update functions for mammalian cell cycle.

Arabidopsis thaliana Differentiation Network:

The dynamics of morphogenesis in the flower *Arabidopsis thaliana* was investigated with a Boolean model by Mendoza et al. [94] which is seen in Fig.4.6(d). In their model they give each edge a weight representing the various strengths of the interactions based on the genetic and molecular data, which leads to the interaction matrix in the form

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -2 & -1 & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 5 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & -2 & 1 & -2 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 & -2 \\ 0 & 0 & 4 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \begin{array}{|c|c|c|} \hline i,j & \text{Node} & \theta_i \\ \hline 1 & \text{EMF1} & 0 \\ 2 & \text{TFL1} & 0 \\ 3 & \text{LFY} & 3 \\ 4 & \text{AP1} & -1 \\ 5 & \text{CAL} & 1 \\ 6 & \text{LUG} & 0 \\ 7 & \text{UFO} & 0 \\ 8 & \text{BFU} & 1 \\ 9 & \text{AG} & -1 \\ 10 & \text{AP3} & 0 \\ 11 & \text{PI} & 0 \\ 12 & \text{SUP} & 0 \\ \hline \end{array}, \tag{4.10}$$

where θ_i are the thresholds for the corresponding nodes, which are used in the time evolution equations given as:

$$\sigma_i(t+1) = \begin{cases} 1, & \sum_j a_{ij}\sigma_j(t) > \theta_i \\ 0, & \sum_j a_{ij}\sigma_j(t) \leq \theta_i. \end{cases} \tag{4.11}$$

Four of the six point attractors arising from the dynamical rule in Eq. 4.11 can be interpreted as the patterns of four different gene activation levels corresponding to the four types of whorls (sepals, petals, stamens, carpels) seen in the flower whereas the other two can also be experimentally observed.

Myeloid Differentiation Network:

A Boolean model was set up by Krumsiek et al. [10] in order to understand the mechanisms underlying myeloid differentiation from common myeloid progenitors to megakaryocytes, erythrocytes, granulocytes and monocytes. The model network (Fig. 4.6(e)) is composed of relevant transcription factors which evolve (in time) under separate logical update functions, again inferred from the available experimental evidence. The logical rules derived from experiments are listed in Table 4.5 for each node. They give rise to 5 point attractors, where 4 are in agreement with microarray expression profiles of the mature cell types. It is pointed out that the fifth attractor cannot be realized during physiological hematopoietic differentiation.

Gene	Boolean Update Function
GATA-2	$\text{GATA-2} \wedge \overline{(\text{GATA-1} \wedge \text{FOG-1})} \wedge \overline{\text{PU.1}}$
GATA-1	$(\text{GATA-1} \vee \text{GATA-2} \vee \text{Fli-1}) \wedge \overline{\text{PU.1}}$
FOG-1	GATA-1
EKLF	$\text{GATA-1} \wedge \overline{\text{Fli-1}}$
Fli-1	$\text{GATA-1} \wedge \overline{\text{EKLF}}$
SCL	$\text{GATA-1} \wedge \overline{\text{PU.1}}$
C/EBP$_\alpha$	$\text{C/EBP}_\alpha \wedge \overline{(\text{GATA-1} \wedge \text{FOG-1} \wedge \text{SCL})}$
PU.1	$(\text{C/EBP}_\alpha \vee \text{PU.1}) \wedge \overline{(\text{GATA-1} \vee \text{GATA-2})}$
cJun	$\text{PU.1} \wedge \overline{\text{Gfi-1}}$
EgrNab	$(\text{PU.1} \wedge \text{cJun}) \wedge \overline{\text{Gfi-1}}$
Gfi-1	$\text{C/EBP}_\alpha \wedge \overline{\text{EgrNab}}$

Table 4.5: Update functions for myeloid differentiation.

Th Lymphocyte Differentiation Network:

Remy et al. [2] proposed this regulatory network model for the differentiation of T-helper lymphocytes (Th0) cells into Th1 and Th2 in the vertebrate immune system. Discrete-time evolution of each node in the model network is governed by node-specific logical rules given in Table 4.6. The network structure in Fig. 4.6(f) is dominated by activatory interactions. The dynamics settles into 3 steady states, in agreement with the gene expression levels in the Th0, Th1 and Th2 cells.

Below in Table 4.7 we list the structural and dynamical features of the considered networks, whose k varies in the range [1.67, 3.1], whereas p takes values in the interval [0.32, 0.65]. The network size is almost the same in each of them. Recall that although the parameters k and p are important determinants of global coherence, networks size n influences α_c only weakly.

Gene	Boolean Update Function†
IFN-γ	$K_1(9), K_1(11), K_1(9,11)$
IL-4	$K_2(12)$
IL-12	
IFN-γR	$K_4(1)$
IL-4R	$K_5(2)$
IL-12R	$K_6(3)$
STAT1	$K_7(4)$
STAT6	$K_8(5)$
STAT4	$K_9(6)$
SOCS1	$K_{10}(7), K_{10}(11), K_{10}(7,11)$
T-bet	$K_{11}(7), K_{11}(11), K_{11}(7,11)$
GATA-3	$K_{12}(8)$

Table 4.6: Update functions for lymphocyte differentiation. Here $K_i(\{j\})$ is the set of nodes j that when simultaneously active turn the node i on. For details one can refer to [2].

Organism/Cell type	**Func.**	**Network parameters**			**# of attr.**	
		n	k	p	all	bio.
S. pombe	c.c.	10	2.3	0.35	13	1
Mammalian cell	c.c.	10	3.1	0.32	2	2
S. cerevisiae	c.c.	11	2.64	0.51	6	1
A. thaliana whorls	diff.	11	2	0.55	6	4
Myeloid progenitor	diff.	11	2.36	0.42	5	4
Th-lymphocyte	diff.	12	1.67	0.65	6	2

Table 4.7: Structural and dynamical features associated with the six regulatory networks adopted from the literature. "c.c" and "diff." refer to regulation of cell cycle and differentiation respectively. Last two columns exclude the trivial point attractor $\boldsymbol{\sigma} = 0$.

4.3.2 Coherence: Model Networks versus Random Networks

First we calculated the degree of global coherence for each of the six biological networks over the full set of their attractors, including fixed points and cycles, (α_c) and over the subset of biologically relevant attractors (α_c^{bio}) as indicated in the corresponding references. These results are then compared with the mean of the respective randomized ensembles (α_c^{rand}), each containing 10^4 networks of the first type, where RN_1(biological network) have the same topological structure as the original one as described in Section 3.4.1. The outcomes, with uniform and basin-size weights, are summarized in Table 4.8 and in Fig. 4.7 one can see the corresponding distributions for a better judgement.

Organism/Cell type	α_c **(uniformly weighted)**			α_c **(basin-size weighted)**		
	α_c	α_c^{bio}	α_c^{rand}	α_c	α_c^{bio}	α_c^{rand}
S. pombe	0.95	1	0.94 ± 0.04	0.99	1	0.93± 0.05
Mammalian cell	0.88	0.88	0.85 ± 0.07	0.88	0.88	0.86 ± 0.07
S. cerevisiae	0.97	1	0.80 ± 0.08	0.99	1	0.75± 0.10
A. thaliana whorls	1	1	0.97 ± 0.04	1	1	0.96 ± 0.04
Myeloid progenitor	0.89	0.89	0.82 ± 0.09	0.90	0.88	0.84 ± 0.09
Th-lymphocyte	0.99	0.96	0.94 ± 0.05	0.94	0.92	0.94 ± 0.05

Table 4.8: Degrees of coherence associated with the regulatory networks shown in Fig. 4.6. α_c^{all}, α_c^{bio}, α_c^{rand}, denote the coherence value calculated by using Eq.(3.9) over all attractors of the model dynamics, only the attractors observed in the wild type, and all networks in the corresponding randomized ensemble, respectively.

The primary observation of the analysis is the bias towards high coherence of the original networks in comparison with the random networks of similar structure, even with the variability in k and p. This difference can be seen in Fig. 4.7 but for most examples they lie in the acceptable range. However, all deviations from the mean are in the positive direction suggesting an overall preference for coherence. To quantify this predisposition, the Fisher's Method [106] for statistical significance is used on the data. It yielded a likelihood of 0.044 and 0.075 of a chance to observe this much or larger deviation from the mean in uniformly and basin-weighted cases, respectively.

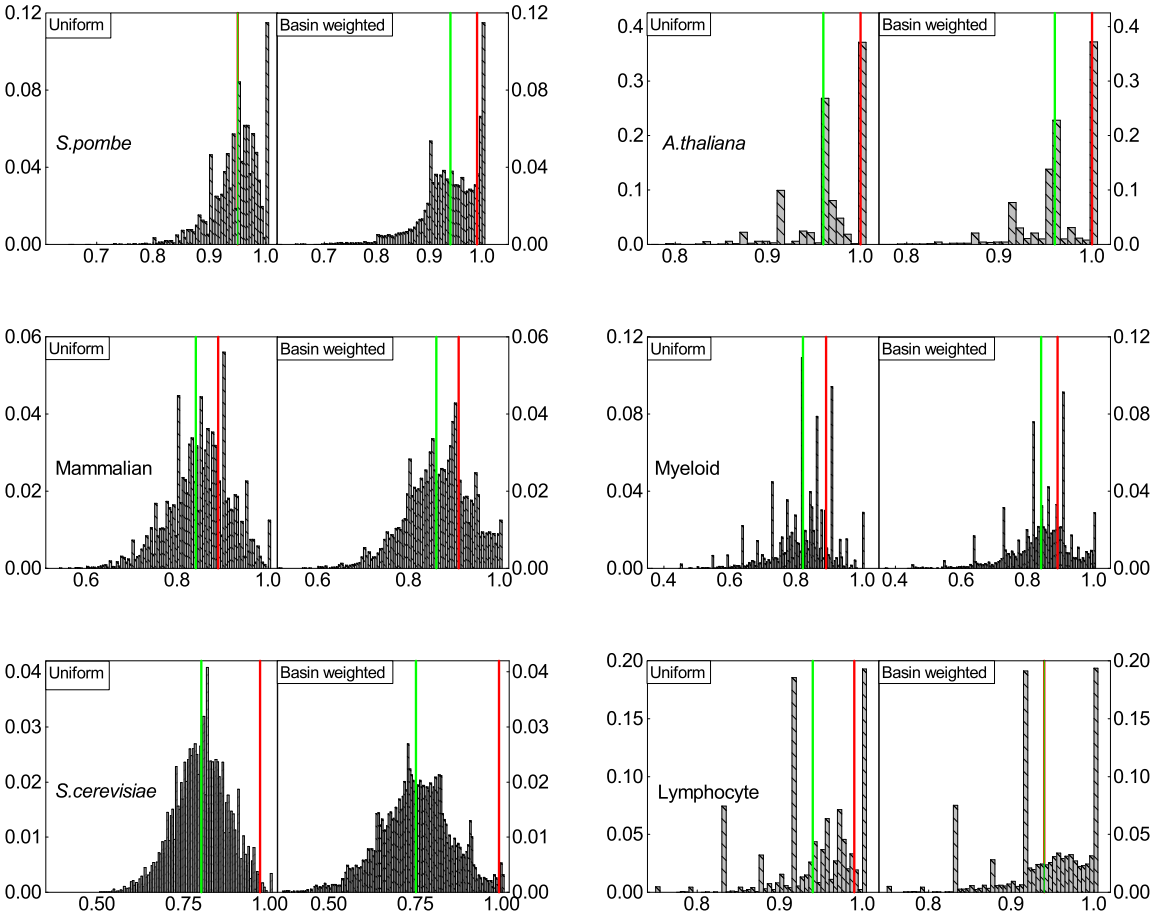


Figure 4.7: Degree of coherent regulation for biological *vs* random networks. The histograms of α_c for random ensembles of 10^4 networks having similar structure is shown with their mean values (green line). The red line indicates the α_c of the biological networks, whose exact values can be seen on Table 4.8.

The degree of bias is significant, considering the fact that besides having the same number of nodes, edges and ratio of positive and negative regulators, the random networks in comparison also have the same local architecture.

4.3.3 Coherence versus Structural Parameters

The results of the comparison with topologically similar networks brings the motivation to investigate how coherence depends on other architectural and dynamical aspects of the underlying networks. Distinguishing such relationships would help one to identify coherent systems by means of simple guiding principles without the demand of detailed knowledge of the system. Below we mention some of the possible instances.

Edge Number and Type:

To expose the dependence of coherence on the total number of edges along with the ratio of positive edges, we performed an analysis implementing the majority rule for the YCC network given in Eq. 4.2 on random networks of size 11. To do this, for each possible (k, p) pair in the range $0.909 \leq k \leq 10 \ \otimes \ 0 \leq p \leq 1$, an ensemble of 10^3 random networks of the second type, $RN_2(11, k, p)$ is generated, whose mean $\overline{\alpha}_c$ serves as the representative value for coherence in networks with parameters k and p.

In Fig. 4.8 one can observe that, independent of the average connectivity k, minimum coherence occurs when approximately two-thirds (60-72%) of the edges are negative. An intuitive understanding for this behavior was given in Section 4.2.2, where a subset of the results in Fig. 4.8 was demonstrated for the YCC network corresponding to $k = k_{M1} = 2.64$. On the same figure colored spheres indicate the six studied biological networks, which interestingly lie on the low-p slope of the minimum coherence valley.

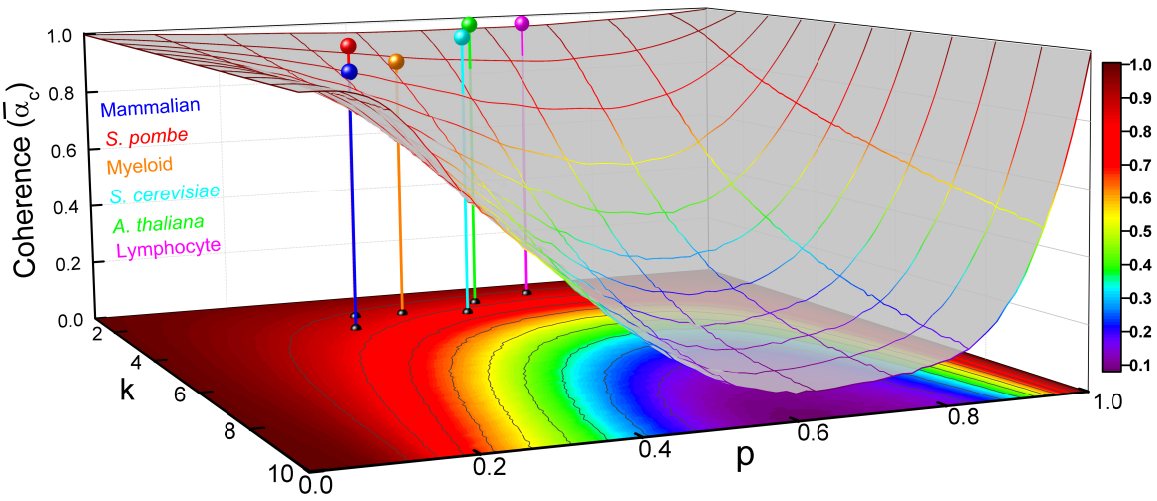


Figure 4.8: Dependence of coherence value ($\overline{\alpha}_c$) on the network parameters (k, p) with $n = 11$. As k increases from 1 to 10, the p value where $\overline{\alpha}_c$ is minimum for the fixed k shifts from 0.72 to 0.6. Coherence values corresponding to the biological networks are shown by spheres.

Number of Active Genes:

Since coherence is dictated by the active nodes, the next question to ask is whether the number of active genes (n_a) in an attractor is a determinant factor of coherent behavior. To answer this question we considered the attractors of the networks in the ensembles generated by rewiring the edges (RN_1) of each GRN in Fig. 4.6. These attractors were grouped according to their n_a values and within each group the average of coherence was calculated. The results are shown in Fig. 4.9, showing that attractors with high number of active nodes are less coherent. This can be expected since it is more difficult to achieve concordance when more components are involved. It is worth noting here that this correlation is only valid within the restricted ensembles of the first type (RN_1) random networks. Generally coherence does not depend monotonically on n_a, which is the case, for example, when the difference in n_a between two networks is caused by the difference in their respective p values. This is revealed in Fig. 4.8 where, for a fixed k, $\overline{\alpha}_c$ is non-monotonic with respect to p although $\overline{n}_a(p)$ calculated over the ensembles trivially increases with p.

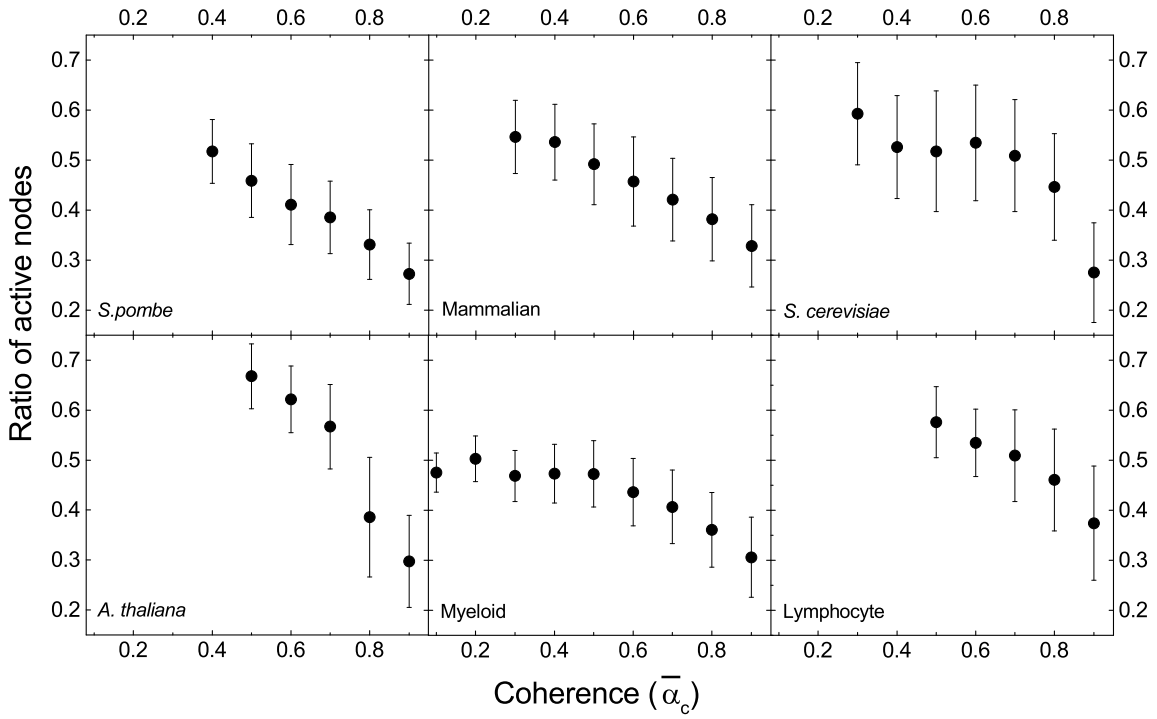


Figure 4.9: Average ratio of active nodes in the point-attractors vs. average coherence in ensembles of model networks. The average is calculated over 10^4 networks chosen randomly from the null-model ensemble described in section 3.4.1.

Basin Size:

When Table 4.8 is examined, one notices that basin weighted averages for coherence are not always higher than the uniformly weighted case, claiming that coherent attractors do not always have the biggest basins of attraction. An examination over the attractors of random ensembles for each model network does not evince a consistent dependence of coherence of an attractor on its basin-size as it is seen in Fig. 4.10. In the same way, biologically relevant attractors of the model GRNs [6, 7, 8, 94, 10, 2] are generally not the ones with the biggest basins, which is not surprising since the initial state of a cell before division or differentiation is never determined randomly.

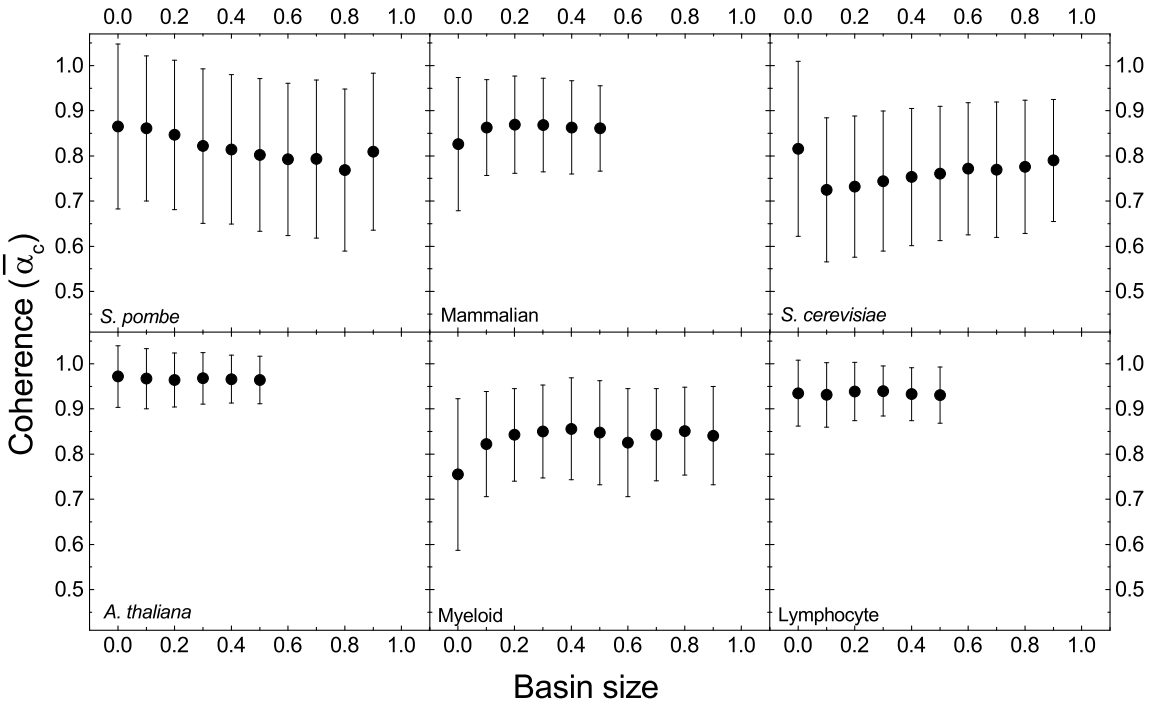


Figure 4.10: Average coherence vs. average basin size for the attractors (point or cycles) of 10^4 networks with similar structure. Note that, proposed rules of dynamics in some models appear to prohibit appearance of excessively dominant attractors.

Robustness:

Previously, in Section 3.1, we suggested that coherent attractors might me more robust due to the absence of conflicting inputs trying to change its state. Here we define the robustness of an attractor to be the ability to return to itself when one of its nodes is perturbed. For example, changing one node at a time the state $\vec{\mathbf{S}} = \{1, 1, 0\}$ can be perturbed into three different states being $\vec{\mathbf{S}}_1 = \{0, 1, 0\}$, $\vec{\mathbf{S}}_2 = \{1, 0, 0\}$ and $\vec{\mathbf{S}}_3 = \{1, 1, 1\}$. If the state $\vec{\mathbf{S}}_i$ $(i = 1, 2, 3)$ returns to the initial state $\vec{\mathbf{S}}$ under the dynamical rule of the underlying network, it is robust and has the robustness value 1 and if not its robustness value is 0. For an attractor of a network with n nodes the robustness value can then be calculated as the average of robustness values of all the n perturbed states.

The idea that robustness and coherence can be related is tested in random net-

works of type $RN_2(11, 6, 0.5)$ by using the Pearson correlation coefficient, which measures the linear correlation between to data sets. Its value changes in the range $[-1, 1]$, where -1 and 1 correspond to total negative and positive correlations respectively, where as no correlation is indicated by the value 0. The correlation of two data sets X and Y is calculated as

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}, \tag{4.12}$$

where $\mathrm{cov}(X, Y)$ is the covariance of two sets and σ_X and σ_Y denote the standard deviations. In our case the sets X and Y correspond to the sets of coherence and robustness values and the obtained Pearson coefficient is $\rho = 0.3$, which can be seen as the slope of the black line in Fig. 4.11, where the results are shown as a scatter plot and heat map of coherence versus robustness for the encountered attractors. In this figure the black squares correspond to mean coherence versus mean robustness. Regarding the correlation coefficient the results remain inconclusive to suggest a strong dependence of coherence and robustness onto each other.

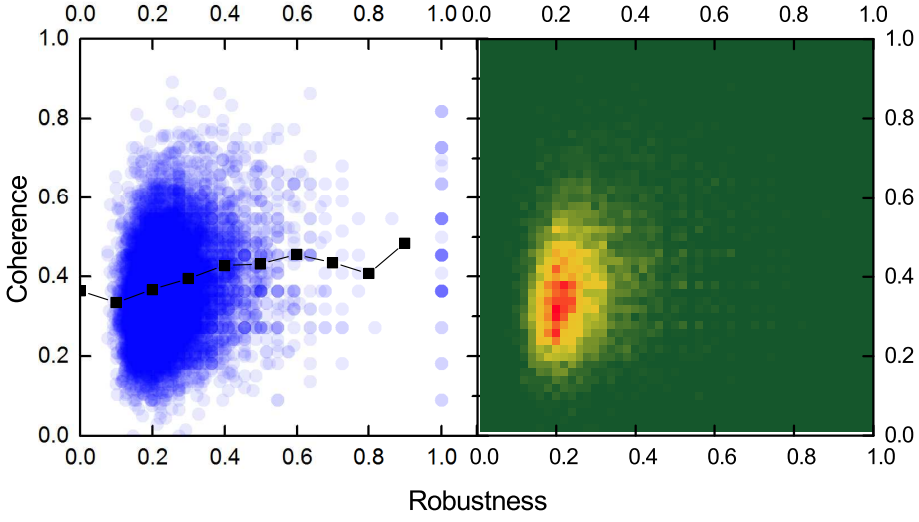


Figure 4.11: Scatter plot and heat map for coherence vs. robustness of the attractors encountered in random networks, indicating no dependence between the two.

Functional Ensembles:

For the YCC network ($M1$) we also compared its coherence with the networks that perform the same function. Functional ensembles consist of networks which, given the same initial condition, follow the identical trajectory. These networks are constructed as explained in [107]. Their coherence distribution is shown in Fig. 4.12 and it is observed that the networks with the same function as YCC ($M1$) have lower coherence values in comparison with the topologically similar networks.

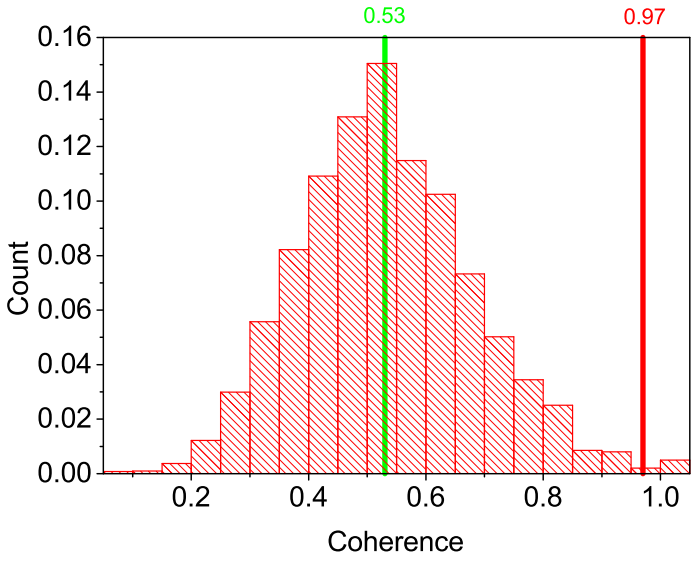


Figure 4.12: Coherence distribution for functional ensembles of YCC ($M1$). Green line indicates the mean of distribution and red line belongs to YCC.

4.4 Future Prospects

Our findings so far support that there is a selection through evolution towards coherent regulation in GRNs. To understand how this selection mechanism works one can think of a Hebbian-like mechanism [108], where interaction get eliminated over time if they lead to incoherence in the expression profile. Transcriptional regulatory networks may also be a good candidate for such a selection due to their high mutation rates among various biological networks [109].

As a future study we also plan to investigate the dependence of coherence on other structural features, such as motif statistics, directed cycles, etc. along with its relationship with controllability in networks. Understanding these aspects can enable one to build a coherent system from scratch or increase coherence in a given system with minimal changes.

BIBLIOGRAPHY

[1] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science Signalling*, 298(5594):824, 2002.

[2] Elisabeth Remy, Paul Ruet, Luis Mendoza, Denis Thieffry, and Claudine Chaouiya. From logical regulatory graphs to standard petri nets: dynamical roles and functionality of feedback circuits. *Transactions on Computational Systems Biology VII*, pages 56–72, 2006.

[3] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[4] Albert-László Barabási and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[5] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Molecular biology of the cell. new york: Garland science; 2002. *Classic textbook now in its 5th Edition*, 2002.

[6] Maria I Davidich and Stefan Bornholdt. Boolean network model predicts cell cycle sequence of fission yeast. *PLoS One*, 3(2):e1672, 2008.

[7] Adrien Fauré, Aurélien Naldi, Claudine Chaouiya, and Denis Thieffry. Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14):e124–e131, 2006.

[8] Fangting Li, Tao Long, Ying Lu, Qi Ouyang, and Chao Tang. The yeast cell-

cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14):4781–4786, 2004.

[9] Luis Mendoza, Denis Thieffry, and Elena R Alvarez-Buylla. Genetic control of flower morphogenesis in arabidopsis thaliana: a logical analysis. *Bioinformatics*, 15(7):593–606, 1999.

[10] Jan Krumsiek, Carsten Marr, Timm Schroeder, and Fabian J Theis. Hierarchical differentiation of myeloid progenitors is encoded in the transcription factor network. *PloS one*, 6(8):e22649, 2011.

[11] Elisabeth Remy and Paul Ruet. From minimal signed circuits to the dynamics of boolean regulatory networks. *Bioinformatics*, 24(16):i220–i226, 2008.

[12] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. CRC press, 2006.

[13] Kai Willadsen, Jochen Triesch, and Janet Wiles. Understanding robustness in random boolean networks. In *ALIFE*, pages 694–701, 2008.

[14] Nicholas T Ingolia. Topology and robustness in the drosophila segment polarity network. *PLoS biology*, 2(6):e123, 2004.

[15] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

[16] Isabelle S Peter and Eric H Davidson. Modularity and design principles in the sea urchin embryo gene regulatory network. *FEBS letters*, 583(24):3948–3958, 2009.

[17] René Thomas and Morten Kaufman. Multistationarity, the basis of cell differentiation and memory. i. structural conditions of multistationarity and other

nontrivial behavior. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 11(1):170–179, 2001.

[18] Yangle Wu, Xiaomeng Zhang, Jianglei Yu, and Qi Ouyang. Identification of a topological characteristic responsible for the biological robustness of regulatory networks. *PLoS computational biology*, 5(7):e1000442, 2009.

[19] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008.

[20] Shisong Ma, Qingqiu Gong, and Hans J Bohnert. An arabidopsis gene network based on the graphical gaussian model. *Genome research*, 17(11):1614–1625, 2007.

[21] Chris J Needham, Iain W Manfield, Andrew J Bulpitt, Philip M Gilmartin, and David R Westhead. From gene expression to gene regulatory networks in arabidopsis thaliana. *BMC systems biology*, 3(1):85, 2009.

[22] Stuart A Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22(3):437–467, 1969.

[23] Stuart A. Kauffman. *The origins of order: Self organization and selection in evolution*. Oxford university press, 1993.

[24] Mark Newman. *Networks: an introduction*. Oxford University Press, 2010.

[25] Paul Erdös and Alfréd Rényi. {On the evolution of random graphs}. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

[26] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[27] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[28] David E Featherstone and Kendal Broadie. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *Bioessays*, 24(3):267–274, 2002.

[29] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[30] Andreas Wagner and David A Fell. The small world inside large metabolic networks. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1478):1803–1810, 2001.

[31] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64–68, 2002.

[32] Réka Albert, István Albert, and Gary L Nakarado. Structural vulnerability of the north american power grid. *Physical review E*, 69(2):025103, 2004.

[33] Shuguang Li, Jianping Yuan, and Juan Cristóbal Zagal. Encouraging networks modularity by seeding motifs. In *Proceedings of the 2011 European Conference on Artificial Life*, page 455, 2011.

[34] John J Tyson and Béla Novák. Functional motifs in biochemical reaction networks. *Annual review of physical chemistry*, 61:219, 2010.

[35] Volkan Sevim and Per Arne Rikvold. Chaotic gene regulatory networks can be robust against mutations and noise. *Journal of theoretical biology*, 253(2):323–332, 2008.

[36] Olivier Cinquin and Jacques Demongeot. Roles of positive and negative feedback in biological systems. *Comptes rendus biologies*, 325(11):1085–1095, 2002.

[37] René Thomas, Denis Thieffry, and Marcelle Kaufman. Dynamical behaviour of biological regulatory networksi. biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bulletin of mathematical biology*, 57(2):247–276, 1995.

[38] Adrien Richard and Jean-Paul Comet. Necessary conditions for multistationarity in discrete dynamical systems. *Discrete Applied Mathematics*, 155(18):2403–2413, 2007.

[39] Christophe Soulé. Graphic requirements for multistationarity. *arXiv preprint q-bio/0403033*, 2004.

[40] Jean-Luc Gouzé. Positive and negative circuits in dynamical systems. *Journal of Biological Systems*, 6(01):11–15, 1998.

[41] Marcelle Kaufman and René Thomas. Emergence of complex behaviour from simple circuit structures. *Comptes rendus biologies*, 326(2):205–214, 2003.

[42] Denis Thieffry. Dynamical roles of biological regulatory circuits. *Briefings in bioinformatics*, 8(4):220–225, 2007.

[43] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[44] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[45] Ernesto Estrada. *The structure of complex networks: theory and applications*. Oxford University Press, 2011.

[46] Björn H Junker and Falk Schreiber. *Analysis of biological networks*, volume 2. John Wiley & Sons, 2011.

[47] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein protein interactions. 1989.

[48] Pengjun Pei. *Graph-based analysis of protein-protein interaction data sets*. ProQuest, 2007.

[49] Peter Uetz, Loic Giot, Gerard Cagney, Traci A Mansfield, Richard S Judson, James R Knight, Daniel Lockshon, Vaibhav Narayan, Maithreyan Srinivasan, Pascale Pochart, et al. A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

[50] Loic Giot, Joel S Bader, C Brouwer, Amitabha Chaudhuri, Bing Kuang, Y Li, YL Hao, CE Ooi, Brian Godwin, E Vitols, et al. A protein interaction map of drosophila melanogaster. *Science*, 302(5651):1727–1736, 2003.

[51] Siming Li, Christopher M Armstrong, Nicolas Bertin, Hui Ge, Stuart Milstein, Mike Boxem, Pierre-Olivier Vidalain, Jing-Dong J Han, Alban Chesneau, Tong Hao, et al. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657):540–543, 2004.

[52] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[53] Jeongah Yoon, Anselm Blumer, and Kyongbum Lee. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics*, 22(24):3106–3108, 2006.

[54] Julien Gagneur, David B Jackson, and Georg Casari. Hierarchical analysis of dependency in metabolic networks. *Bioinformatics*, 19(8):1027–1034, 2003.

[55] Robert D Leclerc. Survival of the sparsest: robust gene networks are parsimonious. *Molecular systems biology*, 4(1):213, 2008.

[56] Z Burda, A Krzywicki, OC Martin, and M Zagorski. Motifs emerge from function in model gene regulatory networks. *Proceedings of the National Academy of Sciences*, 108(42):17263–17268, 2011.

[57] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[58] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, 1995.

[59] International Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[60] Stuart Kauffman. *At home in the universe: The search for the laws of self-organization and complexity*. Oxford University Press, USA, 1995.

[61] Michael PH Stumpf, Thomas Thorne, Eric de Silva, Ronald Stewart, Hyeong Jun An, Michael Lappe, and Carsten Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.

[62] Sequencing Consortium et al. {Genome sequence of the nematode *C.elegans*: A platform for investigating biology}. *Science*, 282:2012–2018, 1998.

[63] Matthew K Vickaryous and Brian K Hall. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biological reviews*, 81(03):425–455, 2006.

[64] Brenton R Graveley. Alternative splicing: increasing diversity in the proteomic world. *TRENDS in Genetics*, 17(2):100–107, 2001.

[65] Eugene V Koonin, Yuri I Wolf, and Georgy P Karev. *Power laws, scale-free networks and genome biology*. Springer, 2006.

[66] Kai Xia, Zheng Fu, Lei Hou, and Jing-Dong J Han. Impacts of protein–protein interaction domains on organism and network complexity. *Genome research*, 18(9):1500–1508, 2008.

[67] Andrew Wood Murray and Tim Hunt. *The cell cycle: an introduction*, volume 251. Oxford University Press New York, 1993.

[68] Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.

[69] Ilya Shmulevich and Edward R Dougherty. *Probabilistic Boolean networks: the modeling and control of gene regulatory networks*. siam, 2010.

[70] Maria Davidich and Stefan Bornholdt. The transition from differential equations to boolean networks: a case study in simplifying a regulatory network model. *Journal of Theoretical Biology*, 255(3):269–277, 2008.

[71] Elena R Alvarez-Buylla, Álvaro Chaos, Maximino Aldana, Mariana Benítez, Yuriria Cortes-Poza, Carlos Espinosa-Soto, Diego A Hartasánchez, R Beau Lotto, David Malkin, Gerardo J Escalera Santos, et al. Floral morphogenesis: stochastic explorations of a gene network epigenetic landscape. *Plos one*, 3(11):e3626, 2008.

[72] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–635, 2014.

[73] James B Russell and Gregory M Cook. Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiological reviews*, 59(1):48–62, 1995.

[74] Frank Buttgereit and Martin D Brand. A hierarchy of atp-consuming processes in mammalian cells. *Biochem. J*, 312:163–167, 1995.

[75] Xueli Zhang, Kaemwich Jantama, Jonathan C Moore, Laura R Jarboe, Keelnatham T Shanmugam, and Lonnie O Ingram. Metabolic evolution of energy-conserving pathways for succinate production in escherichia coli. *Proceedings of the National Academy of Sciences*, 106(48):20180–20185, 2009.

[76] Jeremy E Niven and Simon B Laughlin. Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology*, 211(11):1792–1804, 2008.

[77] Simon B Laughlin, Rob R de Ruyter van Steveninck, and John C Anderson. The metabolic cost of neural information. *Nature neuroscience*, 1(1):36–41, 1998.

[78] Pauli Rämö, Juha Kesseli, and Olli Yli-Harja. Stability of functions in boolean models of gene regulatory networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 15(3):034101–034101, 2005.

[79] Andreas Wagner. Energy constraints on the evolution of gene expression. *Molecular biology and evolution*, 22(6):1365–1374, 2005.

[80] Zheng Eelderink-Chen, Gabriella Mazzotta, Marcel Sturre, Jasper Bosman, Till Roenneberg, and Martha Merrow. A circadian clock in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 107(5):2043–2047, 2010.

[81] Ozgur E Akman, Steven Watterson, Andrew Parton, Nigel Binns, Andrew J Millar, and Peter Ghazal. Digital clocks: simple boolean models can quantitatively describe circadian systems. *Journal of The Royal Society Interface*, 9(74):2365–2382, 2012.

[82] Adrien Fauré and Denis Thieffry. Logical modelling of cell cycle control in eukaryotes: a comparative study. *Mol. BioSyst.*, 5(12):1569–1581, 2009.

[83] Kai-Yeung Lau, Surya Ganguli, and Chao Tang. Function constrains network architecture and dynamics: A case study on the yeast cell cycle boolean network. *Physical Review E*, 75(5):051907, 2007.

[84] Tadao Murata. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580, 1989.

[85] Claudine Chaouiya, Elisabeth Remy, Paul Ruet, and Denis Thieffry. Qualitative modelling of genetic networks: From logical regulatory graphs to standard petri nets. *Applications and Theory of Petri Nets 2004*, pages 137–156, 2004.

[86] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008.

[87] Kuang-Chi Chen, Tse-Yi Wang, Huei-Hun Tseng, Chi-Ying F Huang, and Cheng-Yan Kao. A stochastic differential equation model for quantifying transcriptional regulatory network in saccharomyces cerevisiae. *Bioinformatics*, 21(12):2883–2890, 2005.

[88] Ting Chen, Hongyu L He, George M Church, et al. Modeling gene expression with differential equations. In *Pacific symposium on biocomputing*, volume 4, page 4, 1999.

[89] Katherine C Chen, Laurence Calzone, Attila Csikasz-Nagy, Frederick R Cross, Bela Novak, and John J Tyson. Integrative analysis of cell cycle control in budding yeast. *Molecular biology of the cell*, 15(8):3841–3862, 2004.

[90] Maria Davidich and Stefan Bornholdt. From differential equations to boolean networks: A case study in modeling regulatory networks. *arXiv preprint arXiv:0807.1013*, 2008.

[91] Réka Albert and Hans G Othmer. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in¡ i¿ drosophila melanogaster¡/i¿. *Journal of theoretical biology*, 223(1):1–18, 2003.

[92] Vitaly V Gursky, John Reinitz, and Alexander M Samsonov. How gap genes make their domains: An analytical study based on data driven approximations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 11(1):132–141, 2001.

[93] Jean-Christophe Leloup and Albert Goldbeter. A model for circadian rhythms in drosophila incorporating the formation of a complex between the per and tim proteins. *Journal of biological rhythms*, 13(1):70–87, 1998.

[94] Luis Mendoza and Elena R Alvarez-Buylla. Dynamics of the genetic regulatory network for *Arabidopsis thaliana* flower morphogenesis. *Journal of theoretical biology*, 193(2):307–319, 1998.

[95] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[96] Mark J Hickman and Fred Winston. Heme levels switch the function of hap1 of saccharomyces cerevisiae between transcriptional activator and transcriptional repressor. *Molecular and cellular biology*, 27(21):7414–7424, 2007.

[97] Mark Ptashne. Principles of a switch. *Nature chemical biology*, 7(8):484–487, 2011.

[98] Kazunari Kaizu, Samik Ghosh, Yukiko Matsuoka, Hisao Moriya, Yuki Shimizu-Yoshida, and Hiroaki Kitano. A comprehensive molecular interaction map of the budding yeast cell cycle. *Molecular systems biology*, 6(1), 2010.

[99] Peter E Hodges, Andrew HZ McKee, Brian P Davis, William E Payne, and James I Garrels. The yeast proteome database (ypd): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Research*, 27(1):69–73, 1999.

[100] Gary D Bader, Adrian Heilbut, Brenda Andrews, Mike Tyers, Timothy Hughes, and Charles Boone. Functional genomics and proteomics: charting a multidimensional map of the yeast cell. *Trends in cell biology*, 13(7):344–356, 2003.

[101] Miguel Cacho Teixeira, Pedro Tiago Monteiro, Joana Fernandes Guerreiro, Joana Pinho Gonçalves, Nuno Pereira Mira, Sandra Costa dos Santos, Tânia Rodrigues Cabrito, Margarida Palma, Catarina Costa, Alexandre Paulo Francisco, et al. The yeastract database: an upgraded information system for the analysis of gene and genomic transcription regulation in saccharomyces cerevisiae. *Nucleic acids research*, 42(D1):D161–D166, 2014.

[102] Wen-Bin Lee and Jiun-Yan Huang. Robustness and topology of the yeast cell cycle boolean network. *FEBS letters*, 583(5):927–932, 2009.

[103] Yurie Okabe and Masaki Sasai. Stable stochastic dynamics in yeast cell cycle. *Biophysical journal*, 93(10):3451, 2007.

[104] Karan Mangla, David L Dill, and Mark A Horowitz. Timing robustness in the budding and fission yeast cell cycles. *PloS one*, 5(2):e8906, 2010.

[105] Changki Hong, Minho Lee, Dongsup Kim, Dongsan Kim, Kwang-Hyun Cho, and Insik Shin. A checkpoints capturing timing-robust boolean model of the

budding yeast cell cycle regulatory network. *BMC systems biology*, 6(1):129, 2012.

[106] Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

[107] Kai-Yeung Lau, Surya Ganguli, and Chao Tang. Function constrains network architecture and dynamics: A case study on the yeast cell cycle boolean network. *Physical Review E*, 75(5):051907, 2007.

[108] RGM Morris. Do hebb: The organization of behavior, wiley: New york; 1949. *Brain research bulletin*, 50(5):437, 1999.

[109] Chong Shou, Nitin Bhardwaj, Hugo Y Lam, Koon-Kiu Yan, Philip M Kim, Michael Snyder, and Mark B Gerstein. Measuring the evolutionary rewiring of biological networks. *PLoS Comput Biol*, 7(1):e1001050–e1001050, 2011.

VITA

Neşe Aral was born in Istanbul, Turkey on February 14, 1982. She graduated from Sankt Georg's Austrian High School, Istanbul, in 2001. She received her B.Sc. degrees in Physics and Mathematics from Istanbul Technical University in June 2006. In Koç University, she was accepted to Master program in Physics in August 2006, and to Doctorate program in Physics in January 2009, from which she graduated in 2015.