THE STRUCTURE AND DYNAMICS OF GENE REGULATION NETWORKS

by

Murat Tuğrul

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

Master of Science

in

Computational Sciences & Engineering

Koç University

December, 2007

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Murat Tuğrul

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:

_____

Assist. Prof. Alkan Kabakçıoğlu

_____

Assist. Prof. Deniz Yüret

_____

Assoc. Prof. Attila Gürsoy

Date: _____

To Peace

ABSTRACT

The structure and dynamics of a typical biological system are complex due to strong and inhomogeneous interactions between its constituents. The investigation of such systems with classical mathematical tools, such as differential equations for their dynamics, is not always suitable. The graph theoretical models may serve as a rough but powerful tool in such cases.

In this thesis, I first consider the network modeling for the representation of the biological systems. Both the topological and dynamical investigation tools are developed and applied to the various model networks. In particular, the attractor features' scaling with system size and distributions are explored for model networks. Moreover, the theoretical robustness expressions are discussed and computational studies are done for confirmation.

The main biological research in this thesis is to investigate the transcriptional regulation of gene expression with synchronously and deterministically updated Boolean network models. I explore the attractor structure and the robustness of the known interaction network of the yeast, *Saccharomyces Cerevisiae* and compare with the model networks. Furthermore, I discuss a recent model claiming a possible root to the topology of the yeast's gene regulation network and investigate this model dynamically.

The thesis also included another study which investigates a relation between folding kinetics with a new network representation, namely, the incompatibility network of a protein's native structure. I showed that the conventional topological aspects of these networks are not statistically correlated with the phi-values, for the limited data that is available.

ÖZETÇE

Tipik bir biyolojik sistemin yapısı ve dinamiği, ögelerinin birbirleri ile homojen olmayan ve güçlü etkileşimleri sebebiyle karmaşıktır. Dinamik incelemelerde kullanılan türevli denklemler gibi klasik diyebileceğimiz matematiksel yöntemler, bu tür karmaşık sistemlerin incelenmesinde her zaman uygun olmayabilir. Çizge kuramsal modeller ise daha yüzeysel olsa da bu tür sistemlerin incelenmesi için daha etkili bir yöntem olabilir.

Bu tezde, ilk olarak biyolojik sistemlerin sunumu için ağ modellemesi ele aldım. Topolojik ve dinamik inceleme araçları geliştirilip çeşitli model ağlara uyarlandı. Özelde, model ağlar için çekici özelliklerinin sistem büyüklüğü ile ölçeklenmesi ve dağılımları incelendi. Ayrıca, kuramsal dayanıklılık ifadeleri tartışıllıp ve hesaplamalı olarak doğrulukları sınandı.

Bu tezdeki ana biyoloji araştırması, transkripsiyonel gen ifadesinin düzenlenmesinin eşzamanlı ve deterministik güncellenen Boolyan ağ modeli ile incelenmesi olmuştur. Etkileşim ağı bilinen maya, *Saccharomices Cerevisiae*'nın çekici yapısını ve dayanıklılığını inceledim ve model ağlar ile karşılaştırdım. Ayrıca, mayanın gen ifadesi ağının topolojik muhtemel temellerini irdeleyen yeni bir modeli tartışdım ve bu modeli dinamik olarak inceledim.

Bu tezde ayrıca bir başka ağ modellenmesi olan; asıl protein yapısından elde edilen bağdaşmaz (incompatibility) ağ ile protein kinetiğinin incelenmesi yer almaktadır. Elimizdeki sınırlı veri ile yapılan sınamalarda geleneksel olarak kullanılan belirli topolojik özellikler ile fi-değerleri arasında bağıntı olmadığını gösterdim.

ACKNOWLEDGMENTS

First of all, I would like to express my gratidute to my advisor, Assist. Prof. Alkan Kabakçıoğlu, whose understanding, patience and help added considerably to my graduate experience. I would like to thank other members of my committee Assist. Prof. Deniz Yüret and Assoc. Prof. Attila Gürsoy for critical readings of this thesis and for their valuable comments. A special thank goes to Osman Nuri Yoğurtçu for his effort in reading and critiques. I would like to thank Elif Müjen Şencan for being near me during this master period. Lastly, I would like to thank all my friends, professors, students, workers at Koç University and the people of Sarıyer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PROLOGUE

I remember my scientific interest related to biology at first started in my second year in physics undergraduate. I was very enthusiastic about arranging a sort of scientific article reading group with my classmates at that time. For the first of reading I was searching through the internet for a scientific article that is not very complicated so that our education let us understand the concepts. By luck I found an article[1] about the biology of human hearing and its mathematical modeling. I was impressed by this marvelous organization in the ear and made the article be our initiator reading-piece.

Unfortunately, this reading group did not gather for the second time but helped me understand deeply two important things. The first is that it is not easy and recommendable to do something "social" with the physicists. The second and more related to this context is that biology is not scary as I used to consider, on the contrary, it seems to encompass many bright inquiries about the nature.

In the following period up to the last grade in undergraduate, my interest in biology had increased gradually. I remember some of my popular scientific readings at that period: Schrödinger's book "What is Life?", Watson's book "Double Helix", Dawkins' book "Selfish Gene", etc.. I had taken some courses related to biology and ecology. At last grade I had already been sure to pursue in the life sciences in academy.

Then in September 2005, I started my master degree in Computational Sciences and Engineering program at Koç University. This program has let me appreciate some necessary knowledge about computation and given a chance to do research in biology. In particular, I have investigated the protein folding problem and the transcriptional gene regulation in my thesis. And now I introduce you my studies and research throughout this thesis.

But before passing to the thesis, at this moment, let me ask myself some very basic questions and present the answers so that you can see where I am standing and where I am

[1] Unfortunately, I do not remember exact reference.

looking through the biology.

- *What is Biology?*

For me, the biology is a branch of science which tries to answer the questions gathered around a very philosophical question: "What is life?"

- *How did this branch of science emerge, what is its history in brief?*

I think the history of human knowledge on biology has no starting time since every species must have some information regarding their own and other organisms for surviving in evolutionary period. However, our knowledge on this history at the preliterate time is suspicious and mainly relies on the guesses. If we want to have a more concrete idea emerging from evidences or documents we should go back to Egypt at the era of 3500-1500 BC (at least for the western history). But, for the sake of the reader's wonder it might be worthy to mention that people of preliterate ages were able to classify the animals and plants, to say which plants are/are not toxic, are suitable for some basic medical purposes.

We know that starting from Egypt (3500-1500 BC) the biological knowledge had increased in developed civilizations of the time but merely due to practical needs, i.e. anatomy and agriculture. Also, it was very mixed with mystery, magic and superstition. As long as I know, the first examples of what today we call "the scientific studies" (more abstract questions) regarding the biology belong to ancient Greece. Later, while we see a progress in the knowledge of living things in Arab-Islam civilizations and in Europe of Renaissance I think the actual roots of today's biology belong to Mendel's genetics and Darwin's evolution[2].

When we come to the 20th century with the development in technology, we witness many improvements in biology, such as exploring the DNA as a genetic material. Today, we even decipher the genetic code of many species and possess tremendous data. With an analogy to physics, biology seems to stand at the point of physics at the beginning of 20th century.

- *How about the methodology of biology?*

As a physicist by training, biology seems to be an almost pure empirical discipline to me, however, especially for the last decades there has been a tremendous flow of scientists

[2]See References [1, 2] for further readings.

into the life sciences from other disciplines whose methodology depends mainly upon the theoretical studies. Therefore, we can interpret this as the methodology of biology is just in change.

- *What are the main difficulties of biology today?*

The biological systems, such as cells, are very complex in terms of both structure and dynamics. In other words, there are many variables that are interacting with many other variables in time and space. Today's classical approach in science is limited and it seems to me as the main difficulty in biology. Moreover, there are tremendous data which are waiting to be analyzed, however, there seems to be not a unification for collecting and interpreting these data. That has been at least very confusing and challenging for me while surveying.

- *What is the future of biology?*

There is of course not a direct answer to this question but it is seen from the avalanche of scientists on biology that this century will see many developments in the knowledge of living things.

Murat Tuğrul

Sarıyer, October 2006

Chapter 1

INTRODUCTION

The development and use of technology have accumulated a vast amount of experimental information for biology, such as DNA sequences of different species. However, our knowledge on how a cell works remains largely unexplored [3, 4].

An important component of functional organization in the cell is the regulation of gene expression. Many interacting gene pairs of some organisms were identified with a high accuracy [5], especially for the yeast, *Saccharomyces Cerevisiae* [6]. The networks, or with a mathematical terminology: graphs [7], serve as a simple but powerful mathematical representation of the regulation of the gene expression, i.e. gene regulation networks (GRN). Many topological tools have been developed for the investigation of the networks [8, 9, 10, 11, 12, 4] and they have been already applied to the yeast and other GRNs extensively [13, 14, 5].

It is known that the genes of eukaryotes are not always active [15] and their activation profiles show a very complex dynamical aspect beacuse of the strong and inhomogeneous interactions. As a consequence, the studying the dynamics of a gene regulation is not easy with the classical dynamical investigation tools like differential equations. In gene regulation literature, the deterministically and synchronously updated boolean networks have been used widely for the dynamical investigations [16, 17]. The boolean model is based on a 1/0 binary representation of the individuals at discrete time steps. Though such modelings are approximate [18], it has been shown that some applications predicted the wild and mutant phenotypes correctly [19, 20, 21].

Another challenge for the dynamics of gene regulation is that the rules (functions) that govern the interactions are not known in detail, therefore; the statistical approaches with random functions gain importance. Some experimental studies established some canalazing behaviors in the regulation functions [22]. Later, it was argued that a subset of canalazing

functions which was named as the nested canalazing function exists in yeast gene regulation [23]. A very recent study which depended on a logical formalism (AND and OR) claimed that two subclasses of the nested canalazing functions are dominant in the yeast [24].

Since the size of state space 2^N, where N is the number of individuals, is finite quantity with a boolean approach, a deterministically and synchronously updated dynamics fall into the state cycles which are called *attractors*. The attractors are used for investigation of the dynamics and it is argued that attractors in GRN correspond to some cycles in the cells such as phenotype [19, 20]. In particular, the number and length of attractors, the average transient length to the attractors and the basin of attractions are explored. Another notion for the dynamics is the *robustness* of the system against the perturbations. There is a hypothesis related to robustness of the living systems; "Life at the edge of the chaos" which states that biological systems are to be robust against the perturbations but at the same time must be able to adapt the environments in order to be successful in evolution [17, 25, 26].

In order to simulate gene regulation, artificially created network models have been used in literature, to my knowledge, starting from Kauffman [16]. Many dynamical studies using the boolean approach have been performed with the model networks. A highly used model was called "Kauffman Networks" which has N nodes and exactly 2 incoming edges (in this thesis, this model is the in-NK network with $K = 2$). For these model networks, it was believed that the number and length of attractors scale with $\sqrt{N}$ with random functions, but Socolar & Kauffman recently argued that the number of attractors scales faster than linear [27]. Apart from the attractors, the robustness of in-NK model with random functions which can be explained analytically by Derrida & Pomeau [28] was extensively studied in literature [17].

In this thesis, the theoretical background is discussed first and applied to the artificially constructed model networks; of in-NK type, which has N nodes and exactly K indegree edges for each node; of in-PL type, which has a power-law indegree distribution and argued to be found in the natural systems often [17]; of in-EXP type, which has an exponential indegree distribution as in yeast GRN [13]. Upon investigation with conventional topological tools, the dynamical investigations take part for those model networks. In order to be able to compare the results with the literature, the investigations are performed with the structural

parameter of $\langle k_{in} \rangle = 2.0$ and with the boolean functions parameter $p = 0.5$, which is the probability for assigning a gene to be expressed [29]. The distribution of attractor features show a power-law decay [29, 30]. Also, it is observed that simple random functions produced larger average values of attractor features for these model networks than other canalazing types do. Apart from the attractor structures, the robustness was investigated. It is shown that Derrida's robustness expression for random functions [28, 17] predicts the robustness values of the model networks within a finite-size effect.

The yeast GRN is dynamically investigated with exploring the attractor and robustness structures and compared to the model networks. The average and distribution of the yeast attractor features were compared for different types of functions. It has been seen that the special subclasses of nested canalazing functions produce high number of attractors and short length of attractors and transients in the dynamics realizations. Also, it is observed that the distributions of the number of attractors and entropy are not typical and show a different profile unobserved before. Li *et al.* stated that the yeast GRN is robustly designed [31]. The robustness of the yeast GRN with various functions and p-values is computed. Morover, the model networks whose degree distributions are similar to the yeast are compared with the yeast dynamics and it is shown that those networks fail to mimic the attractor features while predicting the robustness structure correctly. Furthermore, a recent model, which is called in the thesis as Balcan *et al.* model, is discussed and some of the the results were reproduced [32]. It is shown that while Balcan *et al.* model is very successful for producing networks that are topologically similar to the yeast, it fails to mimic both attractor and robustness structures of the yeast GRN.

The thesis also included another study which investigates a relation between folding kinetics with a new network representation, namely, the incompatibility network of a protein's native structure. Starting from the description of the protein and the protein folding problem, I discuss a novel approach to the problem proposed by my thesis advisor Assist. Prof. Alkan Kabakçıoğlu and tried to investigate the relation between proteins's structure and folding kinetics. I showed that the conventional topological aspects of these networks are not statistically correlated with the ϕ-values, for the limited data that is available.

Outline: There are 5 chapters in this thesis. Chapter 2 gives the theoretical and computational backgrounds and some applications with model networks. Chapter 3 consists of topological and dynamical investigation of yeast gene regulation network with a comparison to exponential and Balcan*et al.* model networks. Chapter 4 is for the investigation of a new approach to the protein folding problem. The final part of the thesis includes conclusions and appendices that contain extra informations.

Chapter 2

NETWORK MODELING

This chapter is organized as follows: Section 2.1 is an brief explanation of Graph Theory which is the mathematical roots of networks. Section 2.2 is an overview of conventional tools for topological analyse of networks and gives some applications with model networks. Section 2.3 constructs the necessary modeling and tools for dynamics, and applies to model networks.

2.1 Graph Theory

Considering the dynamics and structure of *complex systems*, we need better mathematical tools than ordinary ones, such as differential equations, to deal with these systems [8]. For such a purpose, *networks* are used in scientific representation of complex systems.

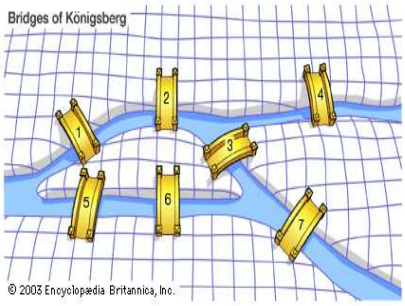

Figure 2.1: The problem is whether it is possible for a citizen of Königsberg (today's Kaliningrad) to start from somewhere in the city and cross the bridges exactly once and return to initial place.

A more mathematical term for "Network" is "Graph" and all tools of network modeling can be originated from *Graph Theory* which is a popular subdiscipline of mathematics at present. It is believed that the starting point of Graph Theory goes back to 1736: Euler's negative proof to the famous *Königsberg Bridge Problem* (Figure 2.1[1]). Many developments

[1]The figure is taken from http://www.britannica.com/eb/article-9384377/Konigsberg-bridge-problem.

in graph theory have been achieved in the previous century with an increasing interest from other sciences [7].

Here, I emphasize only the fundamental concepts related to my thesis. The definitions of Graph Theory are given as follows (as adapted from West [7]):

Graph: A graph (or network) **G** is a triple consisting of a node (or vertex) set **V(G)**, an edge set **E(G)**, and a relation that associates with each edge two vertices (not necessarily distinct) called its endpoints.

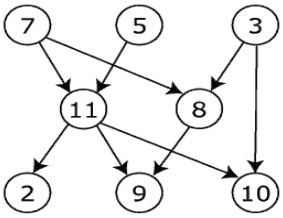

Figure 2.2: A simple graph (network); numbered objects in the figure are vertices (nodes) and the arrows are directed edge.

Loop, Adjacent and Neighbor: A **loop** is an edge whose endpoints are equal. When the nodes u and v are the endpoints of an edge, they are **adjacent** and **neighbors**, and will be shown as $u \leftrightarrow v$

Path and Cycle: A **path** is a simple graph whose nodes can be ordered so that two nodes are adjacent if and only if they are consecutive in the list. A **cycle** is a graph with an equal number of nodes and edges whose nodes can be placed around a circle so that two nodes are adjacent if and only if if they appear consecutively along the circle.

Subgraph, Connectedness, Cluster: A **subgraph** of a graph G is a graph H such that $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$ and the assignment of endpoints to edges in H is the same as in G. We then write $H \subseteq G$ and say that "G contains H". A graph is connected if each pair of nodes in G belongs to a path; otherwise, G is disconnected. A **cluster** is a connected subgraph in a graph and has no edge to nodes which are not in this subgraph.

Incident, Degree, Isolated node: If node v is an endpoint of edge e, then v and e are **incident**. The **degree** of node v, $d(v)$, is the number of incident

edges. An **isolated node** is a node of degree 0.

Directed Graph (or Digraph): A **directed graph** or **digraph** G is a triple consisting of a node set V(G), an edge set E(G), and a function assigning each edge to an ordered pair of nodes. We say that there is an edge **from** v_i **to** v_j and show it as $v_i \rightarrow v_j$.

Component, Adjacency Matrix: The **components** of a graph G are its maximal connected subgraphs. The **adjacency matrix** of G, written A(G), is the n-by-n matrix in which entry $A_{i,j}$ is the number of edges in G with $v_i \rightarrow v_j$.

Outdegree, Indegree: Indegree of a node v, $d(v)_{in}$, is the number of edges into v. Outdegree of v, $d(v)_{out}$, is the number of edges from v.

2.2 Network Topology

Networks are mathematical objects that represent the real complex systems. In order to classify the network structures, here I tried to examine the topological properties and quantifiers at first and later gave some examples with model networks.

2.2.1 Topological Properties and Quantifiers

Other than the number of nodes N and of edges $N(e)$, one needs some features for both comparing and classifying the topologies of networks. Here, I listed some which were used throughout this thesis. For more detailed readings one can see the review articles [4, 8].

a- Degree Distribution, $P(k)$

The degree probability distribution is the probability distribution function for finding a node of network with degree k.

If the network is non-directed then there is only one degree notion, however, if the network is directed then one can define three different degree distributions. *Total-Degree Distribution, $P(k_{tot})$*: In this case the network is pretended to be non-directed, in other words, directions of the edges are removed, and the degree distribution is explored. *Out-Degree Distribution, $P(k_{out})$*: In this case, one counts only the edges *outgoing* from node and calculate their distribution. *In-Degree Distribution, $P(k_{in})$*: In this case, one counts only the edges *incoming* to node and calculate their distribution.

b- Degree-degree correlation, $k_{nn}(k)$

Degree-degree correlation function gives us the average degree of a node which our k-degree node connects [32, 13].

$$k_{nn}(k) = \sum_{k'} k' p(k|k') \tag{2.1}$$

where $p(k|k')$ is the conditional probability that the node with degree k is connected to a node with degree k'.

c- Clustering Coefficient, C_i

The clustering coefficient of a node is the fraction of the existing triangles including the node in quest to maximum possible number of triangles including this node [8]. Using the notation Δ_i for the number of triangles including v_i and knowing the fact that the maximum number of triangles is $\frac{k_i(k_i-1)}{2}$, one can state the clustering coefficient of v_i C_i as follows:

$$C_i = \frac{2\Delta_i}{k_i(k_i-1)} \tag{2.2}$$

One can define another quantity $C(k)$ which give us the average C_i of the nodes whose degree is k [32].

$$C(k) = \langle C_i \rangle_{d(v_i)=k} \tag{2.3}$$

d- Rich-Club Coefficient, $r(k)$

One can define $N_{>k}$ as the number of nodes whose degree is greater than k and $N(e_{>k})$ as the number of edges between those nodes. Then, rich-club coefficient $r(k)$ [11]:

$$r(k) = \frac{2N(e_{>k})}{N_{>k}(N_{>k}-1)}. \tag{2.4}$$

e- K-core

K-core was proposed by Bollobas [9] as a quantity that reflects hierarchical structuring in a network. Starting from $k = 0$, recursively each node whose degree is less than or equal to k is labeled as the member of this k-core and then pruned with its edges from the network. This procedure is repeated until no node whose degree is less than k remains.[2]

[2]For this topological feature one can use the online tool: http://xavier.informatics.indiana.edu/lanet-vi/, October 31st 2007

f- Motifs

"Motifs" have been recently proposed by Milo *et al.* [10] to capture simple subnetwork structures in directed networks. Some motifs may appear more frequently in network at hand.[3]

g- Shortest Path, sp

Another important feature in the topology of networks is *geodesic* or shortest path from one node to another. There might be not a single path from v_i to v_j in the network and in this case the smallest length path is called shortest path, sp_{ij} [8, 12]. One can define a probability function, $P(l)$, for finding a shortest path of length l. Similarly, one can also define sp_i, average shortest path of v_i to other nodes in the network.

It should be noted that if two nodes are at different clusters, then their shortest path are either taken as infinity, or ignored in the calculations as in this thesis. Networks with small shortest paths are in special attention in literature and are named as **Small World**[33] networks [8].

h- Betweenness, b_i

Betweenness b_i for v_i is defined as the total number of shortest paths passing through v_i [12]. Therefore, one can define a probability function $P(b)$ for finding a node with betweenness b. For finding b_i I have used Newman's algorithm given in Reference [12].

2.2.2 Topological Investigations on Some Model Networks

First of all, let me define what kind of model networks were used in this chapter. These network types were named according to their **indegree** distributions and this was stressed by inserting an **in-** as a prefix to the model name, such as in-XXX network. Although I did not use in this thesis, one should be aware of other sorts of model topologies, like random or gaussian.

in-NK Network: In this model, every node has exactly **K** incoming edges. In-NK model has been used and investigated widely in literature.

[3]For this topological tool one can use the free-software **mfinder** from Uri Alon's webpage http://www.weizmann.ac.il/mcb/UriAlon/

in-Power-Law (in-PL) Network: A network whose degree distribution is given by power-law, i.e. $P(k) \sim k^{-\alpha}$. It has been observed that many complex systems in nature have a PL behavior in their degree distributions[4]. Those PL networks found in nature come with taking the exponents α as shown in Figure 2.3 [17].

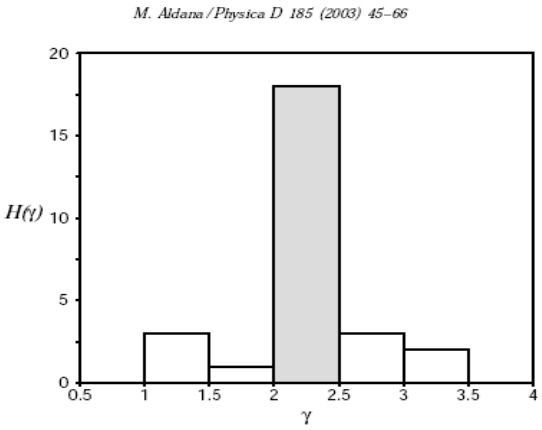

Figure 2.3: Histogram shows the distribution of Power-Law exponent of 30 networks found in nature. Taken from Reference [17]. γ in figure equals to α in my notation.

in-Exponential (in-EXP) Network: A network whose degree distribution is exponential, i.e $P(k) \sim e^{-\lambda k}$. Although to my knowledge many real-life networks are PL networks, some show an exponential behavior, in particular the gene regulation network of yeast shown in Figure 3.3.

a- in-NK, in-PL and in-EXP Network Topologies

In order to be able to compare different topologies, some of the properties needed to be fixed. For the sake of comparison to the network literature, I fixed the number of nodes to $N = 100$ and the average indegree to $\langle k_{in} \rangle \cong 2.00$ so that in-PL exponent $\alpha = 2.25$ and in-EXP exponent $\lambda = 0.7$ were chosen. Figure 2.4 shows and discuss the correspondence of exponents to average indegree of in-EXP and in-PL networks. Next, ensembles of 100 model networks of type in-NK, in-PL and in-EXP were created and investigated. The topological distributions of these ensembles are in Figure 2.5, Figure 2.6 for in-NK networks,

[4]In literature the term **scale-free network** is also used for power-law network

in Figure 2.7, Figure 2.8 for in-PL networks, in Figure 2.9, Figure 2.10 for in-EXP networks.

I found out that in general, the topological features of in-EXP networks are between in-NK and in-PL networks'.

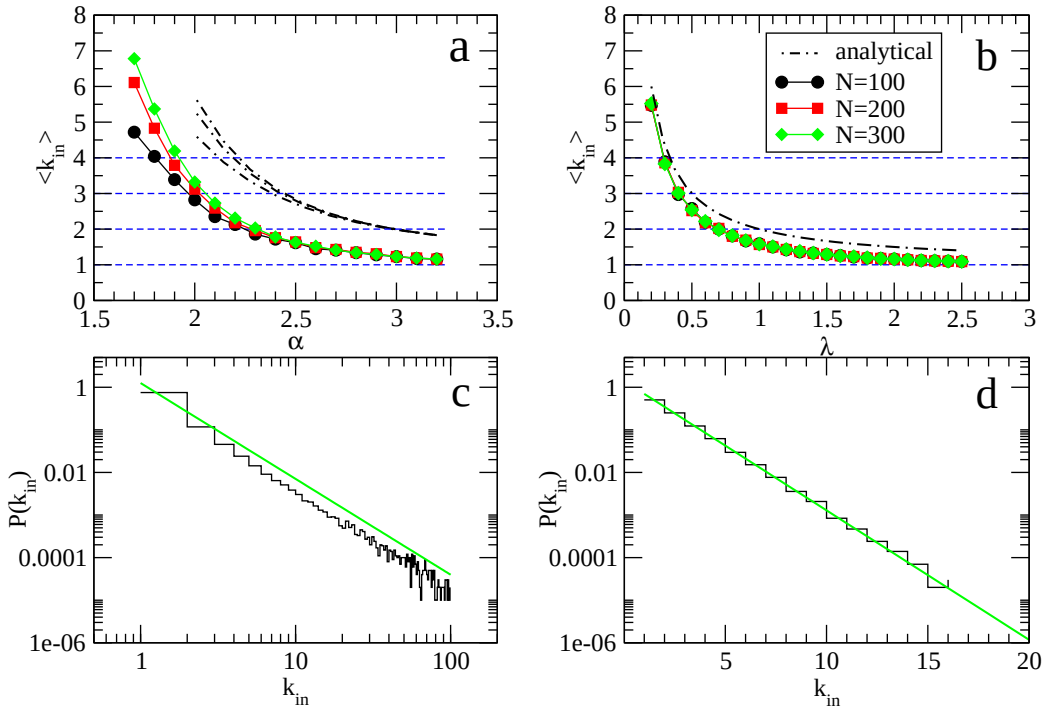

Figure 2.4: **a-)** shows the correspondence of PL exponent to $\langle k_{in} \rangle$ both for the *computational* and *analytical* for $N = 100, 200, 300$ (from bottom to top). It is seen that deviation is at an important level for comp. and anal. cases. Computational results are used in this thesis. **b-)** shows the correspondence of EXP exponent to $\langle k_{in} \rangle$ both for the *computational* and *analytical* for $N = 100, 200, 300$ (from bottom to top). It is seen that deviation has little effect comparing to PL case. Computational results are used in this thesis. **c-)** Indegree prob. distribution of computational and analytical cases for PL exponent $\alpha = 2.25$ $N = 100$; this study shows that the difference of computational and analytical cases comes from the contribution of $k_{in} = 1$ in the sum. **d-)** Indegree distribution of computational and continuous analytical cases for EXP exponent $\alpha = 0.7$ $N = 100$; comparing the PL case, the contribution of $k_{in} = 1$ to sum is small so it does not deviate from analytical case. See Appendix A for analytical $\langle k_{in} \rangle$ expressions.

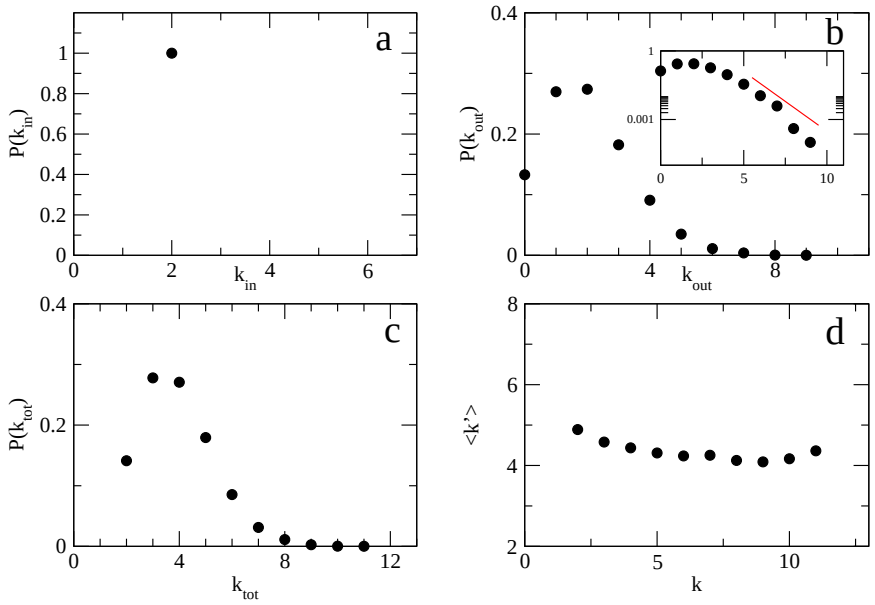

Figure 2.5: 100 in-NK ($N = 100$, $K = 2$) networks were created. Their topological features were investigated and averaged. Figure shows the corresponding; **a-)** indegree probability distribution, **b-)** outdegree probability distribution, **c-)** totaldegree probability distribution and **d-)** degree-degree correlation.

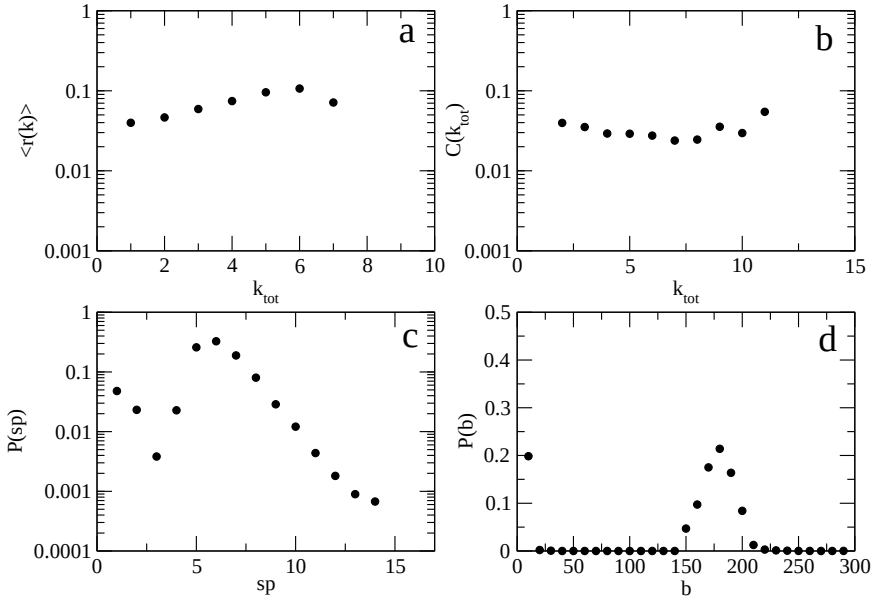

Figure 2.6: 100 in-NK ($N = 100$, $K = 2$) networks were created. Their topological features were investigated and averaged. Figure shows the corresponding; **a-)** rich-club coefficient, **b-)** clustering coefficient distribution, **c-)** average shortest path prob. distribution and **d-)** betweenness prob. distribution.

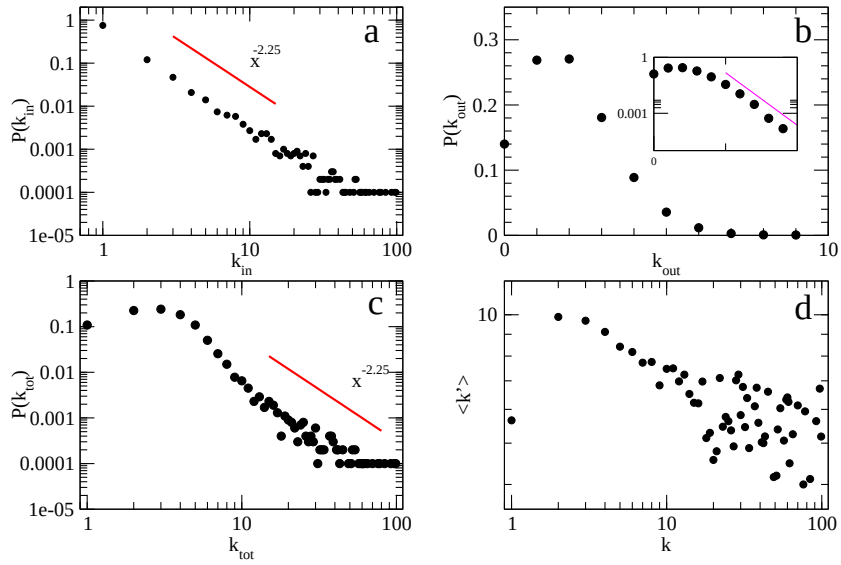

Figure 2.7: 100 in-PL ($N = 100$, $\alpha = 2.25$, $\langle k_i n \rangle = 2.0$) networks were created. Their topological features were investigated and averaged. Figure shows the corresponding; **a-**) indegree probability distribution, **b-)** outdegree probability distribution, **c-)** totaldegree probability distribution and **d-)** degree-degree correlation. It should be noted that $x^{-2.25}$ function was drawn in order to help the reader and is not a fitting.

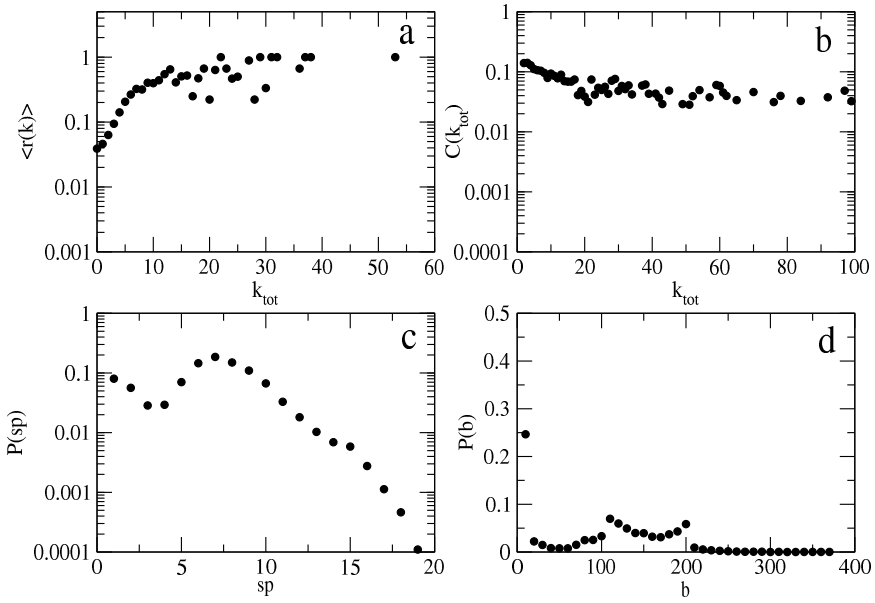

Figure 2.8: 100 in-PL ($N = 100$, $\alpha = 2.25$, $\langle k_i n \rangle = 2.0$) networks were created. Their topological features were investigated and averaged. Figure shows the corresponding; **b-**) clustering coefficient distribution, **c-)** average shortest path prob. distribution and **d-)** betweenness prob. distribution.

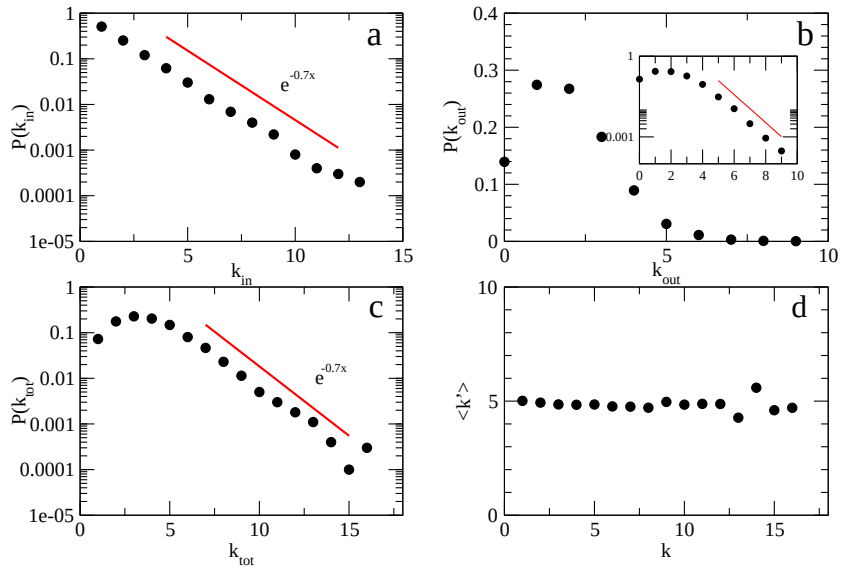

Figure 2.9: 100 in-EXP ($N = 100$, $\lambda = 0.7$, $\langle k_{in} \rangle = 2.0$) networks were created. Their topological features were investigated and averaged. Figure shows the corresponding; **a-)** indegree probability distribution, **b-)** outdegree probability distribution, **c-)** totaldegree probability distribution and **d-)** degree-degree correlation. It should be noted that $x^{-0.7}$ function was drawn in order to help the reader and is not a fitting.

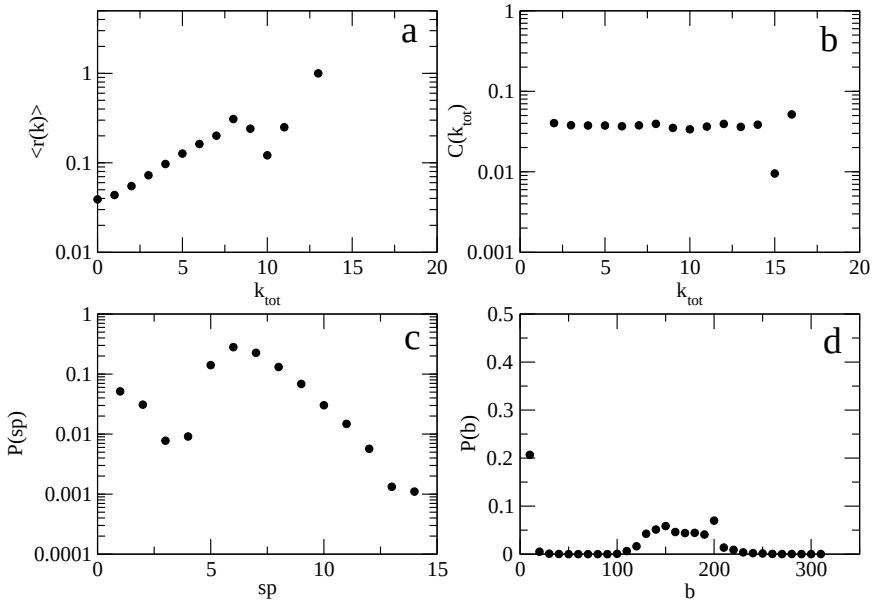

Figure 2.10: 100 in-EXP ($N = 100$, $\lambda = 0.7$, $\langle k_{in} \rangle = 2.0$) networks were created. Their topological features were investigated and averaged. Figure shows the corresponding; **a-)** rich-club coefficient, **b-)** clustering coefficient distribution, **c-)** average shortest path prob. distribution and **d-)** betweenness prob. distribution

2.3 Network Dynamics

For the investigation of dynamics in networks the **Boolean** model with *synchronously* and *deterministic update* was used. In this model, each node v_i has a **node state** $\sigma_i(t)$ at a particular time **t** where $\sigma_i(t)$ is either 1 (on) or 0 (off). The **network state** $S(t)$ is the set of individual node states: $S(t) = \{\sigma_1(t), \sigma_2(t), .., \sigma_N(t)\}$. $\sigma_i(t+1)$ is determined by the **Boolean Function** B_i of the node v_i. B_i depends on only to indegree neighbors of the node v_i. It should be noted that in the case of zero indegree, σ_i is fixed to either to 1 or 0 for every t.

$$\sigma_i(t+1) = \begin{cases} B_i(\sigma_{i,1}(t), \sigma_{i,2}(t), .., \sigma_{i,k_{in}}(t)), & k_{in} > 0 \\ 0 \; or \; 1 \; (fixed), & k_{in} = 0 \end{cases} \tag{2.5}$$

where $\sigma_{i,j}$ is the node state of jth in-neighbor of v_i and k_{in} is the indegree of v_i.

During the dynamics, the node states $\{\sigma_i(t)\}$ are collected at time step **t** and inserted into each Boolean Function synchronously. The output of B_i is assigned to the $\sigma_i(t+1)$ for time **t+1**.

$\sigma_{i,1}(t)$	$\sigma_{i,2}(t)$	...	$\sigma_{i,k_{in}}(t)$	$\sigma_i(t+1)$			
0	0	...	0	0	1	...	1
0	0	...	1	0	0	...	1
.	.	...	.	.	.	...	1
1	1	...	0	0	0	...	1
1	1	...	1	0	0	...	1

Table 2.1: An example of the ruletable expression for B_i where v_i has k_{in} indegree nodes.

Let me explain these processes by constructing a table which includes all possible combination of incoming nodes' states and assigns an output to them for a particular node state. Such a table is called **Ruletable** as shown in Table 2.1. If there is k_{in} incoming nodes, then there are $2^{k_{in}}$ input combinations, i.e $2^{k_{in}}$ rows in the ruletable. The output of each combination is either 1 or 0, which makes $2^{2^k_{in}}$ different ways to construct the output

column[5]. Before the dynamics starts, a ruletable out of $2^{2^{k_{in}}}$ possibilities is chosen for each node, so that dynamics runs deterministically. I used the term **network realization** for one network topology with its all assigned ruletables (functions) in this thesis.

2.3.1 Some Boolean Function Types

Although in many systems we may know the interacting pairs of individuals very well, we have usually little information about which rules govern the dynamics, as in gene regulation networks [23]. In other words, we do not know which combination out of $2^{2^{k_{in}}}$ to choose in the ruletable for each v_i. However, we are able to shape the structure of the Boolean Functions. I use 4 types of these *random* function structures found in literature.

a- Simple Random Function, ***RF***

A function type whose each input combination (each row in the ruletable) is assigned to an output value of 1 with a probability **p**.

b- Canalyzing Random Function, ***CF***

A *Canalyzing Random Function* has at least one canalyzing input variable, such that for at least one certain canalyzing value of that variable, the output value is fixed [23, 25].

$$B_i(\sigma_{i,1},..,\sigma_{i,j},..,\sigma_{i,k_{in}}) = \begin{cases} s_i & \sigma_{i,j} = s_j \\ B_i(\sigma_{i,1},..,\overline{s_j},..,\sigma_{i,k_{in}}) & \sigma_{i,j} \neq s_j \end{cases} \tag{2.6}$$

where jth in-neighbor is the canalyzing node with s_j as the canalyzing value and s_i as the canalazing output. As in RF only one parameter **p** is used whenever an output value is needed to be determined. It is should be mentioned that $B_i(\sigma_{i,1},..,\overline{s_j},..,\sigma_{i,k_{in}})$ in Exps. 2.6 is considered to be RF.

c- Nested Canalyzing Random Function, ***NCF***

Having investigated the Harris *et al.*'s work on gene regulation [22], Kauffman *et al.* have proposed a new function type known as *Nested Canalyzing* or *Hierarchically Canalyzing*

[5]For a better understanding of the magnitude, $2^{2^{k_{in}}} = 4, 16, 256, 65536, 4294967296$ for $K = 1, 2, 3, 4, 5$, respectively.

function which is argued to be found in the biological systems [23]. In this type, there is a canalazing order in input nodes and the output is determined by first node at its canalyzing value [23] is given:

$$B_i(\sigma_{i,1},..,\sigma_{i,j},..,\sigma_{i,k_{in}}) = \begin{cases} s_{i,1} & \sigma_{i,1} = s_1 \\ s_{i,2} & \sigma_{i,1} \neq s_1 \wedge \sigma_{i,2} = s_2 \\ ... & ... \\ s_{i,j} & \sigma_{i,1} \neq s_1 \wedge \sigma_{i,2} \neq s_2 \wedge ... \wedge \sigma_{i,j} = s_j \\ ... & ... \\ s_{i,k_{in}} & \sigma_{i,1} \neq s_1 \wedge \sigma_{i,2} \neq s_2 \wedge ... \wedge \sigma_{i,k_{in}} = s_{k_{in}} \\ \overline{s_{i,k_{in}}} & \sigma_{i,1} \neq s_1 \wedge \sigma_{i,2} \neq s_2 \wedge ... \wedge \sigma_{i,k_{in}} \neq s_{k_{in}} \end{cases} \tag{2.7}$$

where P($s_{i,j}$=TRUE)=P(s_j=TRUE)=$\frac{exp(-2^j\alpha)}{1+exp(-2^j\alpha)}$ j=1,2..,k_{in} and j is numbered with respect to the canalyzing order.

In this thesis, the definition of NCF is modified for the sake of consistency. Firstly, a **p** parameter as in above functions were adapted. Secondly, the last statement in Exps. 2.7 in determining the output was altered. Instead of using $\overline{s_{i,k_{in}}}$, the output value was determined again by using p.

d- Special Subclasses of Nested Canalyzing Random Function, ***SNCF***

After the proposition of Nested Canalyzing Functions (NCF) by Kauffman *et al.* [23], Nikolejewa *et al.* presented "a new minimal logical expression" for all NCFs [24] as follows,

$$\begin{aligned} \sigma_i &= B_i(\sigma_{i+1}, \sigma_{i+2}, ..., \sigma_{i+k_{in}-1}, \sigma_{i+k_{in}}) \\ &= \sigma_{i+1}^{\Theta} \bigodot (\sigma_{i+2}^{\Theta} \bigodot (... \bigodot (\sigma_{i+k_{in}-1}^{\Theta} \bigodot \sigma_{i+k_{in}}^{\Theta})...)) \end{aligned} \tag{2.8}$$

where $\bigodot$ represents either AND or OR logical function, i.e. $\bigodot \in \{\wedge, \vee\}$ and σ^{Θ} is for possible negation of σ, i.e. $\sigma^{\Theta} \in \{\sigma, \overline{\sigma}\}$

They classified the NCFs according to the possible chances for $\bigodot$. Upon investigation of Harris *et al.* data [22][6] they have found that gene regulatory rules are mainly governed by two subclasses of NCF [24]:

$$\sigma_{i+1}^{\Theta} \wedge (\sigma_{i+2}^{\Theta} \wedge (... \wedge (\sigma_{i+k_{in}-1}^{\Theta} \wedge \sigma_{i+k_{in}}^{\Theta})...)) \tag{2.9}$$

[6]Nikolejewa *et al.* have noted in their paper that they have taken the data from Harris by private communication

and

$$\sigma^{\Theta}_{i+1} \wedge (\sigma^{\Theta}_{i+2} \wedge (... \wedge (\sigma^{\Theta}_{i+k_{in}-1} \vee \sigma^{\Theta}_{i+k_{in}})...)) \tag{2.10}$$

with 66.39% and 29.41% probability of occurrence, respectively.

In this type of function p is not a free parameter and depends on the topology. It is easy to calculate p analytically for in-NK model topologies: $\mathbf{p}\approx\ (2/3) \times (1/2^K) + (1/3) \times (3/2^K) = 1.66 \times 2^{-K}$ [24]. For instance, for $K = 1, 2, 3, 4, 5, 6$, one finds $p = 0.83, 0.41, 0.21, 0.10, 0.05, 0.03, 0.01$, respectively. Figure 2.11 shows the validity of this formula for in-NK model where $K > 1$ and presents also p values for other topologies.

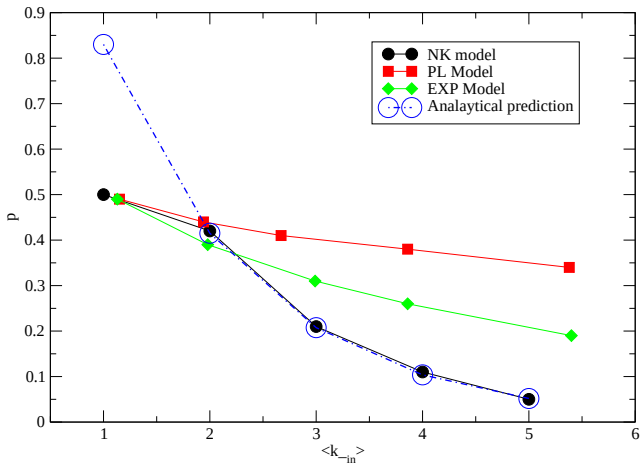

Figure 2.11: p value investigation of Special Subclasses of Nested Canalyzing Random Function for model networks. The analytical expression for in-NK networks is $1.66 \times 2^{-k_{in}}$ and it is seen that it predicts correctly for $k_{in} > 1$. For other types, p-values are obtained from this figure in this thesis.

2.3.2 Dynamical Properties and Quantifiers

In order to compare the dynamics of various networks, one needs quantitative measures. Here I provide quantities related to two notions: *Attractor* and *Robustness*.

a- Attractor

Remembering that each node state can be either 1 or 0, the size of state space is 2^N. Once the network realization (network topology and ruletables) are fixed, the dynamics is deterministic. In other words, if one chooses a network state S_i at time t in 2^N states, s/he

arrives at exactly *one* network state at next time step *t+1*. Also, since 2^N is a finite number, at most after traversing all the states, the dynamics starts to fall in a cycle (Figure 2.12). Such a cycle is called *attractor* and it is an important feature of the boolean dynamics.

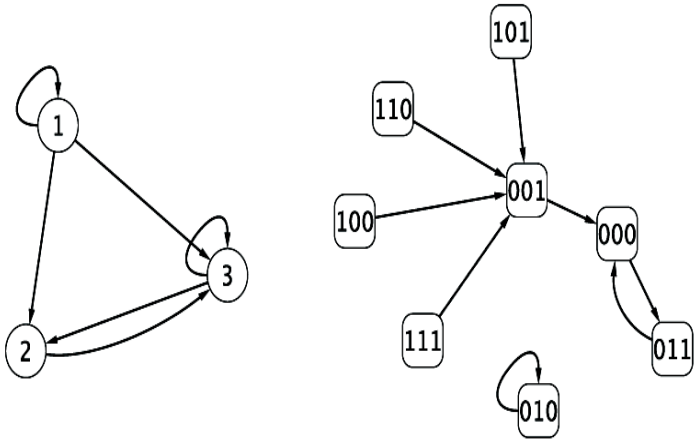

Figure 2.12: A simple network of 3 nodes is shown on the left. Its realization with ruletables shown in Table 2.2 produces two attractor as shown on the right.

$\sigma_1(t)$	$\sigma_1(t+1)$	$\sigma_1(t)$	$\sigma_3(t)$	$\sigma_2(t+1)$	$\sigma_1(t)$	$\sigma_2(t)$	$\sigma_3(t)$	$\sigma_3(t+1)$
0	0	0	0	1	0	0	0	1
1	0	0	1	0	0	0	1	0
		1	0	0	0	1	0	0
		1	1	0	0	1	1	0
					1	0	0	1
					1	0	1	1
					1	1	0	1
					1	1	1	1

Table 2.2: The ruletables of the nodes in the network shown in Figure 2.12.

It is believed that such attractors correspond to some process cycles in the systems, such

as phenotype in the cells [25, 19]. Mendoza *et al.* showed the correspondence between some attractors and known phenotypes in *Arabidopsis thaliana* by using a similar approach. They also predicted some mutant phenotypes and confirmed them by experiments [19]. Some other studies also reported similar conclusions [20, 21]

There are some quantifiers for the attractors in boolean systems. The first one is **the number of attractors** N_{attr} in the network realization. Second one is the number of network states of an attractor possesses, **length of the attractor** L_{attr}. Third one is the average number of network states to arrive at an attractor, **transient to attractor** τ_{attr}. The last quantifier is related to the notion of **the basin of attraction**. The set of network states which go to a particular attractor is called the basin of attraction of that attractor. The size of the basin of attraction normalized by 2^N is w_{attr}. Recently, Kravitz and Schumulevich [34] have proposed an entropy h definition for boolean dynamics:

$$h = -\sum_i w_i ln w_i \tag{2.11}$$

and h was used here to compare the basin of attractions of the network realizations.

Attractors were found in this thesis by using a heuristic algorithm (See Appendix B for more details about attractor finding algorithms.).

b- Robustness

For a system to be sustainable, its dynamics should not be effected drastically in every intensive or extensive changes, such as errors in individuals or environments. On the other hand, the dynamical systems like gene regulation should be open to some changes in order to survive through evolution. These arguments brings a hypothesis called "Life at the edge of chaos" that states the life systems should be at some where between chaotic and ordered phases [17, 25, 26].

So, how *robust* system is a valuable to detect for the dynamics and urges us to quantify *robustness*. It is presented here as follows [17]. Consider two network states $S(t)$ and $S^{'}(t)$. Their **Hamming Distance** $HD(t)$ is the number of nodes that are different in their states at time **t** [26, 17]:

$$HD(t) = \sum_{i=1}^{N} \mid \sigma_i(t) - \sigma_i^{'}(t) \mid \tag{2.12}$$

Let me define two other quantities which are the **overlapping functions** at time step **t** and **t+1**, respectively:

$$x(t) = 1 - \frac{HD(t)}{N} \tag{2.13}$$

$$M(x(t)) \equiv x(t+1) \tag{2.14}$$

The robustness under small perturbations is measured at the attractor of the system. In other words, we have

$$x \quad = \lim_{t\to\infty} x(t) \tag{2.15}$$

while $x \to 1^{-}$.

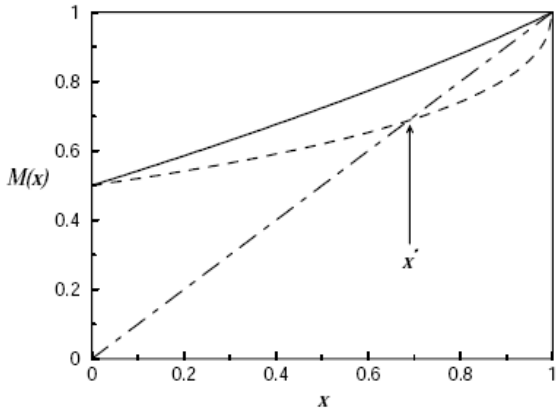

Figure 2.13: Robustness criteria [17]

Referring the Figure 2.13, we know $M(x) \geq 0$ when $x = 0$ and assume a monotic increase in $M(x)$ function. If $\lim_{x\to 1^-} \frac{dM(x)}{dx}$ is gerater than 1 then M(x) function crosses $M(x) = x$ line at some $x* < 1$, which is a stable fixed point other than $x = 1$. In this case, the system is forbidden to arrive at $x = 1$ unless $x = 1$. If $\lim_{x\to 1^-} \frac{dM(x)}{dx}$ is less than 1 then there is no stable fixed point other than $x = 1$ and two network states S and $S^{'}$ converge soon or later. These two cases are named as **chaotic**, **ordered** respectively and the case of $\lim_{x\to 1^-} \frac{dM(x)}{dx}$ equals 1 is named as **critical transition border/boundary** in the corresponding literature [17, 25, 26].

In sum, with showing the *robustness* quantity with s,

$$s = \lim_{x\to 1^-} \frac{dM(x)}{dx} \tag{2.16}$$

three important phase are summarized as follows,

$$\begin{aligned} s &< 1 \quad Ordered \\ s &= 1 \quad Critical\ Boundary \\ s &> 1 \quad Chaotic. \end{aligned} \tag{2.17}$$

and it is concluded that if $s < 1$, the system is robust against perturbations while if $s > 1$, the system is very sensitive to them [26].

2.3.3 Dynamically Relevant Subnetwork

Some of the nodes are irrelevant to the attractor results due to topology or functions of the network realization. These nodes only serve as computational challenges and some of the nodes with their edges can be recursively removed from the network without any general change in the dynamics [27]. Such nodes were labeled as **irrelevant** and the rest of the network/nodes were called the **dynamically relevant subnetwork/nodes** in this thesis.

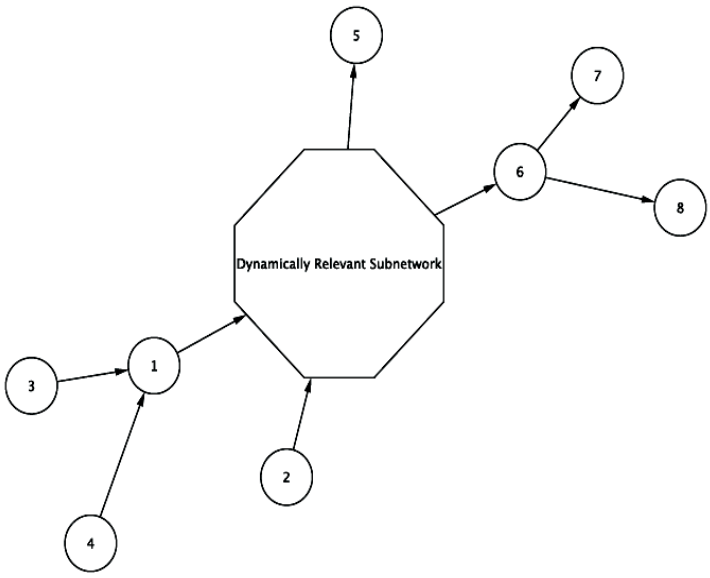

Figure 2.14: The subnetwork yielded after pruning recursively the nodes with either zero outdegree or zero indegree is named *dynamically relevant subnetwork* and this subnetwork was used in this thesis during the dynamics runs.

The attractor computations in this thesis were done with a minimal dynamically relevant subnetwork which is found by use of a procedure which considers only the topology of the

network to obtain it . The procedure depends on the fact that a node with zero indegree stays at a fix state all time steps. Also, a node with zero outdegree does not affect any node in the system although its state may fluctuate. Removing these two types of nodes with their edges recursively results in the dynamically relevant subnetworks used in this thesis.

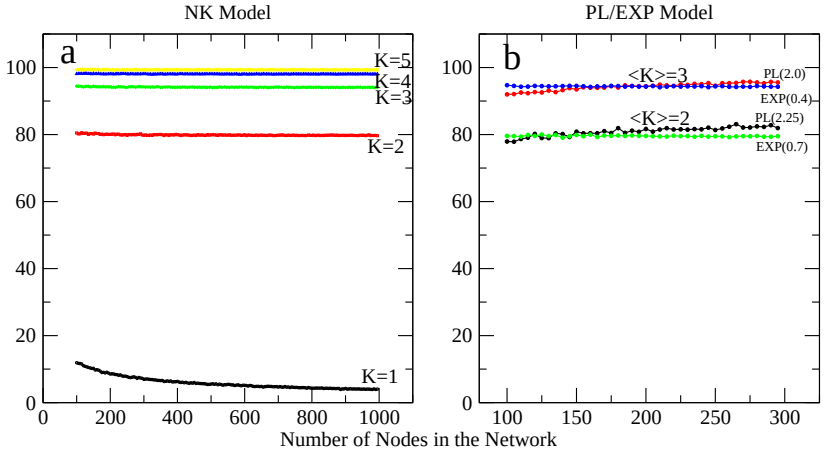

Figure 2.15: The fraction of dynamically relevant nodes to network size N at **a-)** in-NK, **b-)** in-PL and in-EXP networks. For in-NK, the values were found by averaging over $1000, 800, 500, 300, 200$ networks of $K = 1, 2, 3, 4, 5$, respectively. For in-PL and in-EXP, the values were found over $100, 500$ networks, respectively.

Figure 2.15 presents a computational study which investigate the percentage size of the dynamically relevant nodes in the in-NK, in-PL and in-EXP networks for different $\langle k_{in} \rangle$ and N. Main conclusion of this study is that the percentage of relevant nodes depends only on the $\langle k_{in} \rangle$, not on the network topology.

2.3.4 Dynamical Investigations on Some Model Networks

In order to investigate the dynamics of different model networks and the effect of their topologies on the dynamics, in-NK, in-PL and in-EXP network ensembles with $\langle k_{in} \rangle \cong 2.0$ were studied (Figure 2.4 shows that in-PL exponent $\alpha = 2.25$ and in-EXP exponent $\lambda = 0.7$ give $\langle k_{in} \rangle \cong 2.00$). The reason of choosing $\langle k_{in} \rangle \cong 2.0$ was its extensively use in related literature.

a- Attractors of in-NK, in-PL and in-EXP Model Networks

For distribution of the attractor features, N and p are fixed to 100 and 0.5, respectively. The reason behind using this p value was to compare the results to literature. However, since p was fixed, the SNCF whose p is not a free parameter was not used in this part. For each model type, I used 200 networks with 10 realization for each network in computations. Attractors were obtained after sampling 1000 initial conditions with the limits of maximum step and maximum length of an attractor as 1000 and 200, respectively.

The distribution of the number of attractors N_{attr}, the length of attractor L_{attr}, transient τ_{attr} and the entropy h_{attr} are shown in Figure 2.16, Figure 2.17, Figure 2.18 and Figure 2.19, respectively. Apart from the distributions, averages of these features are given in Table 2.3, Table 2.4, Table 2.5 and Table 2.6, respectively.

I found out that the probability distribution functions for N_{attr}, L_{attr}, τ_{attr} and h_{attr} in a network realization with in-NK, in-PL and in-EXP networks and RF, CF and NCF decay as a power-law function as stated for some topology and function types in References [29, 30]. Also, is was noted that both N_{attr} and L_{attr} shows a strange odd-even oscillations in the distributions which was also stated in Reference [30]. After some discussions, it was considered that these odd-evenness due to from artificial effects, for instance, the combinations of 2-, 3-, etc. node *partial* network states tends to create evenness. Furthermore, it should be noted that RF gives out considerably greater average values of those features than CF's and NCF's. While the average values were closer to each other with CF and NCF for all types of topologies, the averages with RF are higher than other for in-NK topology.

I also checked the scaling of the average values of the quantities above with number of nodes N for random, canalazing and nested canalazing functions with $p = 0.5$. N were chosen as 50, 55, 60, 66, 74, 82, 92, 100, 113, 124, 136, 149, 163, 179, 200, 215, 236, 259, 284, 300, 343, 377, 414, 455, 500, 550, 605, 665, 731, 804, 884, 972, 1000 in order to have a more accurate scaling behavior at small Ns but also to check big Ns. For $N \leq 100$, 200 networks were used for all function types. For $N > 100$, 100 networks for CF and NCF, and 50 networks for RF were used (RF runs slowly than others). The other parameters for dynamics were the same with $N = 100$ case above. The results for $\langle N_{attr} \rangle$, $\langle L_{attr} \rangle$, $\langle \tau_{attr} \rangle$ and $\langle h_{attr} \rangle$ scalings with N can be seen in Figure 2.20, Figure 2.21, Figure 2.22 and Figure 2.23, respectively.

$\langle N_{attr} \rangle$	in-NK	in-PL	in-EXP
RF	12.00 ∓ 24.06	5.28 ∓ 9.49	9.20 ∓ 21.95
CF	3.97 ∓ 7.92	3.11 ∓ 5.06	3.81 ∓ 6.49
NCF	4.59 ∓ 11.18	3.54 ∓ 17.17	2.86 ∓ 4.69

Table 2.3: The average number of attractors $\langle N_{attr} \rangle$ of model networks for random **RF**, canalyzing **CF**, nested canalyzing **NCF** functions. For 200 networks with $N = 100$ and 10 realisations for each network, the attractors were found by initiating from 1000 initials conditions with limits of 1000 maximum step size and 200 attractor length.

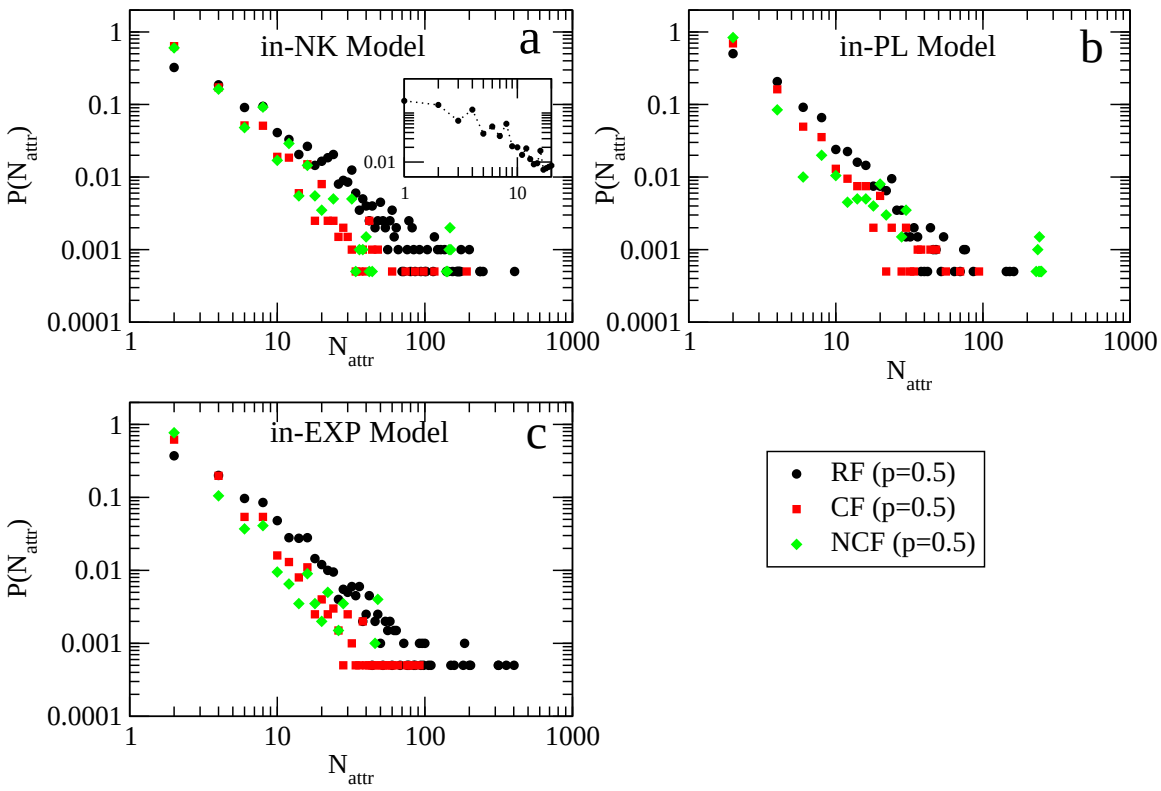

Figure 2.16: The **number of attractors** prob. distribution for model networks. **RF**: Random Function, **CF**: Canalazing Function, **NCF**: Nested Canalazing Function. For 200 networks with $N = 100$ and 10 realizations for each network, the attractors were found by initiating from 1000 initials conditions with limits of 1000 maximum step size and 200 attractor length. In order to uncover the artificial odd-even effect as shown in small frame, the data were binned in 2 units.

$\langle L_{attr} \rangle$	in-NK	in-PL	in-EXP
RF	11.69 ∓ 22.36	12.13 ∓ 71.60	20.15 ∓ 101.82
CF	3.05 ∓ 4.39	3.39 ∓ 23.48	3.77 ∓ 26.58
NCF	2.98 ∓ 3.83	2.02 ∓ 2.60	2.12 ∓ 1.92

Table 2.4: The average length of attractors of model networks. **RF**: Random Function, **CF**: Canalazing Function, **NCF**: Nested Canalazing Function. For 200 networks with $N = 100$ and 10 realizations for each network, the attractors were found by initiating from 1000 initials conditions with limits of 1000 maximum step size and 200 attractor length.

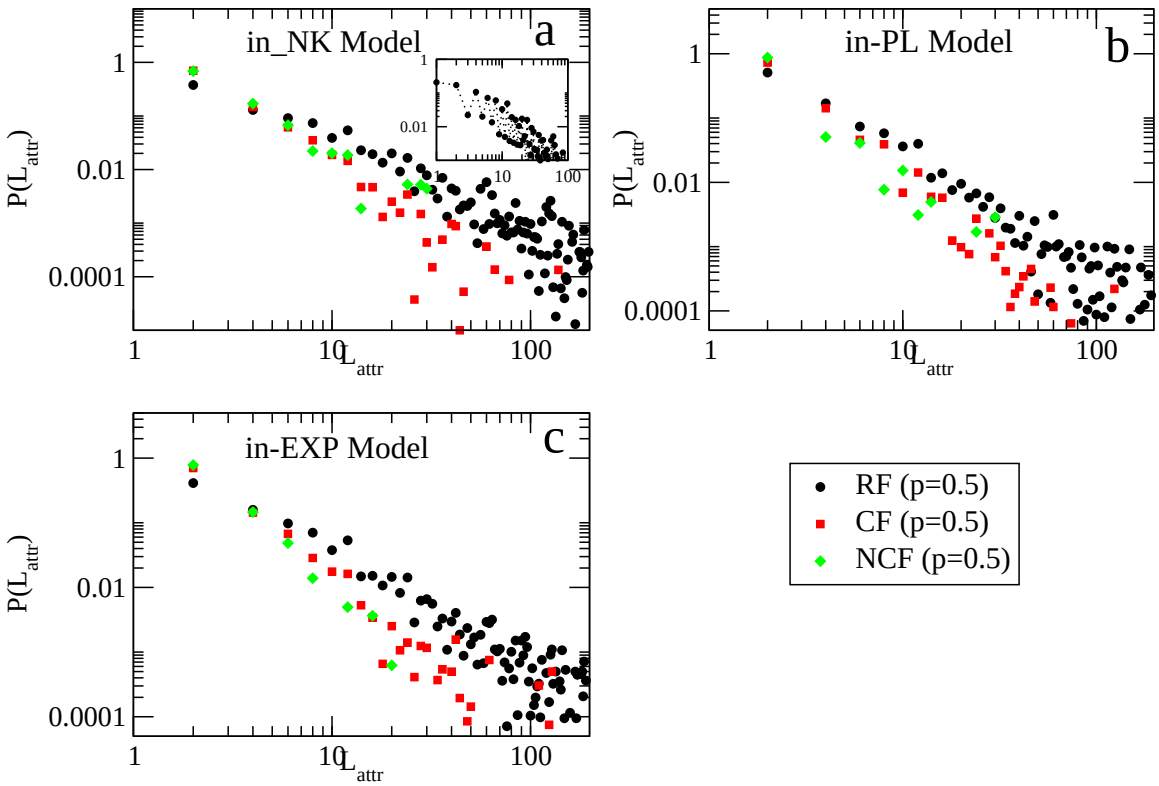

Figure 2.17: The **length of attractors** $\langle L_{attr} \rangle$ prob. distribution of model networks. **RF**: Random Function, **CF**: Canalyzing Function, **NCF**: Nested Canalyzing Function. For 200 networks with $N = 100$ and 10 realizations for each network, the attractors were found by initiating from 1000 initials conditions with limits of 1000 maximum step size and 200 attractor length. In order to uncover the artificial odd-even effect as shown in small frame, the data were binned in 2 units.

$\langle \tau_{attr} \rangle$	in-NK	in-PL	in-EXP
RF	58.44 ∓ 164.61	23.52 ∓ 74.84	35.87 ∓ 104.79
CF	10.55 ∓ 5.33	9.63 ∓ 23.86	10.64 ∓ 26.70
NCF	10.64 ∓ 4.68	7.29 ∓ 3.02	8.35 ∓ 2.91

Table 2.5: Average transient to attractors of model networks. **RF**: Random Function, **CF**: Canalyzing Function, **NCF**: Nested Canalyzing Function. For 200 networks with $N = 100$ and 10 realizations for each network, the attractors were found by initiating from 1000 initials conditions with limits of 1000 maximum step size and 200 attractor length.

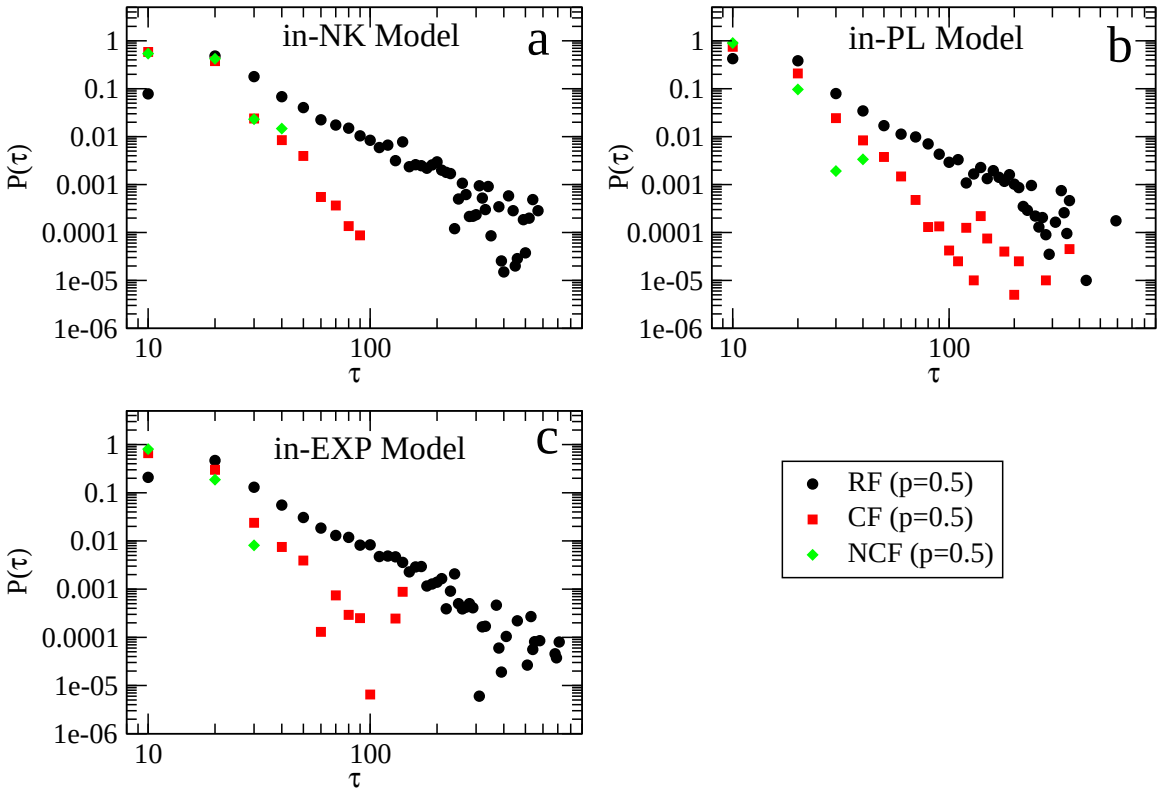

Figure 2.18: **Transient to Attractors** distribution of model networks. **RF**: Random Function, **CF**: Canalyzing Function, **NCF**: Nested Canalyzing Function. For 200 networks with $N = 100$ and 10 realizations for each network, the attractors were found by initiating from 1000 initials conditions with limits of 1000 maximum step size and 200 attractor length. The data were binned in 10 units in order to have clearer distribution.

$\langle h_{attr} \rangle$	in-NK	in-PL	in-EXP
RF	1.30 ∓ 1.10	0.88 ∓ 0.86	1.17 ∓ 1.00
CF	0.65 ∓ 0.78	0.58 ∓ 0.71	0.68 ∓ 0.75
NCF	0.71 ∓ 0.81	0.43 ∓ 0.72	0.48 ∓ 0.68

Table 2.6: Average values of the entropy of model networks. **RF**: Random Function, **CF**: Canalyzing Function, **NCF**: Nested Canalyzing Function. For 200 networks with $N = 100$ and 10 realizations for each network, the attractors were found by initiating from 1000 initials conditions with limits of 1000 maximum step size and 200 attractor length.

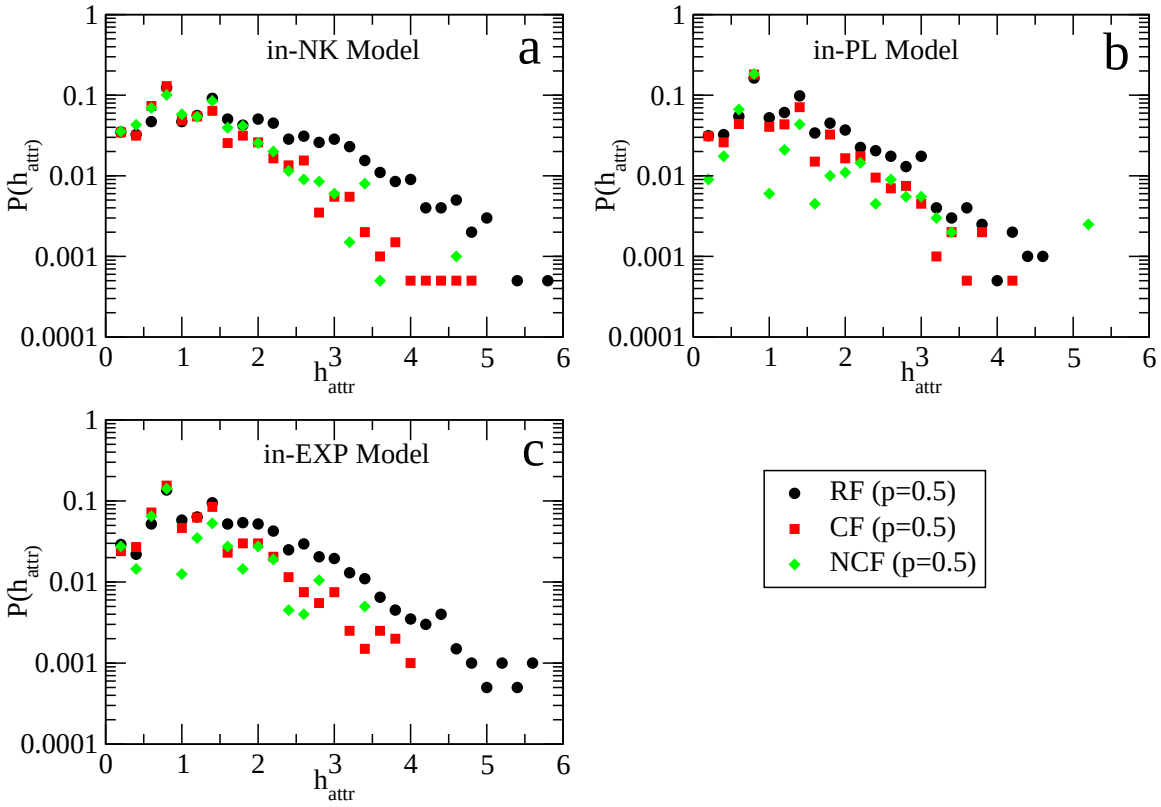

Figure 2.19: Entropy h_{attr} distribution of model networks. **RF**: Random Function, **CF**: Canalazing Function, **NCF**: Nested Canalazing Function. For 200 networks with $N = 100$ and 10 realizations for each network, the attractors were found by initiating from 1000 initials conditions with limits of 1000 maximum step size and 200 attractor length. The data were binned in 0.2 units in order to have clearer distribution.

I found out that for in-NK network with RF; $\langle N_{attr} \rangle$ scales with a fitting $N^{0.53}$, $\langle L_{attr} \rangle$ scales with a fitting $N^{0.87}$, $\langle \tau_{attr} \rangle$ scales with a fitting $N^{1.04}$. For a long time $\langle N_{attr} \rangle$ and $\langle L_{attr} \rangle$ scalings were considered as $\sqrt{N}$ [25] until Socolar & Kauffman published Reference [27] which states that $\langle N_{attr} \rangle$ scales with faster than linear. With this study I have shown that $\sqrt{N}$ scaling is valid for N_{attr} while fails for L_{attr}. Also, it should be noted that for CF and NCF, scalings are very small comparing to RF which might be considered as desirable for the biological systems.

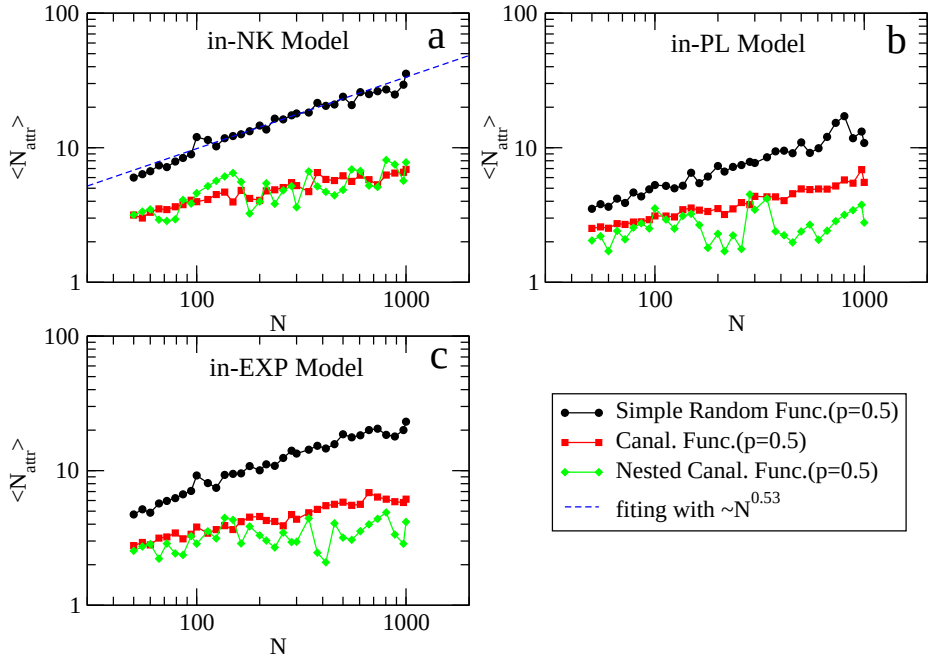

Figure 2.20: The scaling with N of the average number of attractors for RF,CF,NCF with $p = 0.5$.

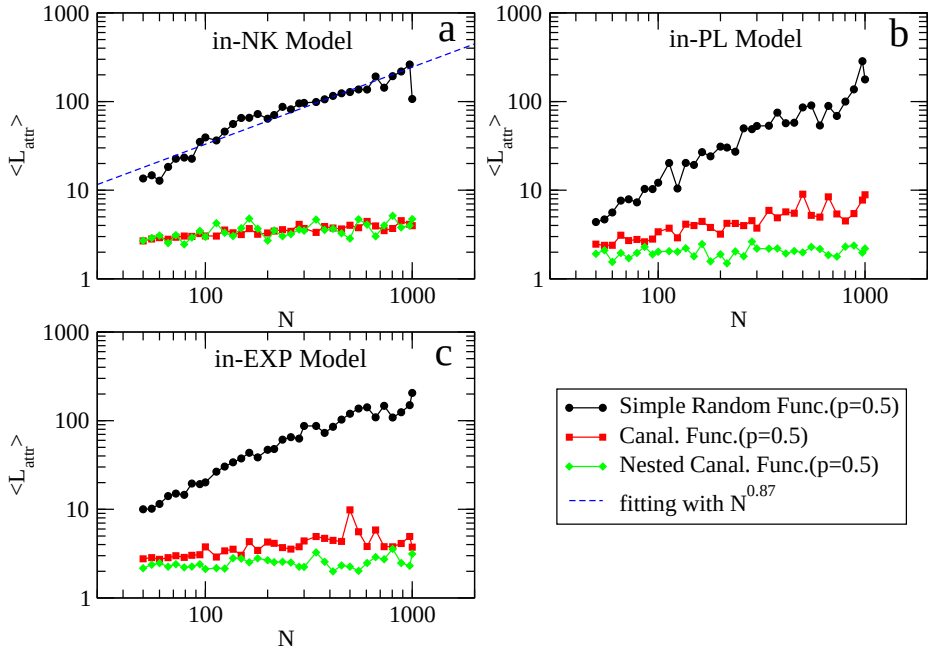

Figure 2.21: The scaling with N of the average length of attractors for RF,CF,NCF with $p = 0.5$.

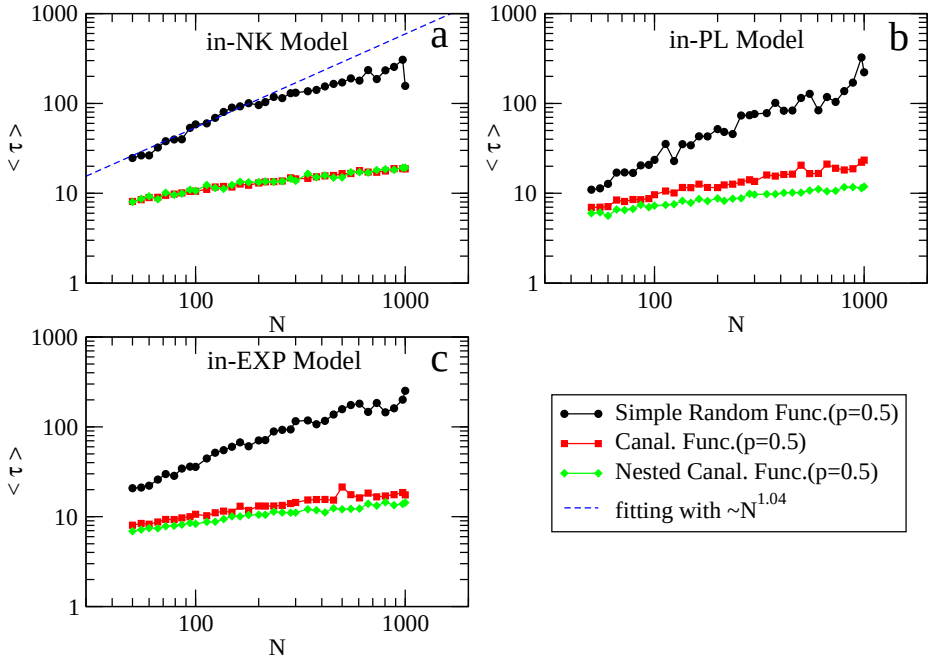

Figure 2.22: The scaling with N of transients to attractors for RF,CF,NCF with $p = 0.5$.

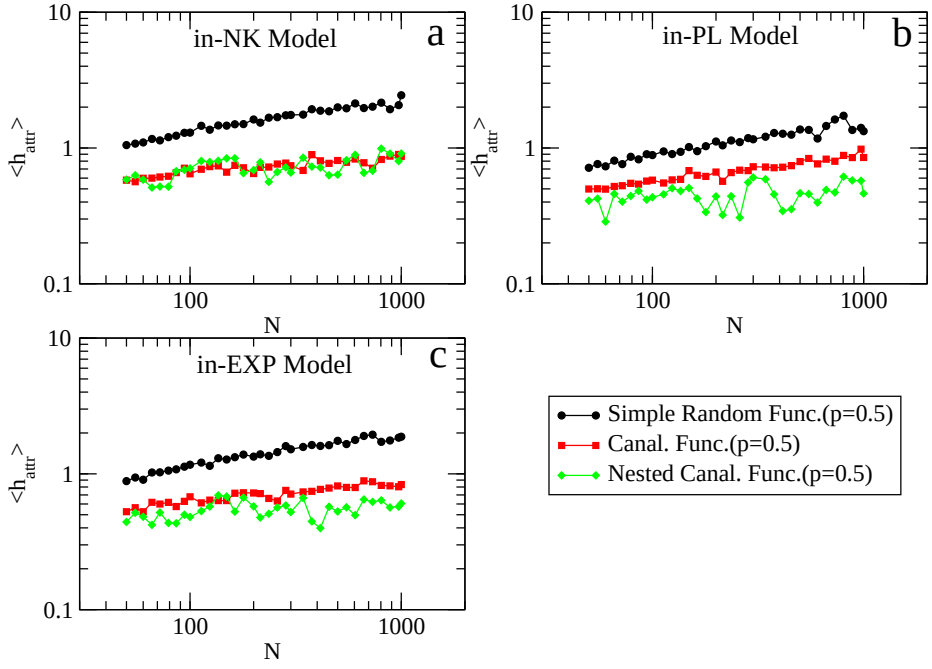

Figure 2.23: The scaling with N of the average entropy for RF,CF,NCF with $p = 0.5$.

b- Robustness of in-NK, in-PL and in-EXP Topologies

Numerical studies have yielded that *in-NK* model networks of $K = 2$ with random functions have privileged dynamical aspects than others, i.e. small, less number of attractors and at the boundary between chaotic and order phase [16]. Derrida & Pomeau were the first to give an analytical argument for why there is a such critical K value [28]. Later, Aldana generalized the argument for other type of topologies and gave more pedagogic expression [17].

Consider $x(t)$ which was discussed in *robustness* expression. One can define two sets $A(t)$ and $B(t)$ such that $A(t)$ is the set of nodes whose incoming edges come only from the nodes whose states are the same in the $S(t)$ and $S^{'}(t)$ and $B(t)$ is vice versa.

$$\begin{aligned} G &= \{v_1, v_2, ..v_N\} \\ A(t) &= \Big\{v_i, v_j : (v_j \rightarrow v_i) \bigwedge (\sigma_j(t) = \sigma'_j(t))\Big\} \\ B(t) &= G - A(t) \end{aligned} \tag{2.18}$$

Then we can express,

$$M(x(t)) = \sum_{k_{in}=1}^{\infty} P(k_{in})\{ \underbrace{\underbrace{[x(t)]^{k_{in}}}_{prob.\ of\ A(t)}}_{Contribution\ from\ A(t)} + \underbrace{\underbrace{(1-[x(t)]^{k_{in}})}_{prob.\ of\ B(t)} \times \underbrace{(p^2+(1-p)^2)}_{prob.\ the\ same\ output}}_{Contribution\ from\ B(t)} \}, \tag{2.19}$$

$$= \sum_{k_i=1}^{\infty} P(k_{in})\{-[x(t)]^{k_{in}}(2p^2-2p) + (2p^2-2p+1)\}. \tag{2.20}$$

If we take the derivative of the both sides with respect to $x(t)$ and use the fact that $\sum_{k_{in}=1}^{\infty} P(k_{in}) = 1$,

$$\frac{dM(x(t))}{dx(t)} = (2p^2-2p)\sum_{k_i=1}^{\infty} k_{in}[x(t)]^{k_{in}-1}P(k_{in}). \tag{2.21}$$

Remembering $\sum_{k_{in}=1}^{\infty} k_{in}P(k_{in}) = \langle k_{in}\rangle$ and Eq. 2.16, one finds;

$$\lim_{x(t)\rightarrow 1^-} \frac{dM(x(t))}{dx(t)} = \lim_{x(t)\rightarrow 1^-} \langle k_{in}\rangle x(t)^{\langle k_{in}\rangle - 1}(2p^2-2p) \tag{2.22}$$

$$= 2p(p-1)\langle k_{in}\rangle. \tag{2.23}$$

I tried to examine the validity of Formula 2.23 for *in-NK*, *in-PL* and *in-EXP* networks with the *simple random function* (RF) by comparing the analytical and computational results. Especially, the critical chaotic-ordered boundary, $s = 1$, was checked. The

robustness computations were done for different sets of $p \in \{0.0, 0.01, 0.02, ..., 0.49, 0.50\}$ and corresponding parameters for $\langle k_{in} \rangle$: for in-NK, $K \in \{1, 2, ..., 5, 6\}$; for in-PL, $\alpha \in \{1.60, 1.65, 1.70, ..., 2.45, 2.5\}$ and for in-EXP, $\lambda \in \{0.30, 0.35, ..., 0.95, 1.0\}$. For the analytical expression (Eq.2.23), the relations between $\langle k_{in} \rangle$ with α, λ were yielded by consulting the Figure 2.4. The results can be seen in Figure 2.24, in Figure 2.25 and in Figure 2.26 for in-NK, in-PL and in-EXP networks, respectively.

I found out that robustness values, especially the critical boundary, match with the analytical values (Formula 2.23) for in-NK Model. The analytical expression also predicted the robustness values of in-PL and in-EXP networks but not as in-NK networks' case. I saw that matching for these networks gets better while N is increased which concluded as finite-size effect. In short, Expression 2.23 is successful for predicting the robustness of the networks for simple random functions.

I also checked the variation of the robustness for all types of functions (RF, CF, NCF, SNCF). Again, each network topology was set to $\langle k_{in} \cong 2.0 \rangle$ and $N = 100$. Each *robustness* value for corresponding $\langle k_{in} \rangle$ parameter and p was yielded by using the following parameters: 100 random initials conditions for each 10 network realizations for each 10 networks. As it can be seen in Figure 2.27, canalazing and nested canalazing functions resulted in more ordered robustness values than simple random and special subclasses of nested canalazing functions for all types of networks.

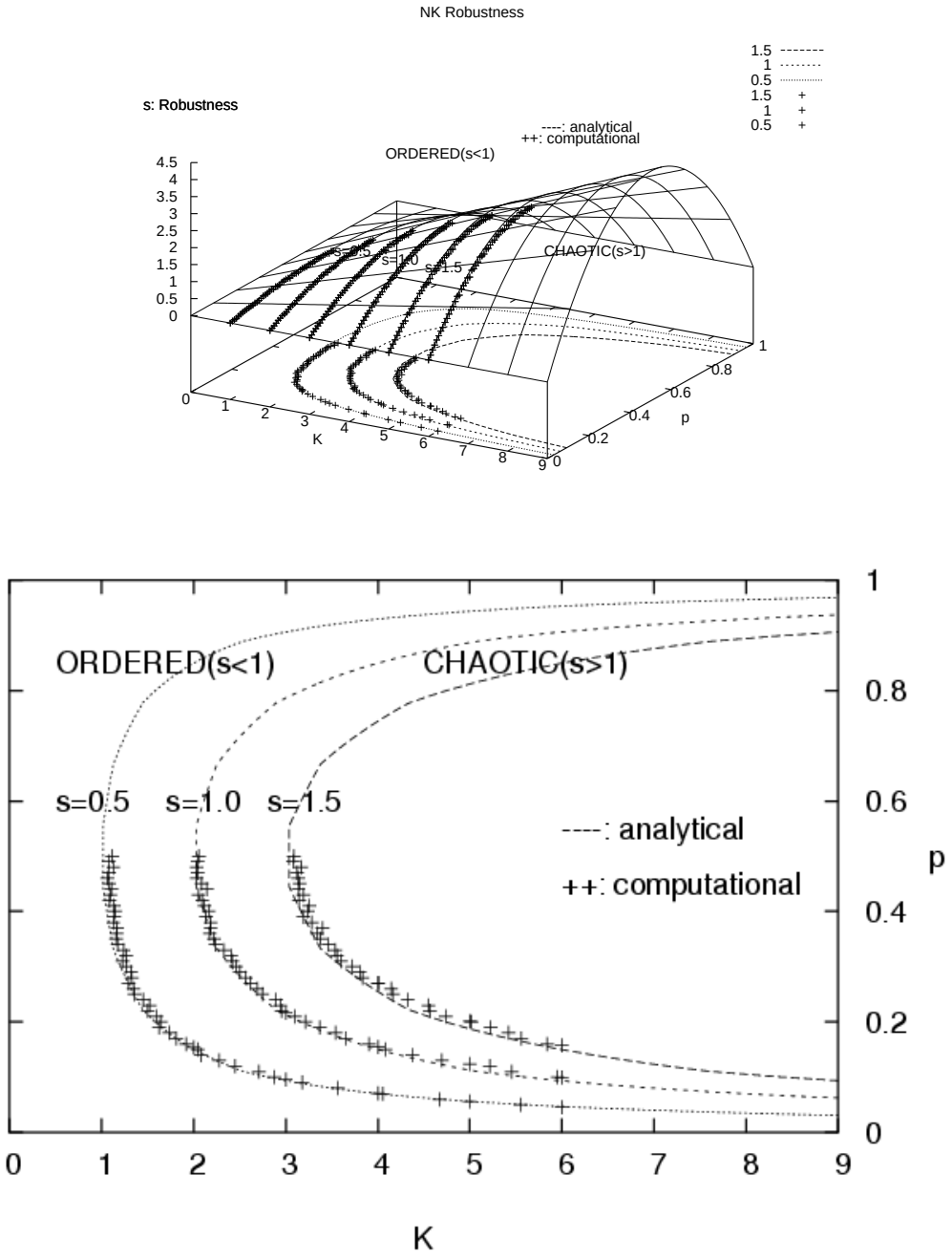

Figure 2.24: Analytical$(---)$ and computational$(+++)$ the robustness values were compared for in-NK model with $N = 100$ for each $K \in \{1, 2, .., 6\}$ and $p \in \{0.0, 0.01, .., 0.49, 0.50\}$ for simple random functions. For each K, p, 10 networks and for each network 10 realization were constructed and the robustness was calculated starting from 100 initials conditions of each realization. Here, s=1 corresponds the critical border for transition from ordered to chaotic regimes. It seems that Derrida's results matches the s=1 border.

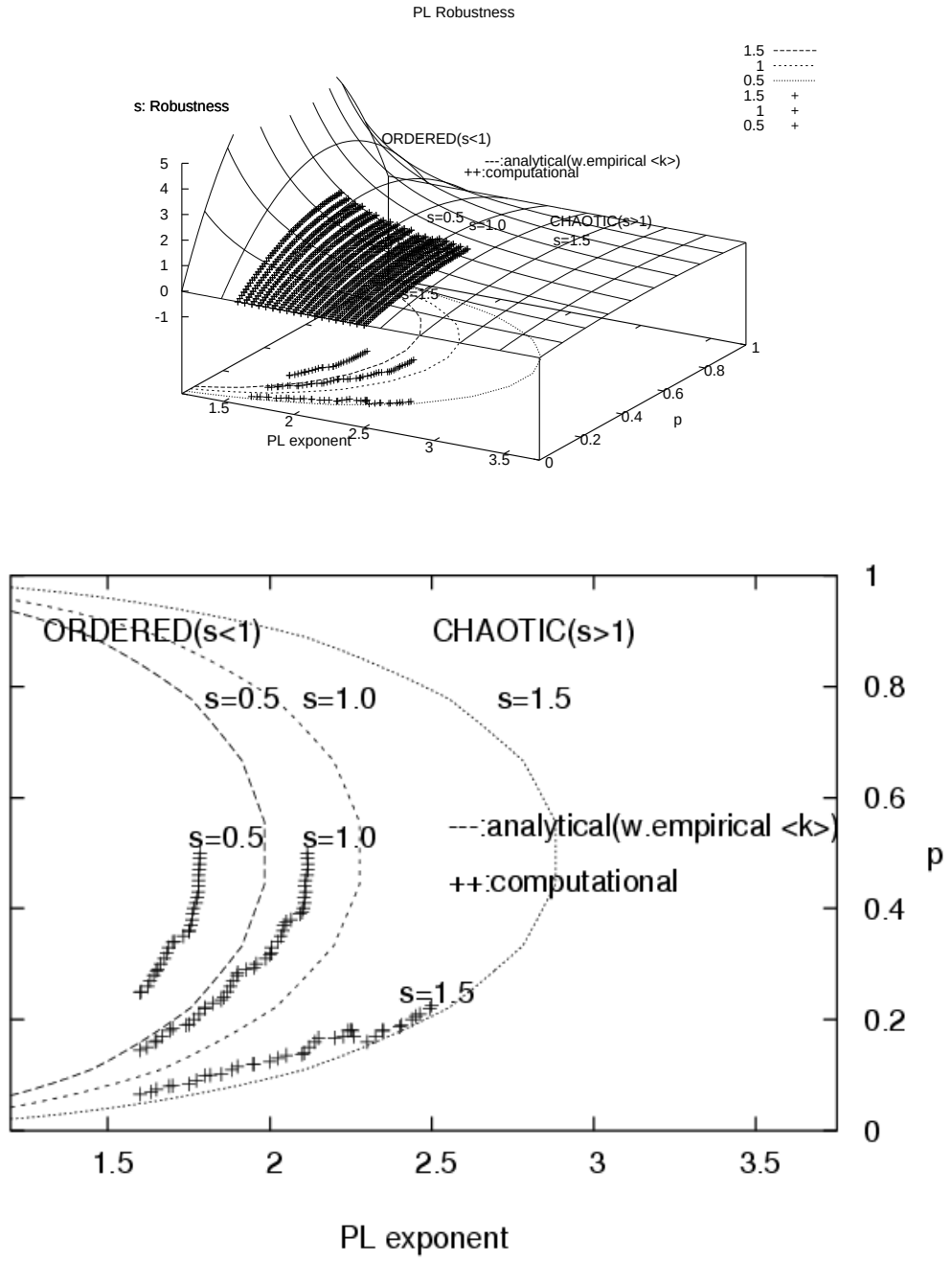

Figure 2.25: As in previous figure, the computational and analytical robustness was compared for in-PL networks with $N = 100, \alpha = 1.7, 1.75, .., 2.5$ and $p \in \{0.0, 0.01, .., 0.49, 0.50\}$ for simple random functions. For each α, p, 10 networks and for each network 10 realization were constructed and the robustness was calculated starting from 100 initials conditions of each realization. The computational and analytical results are close to each other, it was seen with bigger N values, the matching got closer which concludes a finite-size effect.

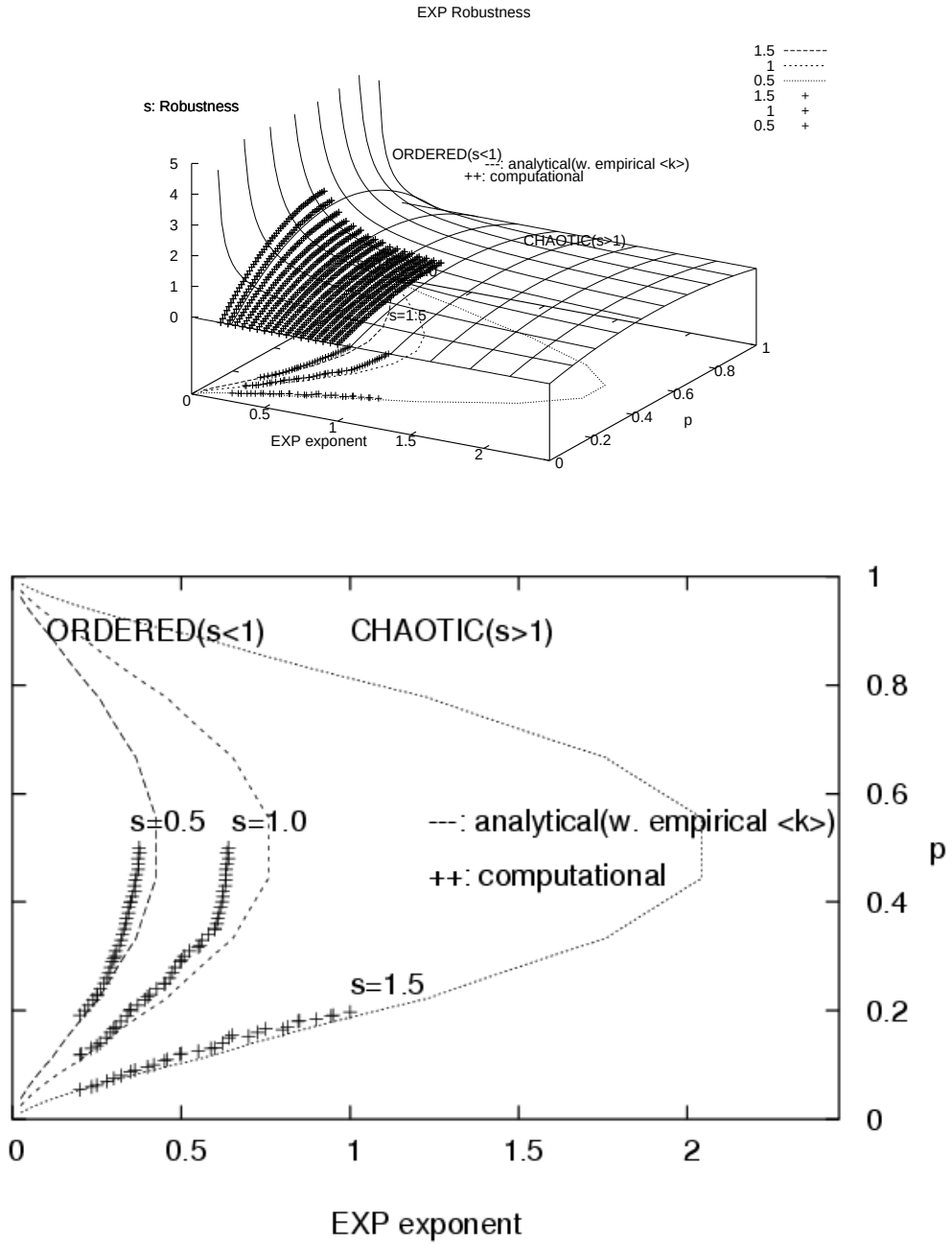

Figure 2.26: As in previous figure, the computational and analytical robustness was compared for in-EXP networks with $N = 100, \lambda = 0.3, 0.35, .., 1.0$ and $p \in \{0.0, 0.01, .., 0.49, 0.50\}$ for simple random functions. For each λ, p, 10 networks and for each network 10 realization were constructed and the robustness was calculated starting from 100 initials conditions of each realization. As in previous case, it was seen with bigger N values, the matching got closer which concludes a finite-size effect.

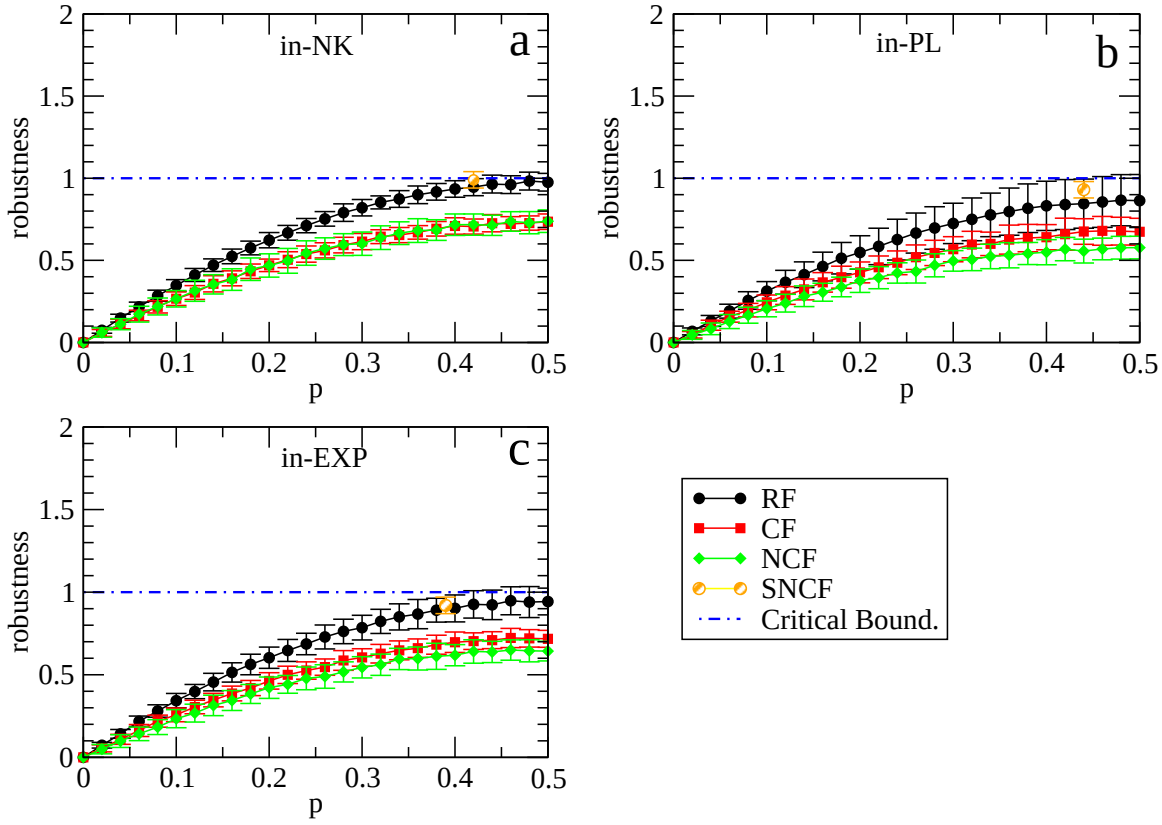

Figure 2.27: Figure shows the robustness values of the **a-)**in-NK, **b-)**in-PL and **c-)**in-EXP networks of $\langle k_{in} \rangle \cong 2.0$ and $N = 100$, $p \in \{0.0, 0.02, .., 0.50\}$ for RF: Simple Random, CF: Canalazing, NCF: Nested Canalazing and SNCF: Special Subclasses Nested Canalazing Functions. SNCF robustness results and corresponding p values for in-NK, in-PL and in-EXP are 0.99 ∓ 0.05, 0.93 ∓ 0.05, 0.92 ∓ 0.05 and 0.42, 0.44, 0.39, respectively.

Chapter 3

GENE REGULATION

This chapter is organized as follows: Section 3.1 introduces the gene regulation concept in biology. Section 3.2 gives the yeast gene regulation network (GRN) with the topological and dynamical investigations. Section 3.3 compares the yeast GRN with model networks whose indegree probability distribution is exponential. Section 3.4 emphasizes a recently proposed model which produce networks that are topologically similar to yeast GRN and discusses this model dynamically.

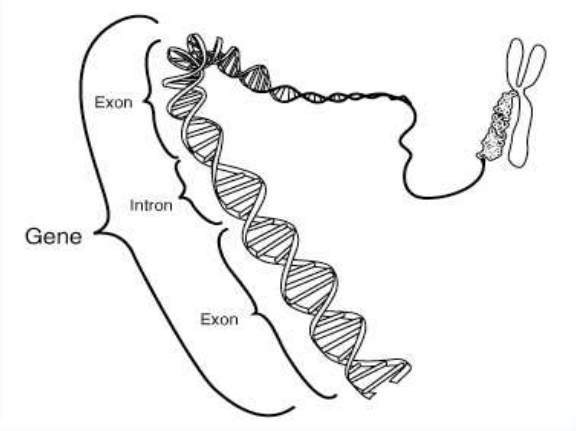

Figure 3.1: The DNA is considered to fulfill three main function in the life systems: 1-) *storage*, 2-) *heritage* and 3-) **expression** of the genetic information in the cells [2]

3.1 Introduction

To my knowledge, the most recent scientific definition of "gene" (Figure 3.1[1]) was proposed by Gerstein *et al.*: "A **gene** is a union of genomic sequences encoding a coherent set of potentially overlapping functional products." [35]. Genes in prokaryotes are always active and express their coded information into functional elements unless they are repressed by

[1]Taken from http://en.wikipedia.org/wiki/Gene

outside factors. However, at a particular time genes of eukaryotes (in Reference [15] it is stated as 2-15%) are generally inactive and need to be activated, therefore, one can talk about *regulation* of gene expression in eukaryotes [15].

There are different types of regulation of gene expression data depending of how it is detected. In this study, I used the **transcriptional** GR since the data used at this thesis was supplied by monitoring the levels of transcription materials, i.e. mRNA. For the details of experiments to yield the data, see References [6, 3, 36].

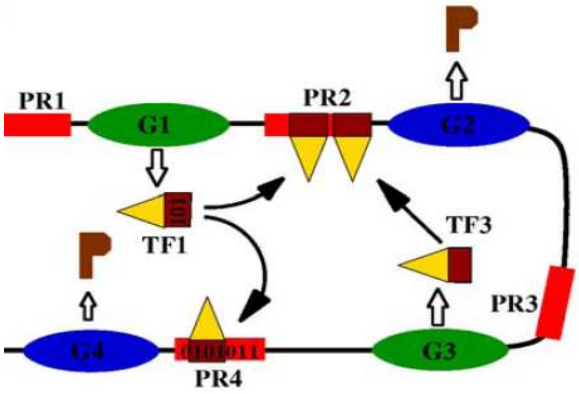

Figure 3.2: A brief explanation of the gene regulation process used in the thesis [32].

The transcriptional regulation of gene expression, or in short, the gene regulation is modeled as follows: Each gene on the DNA possesses two main regions. The first is the **promoter region (PR)** and the second is genetic codes that are transcribed. In order to *activate/inhibit* the transcription, the necessary conditions need to be hold at the PRs. Although these conditions are more complex, they are simplified as *existence/absence* of **transcription factors (TFs)** that are unique proteins bind the PRs. When the conditions hold, the gene is activated/inhibited for transcription depending on the rules of this specific gene. The products of these processes could be either the functional proteins or TFs (See Figure 3.2 for a simple scatching of the GR process [32]). More details about gene regulations can be found in References [15, 3].

My main aim in this part of the thesis is to both topologically and dynamically investigate the yeast gene regulation by using the network tools introduced in Chapter 2.

3.2 The Example at Hand: Saccharomyces Cerevisiae (Yeast)

Saccharomyces Cerevisiae or the yeast is a unicellular eukaryotic microorganism. Its gene regulation data [6] was used in this thesis since it is one of the mostly studied organism. The data was retrieved from YEASTRACT[2] database [37]. When the data was taken it was including 4252 genes (146 of them are TFs) with 12541 interactions.

3.2.1 Topology of the Yeast Gene Regulation Network

My investigations for some conventional topological features of the yeast GRN can be seen in Figure 3.3 and Figure 3.4. Previously, Guelzim *et al.* have topologically investigated the yeast GRN and stated an exponential decay in the indegree distribution with an exponent $\lambda = 0.45$ [13]. However, my indegree distribution was fitted to an exponential decay with an exponent $\lambda = 0.38 \mp 0.01$ by using GNUPLOT[3] with ignoring $k_{in} = 0$. The reason of this difference might be due to that my fitting was done by at first taking log of y-values and then fitting to a linear function. I also did a direct fitting resulting in $\lambda = 0.46 \mp 0.01$ exponential decay which is is almost the same as Guelzim *et al.*'s. Another detailed studies for topological examination of yeast can be found in References [32, 5].

The yeast GR network includes 4252 nodes/genes with 12541 directed edges/interactions (average degree is 2.95). 146 of those genes are TFs and there are 403 interactions between TFs (average degree is 2.76). Dynamically relevant subnetwork of yeast GRN consists of 82 TFs and 254 interactions (average degree is 3.10).

Comparing the artificial networks, the fraction of dynamically relevant nodes to the number of nodes in yeast GRN is very low ($82/4252 \cong 2\%$). As shown in Figure 2.14, the artificial model networks with the same indegree distributions show a fraction of $85 - 90\%$. Figures 3.3 and 3.4 also show the topological features of dynamically relevant subnetwork of the yeast GRN and one can state the topologies are similar. This specialties of yeast GRN might be crucial in its dynamics in real case.

[2] www.yeastract.com

[3] http://www.gnuplot.info/

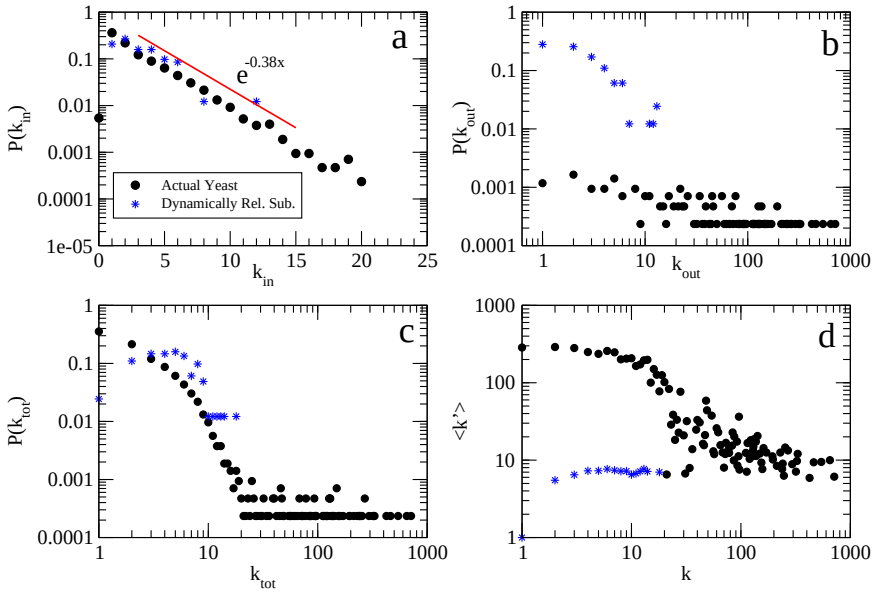

Figure 3.3: Yeast GR actual and dynamically relevant sub- network's **a-)** indegree probability distribution **b-)** outdegree probability distribution, **c-)** total degree probability distribution and **d-)** degree-degree correlation

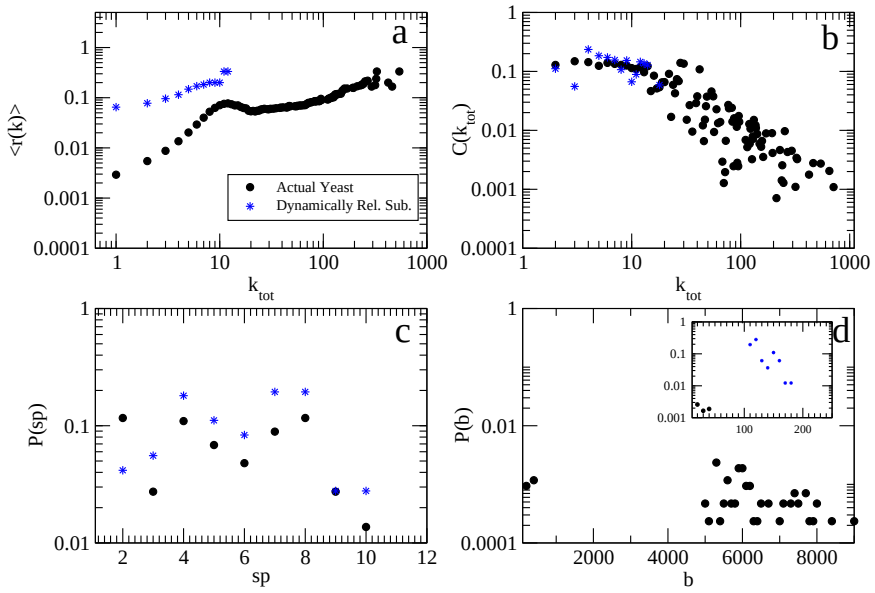

Figure 3.4: Yeast GR actual and dynamically relevant sub- network's **a-)** richclub coefficient **b-)** clustering coefficient, **c-)** shorthest path probability distribution (Binned in 1.0 units) and **d-)** betweenness probability distribution (for actual network it is shown in big frame with binned in 100 units, for dynamically relevant subnetwork it is shown in small frame with binned in 10 units.).

3.2.2 Dynamics of the Yeast Gene Regulation Network

Although the interacting pairs are known very well, the rules governs the interactions in the yeast gene regulation are not identified in detail yet. For this reason, I used the 4 types of random functions introduced in Chapter 2; simple random function (RF), canalazing random function (CF), nested canalazing random function (NCF) and special subclasses of nested canalazing random function (SNCF) with a statistical approach for investigations.

In order to investigate and compare the attractors of Yeast GRN for all function types, one needs to fix the p value. SNCF for dynamically relevant subnetwork of actual Yeast GRN gives out an effective p-value $p = 0.27 \mp 0.05$, therefore, I used this p value as the parameter for other type of functions, too. 2000 network realization were done for investigation. For each realization, attractors were explored by starting from 1000 random initial conditions. In the dynamics, the maximum 1000 steps and the maximum attractor length 200 limits were set. The distributions and averages of the number of attractors N_{attr}, the length of attractor L_{attr}, transient τ_{attr} and the entropy h_{attr} can be seen in Figure 3.5 and Table 3.1.

First of all, it should be noted that N_{attr} and h_{attr} distribution are not decreasing for all functions, especially SNCF type shows a not ordinary profile. This type of the distribution was not observed while studying with model networks in Chapter 2 and should be discussed further. Secondly, the averages with SNCF type shows a significant difference with other types. As it is seen in Table 3.1, it has a big $\langle N_{attr} \rangle$ while having small $\langle L_{attr} \rangle$ and $\langle \tau_{attr} \rangle$ which might be desirable in biological systems [25]. As a conclusion to the attractor results, SNCF type seems to be very appropriate for the dynamics of the yeast gene regulation and should be continued to be investigated.

I also obtained the *robustness* for each p $\in \{0.00, 0.01, .., 0.50\}$ of RF,CF and NCF, and for $p = 0.27$ of SNCF. 10 dynamics realization and for each realization 1000 random initial conditions were created and as explained in chapter 2, dynamics were runned for $10 \times N$ steps ($N = 82$ in this case) in order to be able to achieve an attractor. After these steps, *robustness* was measured numerically and averaged for all values. The results can be seen in Figure 3.6.

It is shown that for $p = 0.27$ (attractor statistics was done at this p-value) RF functions at the chaotic side near critical boundary while others are at the ordered side. Also, it seems that the Derrida's Exp.2.23 predicts RF results quite well. The results also shows

	$\langle N_{attr} \rangle$	$\langle L_{attr} \rangle$	$\langle \tau_{attr} \rangle$	$\langle h_{attr} \rangle$
RF	430.78 ∓ 263.78	8.82 ∓ 13.45	18.33 ∓ 15.86	5.40 ∓ 1.10
CF	222.77 ∓ 201.04	3.77 ∓ 3.12	8.48 ∓ 3.99	4.48 ∓ 1.20
NCF	221.15 ∓ 209.14	2.84 ∓ 1.93	6.68 ∓ 2.36	4.45 ∓ 1.27
SNCF	538.72 ∓ 212.18	3.40 ∓ 2.77	7.84 ∓ 3.50	5.96 ∓ 0.61

Table 3.1: Average values of the number of attractors N_{attr}, the length of attractor L_{attr}, transient τ_{attr} and the entropy h_{attr} of Yeast GRN for **RF**: Random Function, **CF**: Canalyzing Function, **NCF**: Nested Canalyzing Function, **SNCF**: Special Subclasses of Nested Canalyzing Function. For the details of the study, see the caption of Figure 3.9.

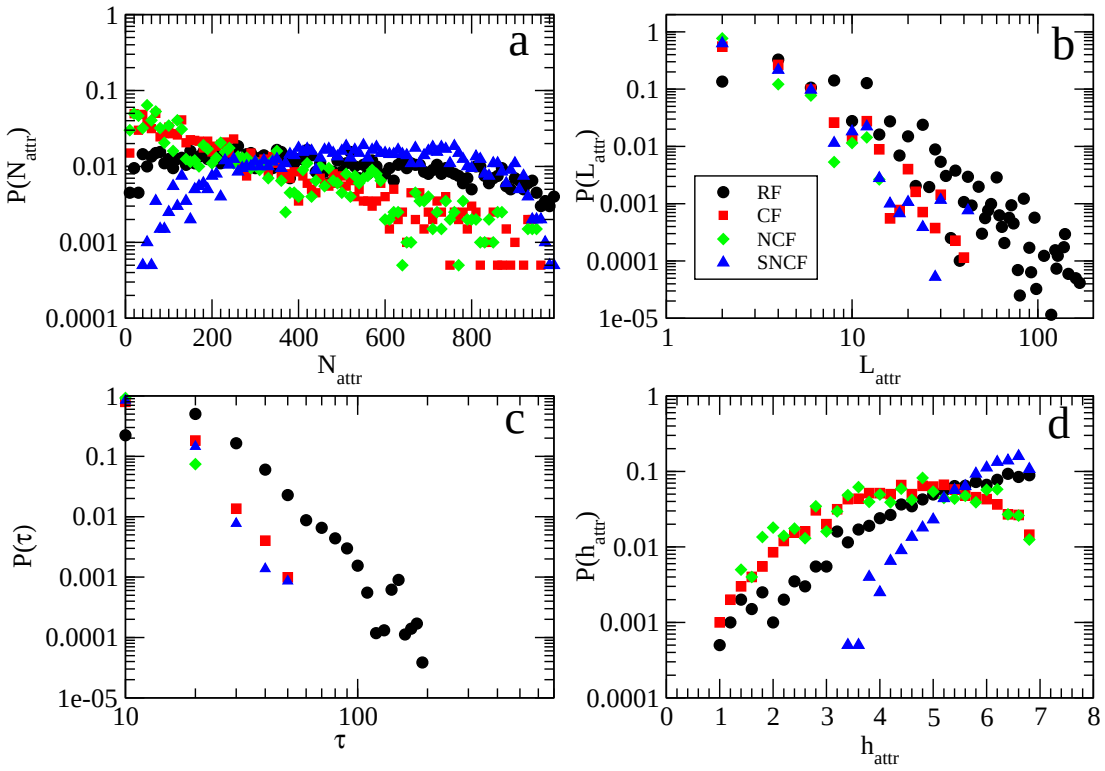

Figure 3.5: Distribution of attractor features of Yeast for Random Function (RF), Canalyzing Function (CF), Nested Canalyzing Function (NCF) and Special Subclasses of Nested Canalyzing Function (SNCF); **a-)** number of attractors N_{attr} probability distribution (Binned in 10 units) **b-)** length of attractor L_{attr} probability distribution (Binned in 2 units), **c-)** transient to attractor τ_{attr} probability distribution (Binned in 10 units), **d-)** entropy h_{attr} probability distribution (Binned in 0.2 units). Attractors were found with starting from 1000 initial conditions of each 2000 network realizations with maximum steps and maximum L_{attr} limits as 1000 and 200, respectively. For RF, CF and NCF, p was fixed to 0.27 which is the p of SNCF for yeast.

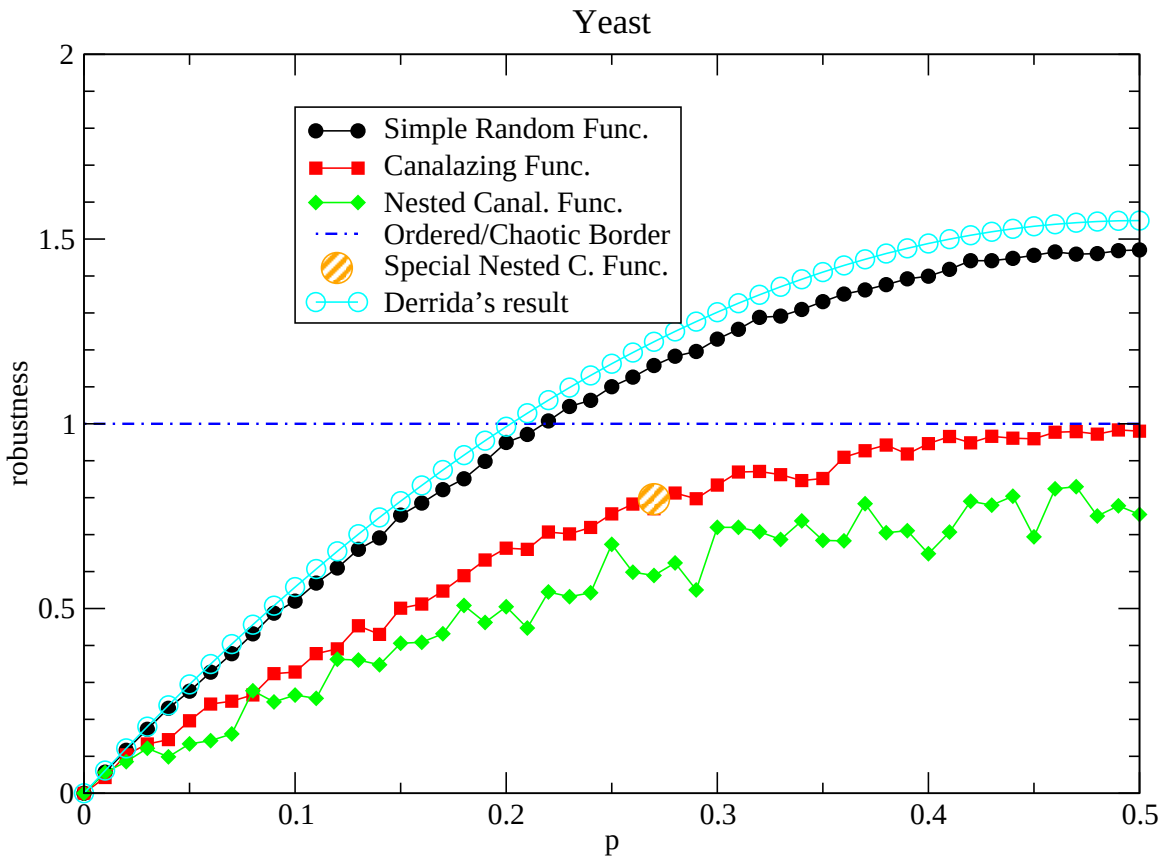

Figure 3.6: Robustness of yeast GRN for all types of functions. For each p value, robustness was computed with averaging over 1000 random initial conditions of each 10 realization. Also, the Derrida's Exp.2.23, $s = 2p(1-p)\langle k_{in}\rangle$ was drawn with $\langle k_{in}\rangle$ of the dyn. rel. subnetwork, 3.1.

that although SNCF and CF gives the same robustness value for the attractor investigation paramater $p = 0.27$, they could produce different attractor structures. In other words, there might be no direct relation between attractor and robustness structure.

3.3 in-EXP Model Networks for Yeast GR

To my knowledge Kauffman is the first who used the model networks to elucidate the gene regulation [16]. He used in-NK networks[4] for modeling the gene regulations. However, in order to compare the dynamics of the yeast, I used 100 in-EXP networks with this exponent since indegree of yeast GRN was an exponential decay with exponent 0.38. I set $N = 94$ for the sake of achieving a similar number of nodes of the dynamically relevant subnetwork of yeast, $N = 82$ (See Figure 2.14 in Chapter 2 for fractions of dynamically relevant nodes to system size for model networks.).

3.3.1 Topological Investigation

Before passing to the dynamics, let me compare the topological features of this model networks with yeast dynamically relevant subnetwork'. Figure 3.7 and 3.8 show the investigations. It can be seen that out- and total-degree distributions and degree-degree correlations resemble that of yeast dynamically relevant subnetwork while others are different.

3.3.2 Dynamical Investigation

Similar to Yeast's case, the dynamics was investigated for $p = 0.27$. It should be noted that the p-value for SNCF is found to be $p = 0.27$ which is the same as Yeast's. In dynamics, 100 networks, for each network 10 network realization and for each network realization 100 initial conditions were created. Again maximum 1000 steps and 200 attractor length were set. The distributions and averages can be found in Figure 3.9 and Table 3.2.

Robustness was also studied for each p $\in \{0.0, 0.02, .., 0.50\}$ for RF, CF and NCF, and $p = 0.27$ for SNCF. The robustness was computed by averaging over 100 networks, 10 dynamics realization for each network and 1000 random initial conditions for each realization. The results are in Figure 3.10.

The yeast and in-EXP model networks produce different attractors features while show similar robustness profiles. The main conclusion of that part is that the model networks which are constructed by knowing only indegree distributions and the network size fail for attractor features predictions while win for robustness.

[4] $K = 2$ case is also known as Kauffman networks in gene regulation literature

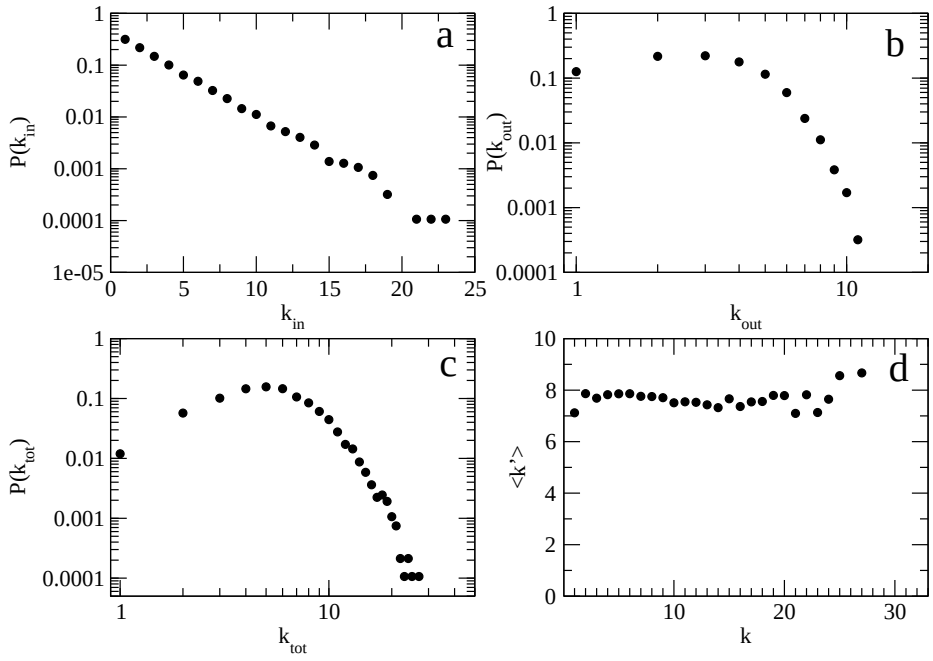

Figure 3.7: in-EXP Model Yeast Network's **a-)** indegree probability distribution **b-)** outdegree probability distribution, **c-)** total degree probability distribution and **d-)** degree-degree correlation

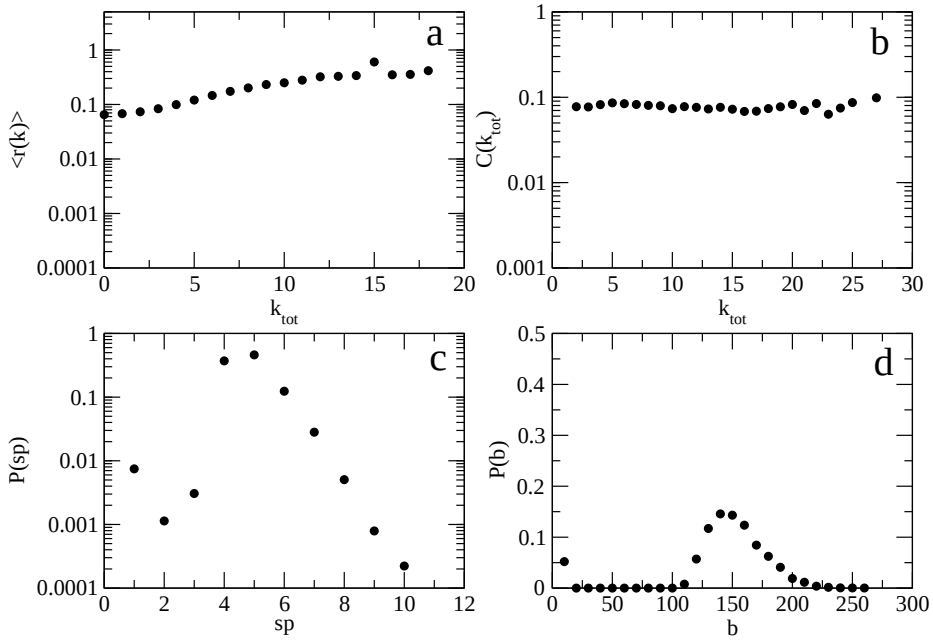

Figure 3.8: in-EXP Model Yeast Network's **a-)** richclub coefficient **b-)** clustering coefficient, **c-)** shorthest path probability distribution (Binned in 1.0 units) and **d-)** betwenness probability distribution (Binned in 10 units)

	$\langle N_{attr} \rangle$	$\langle L_{attr} \rangle$	$\langle \tau_{attr} \rangle$	$\langle h_{attr} \rangle$
RF	5.15 ∓ 5.17	68.29 ∓ 216.64	101.73 ∓ 221.50	0.93 ∓ 0.70
CF	2.87 ∓ 3.50	4.31 ∓ 32.04	11.83 ∓ 32.34	0.53 ∓ 0.63
NCF	1.90 ∓ 1.88	2.07 ∓ 2.37	7.06 ∓ 3.06	0.31 ∓ 0.52
SNCF	3.94 ∓ 5.65	3.70 ∓ 4.56	11.99 ∓ 6.55	0.69 ∓ 0.79

Table 3.2: Average values of attractor features of in-EXP Model Yeast GRN. **RF**: Random Function, **CF**: Canalyzing Function, **NCF**: Nested Canalyzing Function, **SNCF**: Special Subclasses of Nested Canalyzing Function. For the details of the study, see the caption of Figure 3.9.

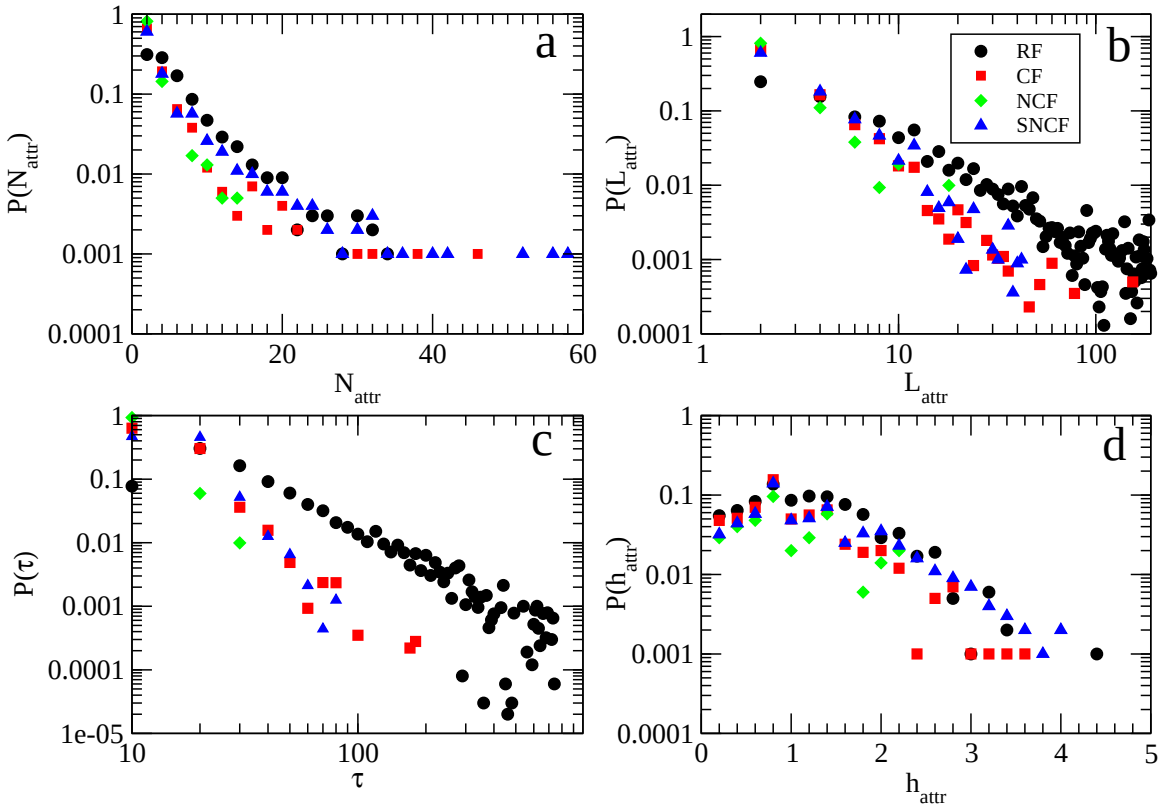

Figure 3.9: Attractor investigation of in-EXP Model Yeast Networks for Random Function (RF), Canalyzing Function (CF), Nested Canalyzing Function (NCF) and Special Subclasses of Nested Canalyzing Function (SNCF); **a-)** Number of Attractors Distribution (binned in 2 units), **b-)** Length of Attractors Distribution (binned in 2 units) **c-)** Transient time distribution (binned in 10 units), **d-)** Entropy distribution (binned in 0.1 units). Attractor were found by starting from 100 initial conditions for each realizations. 10 realizations were done for each of 100 networks. The limits for maximum step size and length of attractors were 1000 and 200, respectively. p of the functions were 0.27

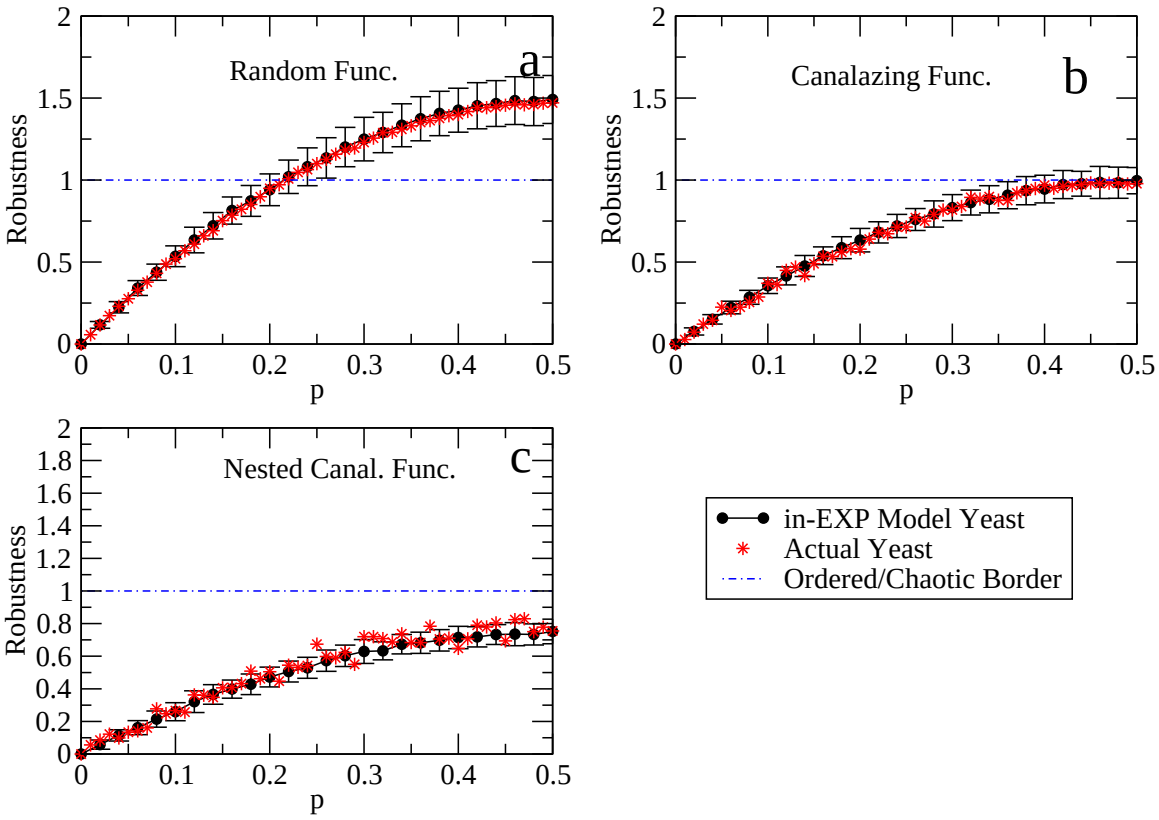

Figure 3.10: Robustness Comparison of Actual Yeast and in-EXP Model Networks. **a-)** Simple Random Func. **b-)** Canalyzing Func. **c-)** Nested Canal. Func.. The robustness was computed by averaging over 100 networks, 10 dynamics realization for each network and 1000 random initial conditions for each realization.

3.4 A Model: Root of the Yeast Gene Regulation Network Topology

Starting from some previous studies [38, 39], Balcan *et al.* have arrived at a novel model produces the complex networks whose *topological* properties resemble that of the yeast gene regulation network [32].

3.4.1 Description of Model

The model is initiated with a starting *fixed number* of **genes**. Next, each gene is assigned to be **Transcription Factor(TF) coding gene** with a probability **p**. The numbers are optimized as 6000 genes at the beginning and $p = 200/6000$ for determining TFs by consulting available actual Yeast data [37].

The model then assigns two types of random binary sequences, i.e. 110100... The former is to all genes being called **Prometer Sequence, PS** and the latter is to only TF coding ones being called **Regulatory Sequence, RS**. Thus, TF coding genes should have two labels where the others have only one.

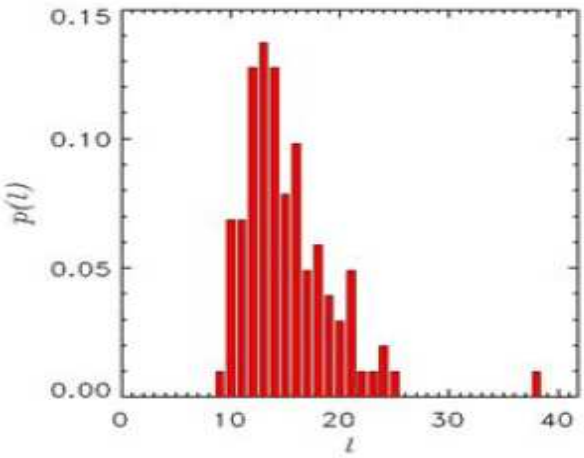

Figure 3.11: Regulatory sequence distribution of yeast [32]

The most important part in assigning these random binary sequences is determining the lengths of the sequence attached to genes. The RS lengths are taken from a distribution yielded from a study of Harbison [40] as shown in Figure 3.11. However, there is no available distribution for the PR lengths; therefore, authors have assumed that the distribution obeys

the power-law with $P(l) \sim l^{-1.1}$. This assumption was based on the investigation of the intergenic portions of the DNA and the result of $P(l) \sim l^{-1-\mu}$ where μ is applicable in $0.0 < \mu < 2.0$ [41]. Morover, again referring the Harbison study [40] which argues that the most of the likelyhood for determining a TF binding site occurs in a 250 bps window on the DNA, they bounded l with the expression $l_{max} - lmin + 1 = 250$ where l_{min} is taken to be pick-value of RS distribution.

After assigning the sequences, a directed edge is placed from TF coding gene (RS) to gene (PS), if and only if a RS is fully inside a PS. These processes can be runned several times and an ensemble can be created.

3.4.2 Topological Investigation (Reproduction of Some Results)

Balcan *et al.* represented in their study that their model creates networks whose topological properties resembles to the actual yeast GRN. I reproduced some of the investigations found in the article [32] as shown in Figure 3.12 in order to be able to elucidate the model. Apart from these topology features, they also stated a similarity in *K-core* structure. In general, the model seems to be very successful for producing the similar topologies. Only topological dissimilarity they found is a detail difference in *motifs*.

I also found a detail mistake in the article which can empower the results. The average number of TFs of the model networks were stated as 202 ∓ 14 in the article, however, according to my computations it was 167 ∓ 14 which is closer to actual number of TFs in the yeast: 146.

3.4.3 Dynamical Investigation of Model

My aim at this part of the thesis is to investigate the Balcan *et al.* model networks in order to detect whether they are also similar with respect to dynamics as they do w.r.t. topology.

For the dynamical investigation 100 Balcan *et al.* model networks were used. As previous cases, investigations were done with using dynamically relevant subnetworks which are found to have 36 ∓ 15 nodes. Comparing to the actual yeast case ($N = 82$), the dynamically relevant nodes are very less in number. Next, both the attractors and robustness were calculated by starting from 100 initial conditions for each of 10 network realization for each network. For finding attractor, the limits for the maximum length of attractor and

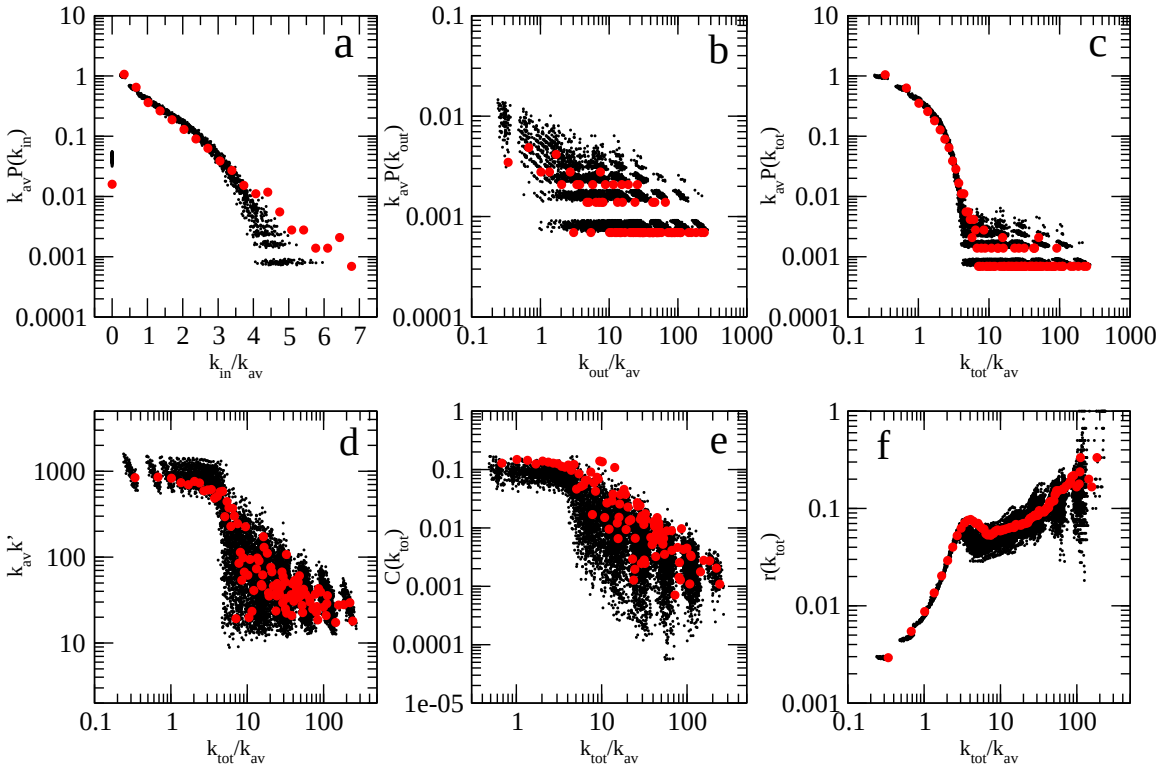

Figure 3.12: Some reproductions of Balcan *et al.* models' results; a-) indegree prob. dist., b-)outdegree prob. dist., c-)total-degree prob. dist., d-)degree-degree correlation, e-)clustering coefficient, f-)richclub coefficent. x- and y-values of the data of corresponding network were multiplied/divided by the average degree.

maximum step size were set to 200 and 1000, respectively. The p value for RF,CF and NCF were set to $p = 0.27$ while p of SNCF was found to be 0.30. The results for attractor features are shown in Table 3.3 and Figure 3.13. The results for robustness is in Figure 3.14.

As a result, Balcan *et al.* model networks failed to mimic the yeast. The results for the average and distributions resemble in-EXP yeast model except that the attractor averages for random functions are considerably bigger. Also, the robustness values for Balcan *et al.* model networks seems to be successful for NCF and SNCF whereas do not give appropriate results for RF and CF. I consider that weak part of the model is that it does not produce a network having as much dense dynamical core as the yeast. This could be related to dissimilarities they stated for the *motifs*. Balcan *et al.* model may be enhanced with considering the dynamcally relevant subnetwork procedure used in this thesis.

	$\langle N_{attr} \rangle$	$\langle L_{attr} \rangle$	$\langle \tau_{attr} \rangle$	$\langle h_{attr} \rangle$
RF	3.14 ∓ 3.51	4.28 ∓ 8.48	10.31 ∓ 12.28	0.57 ∓ 0.62
CF	2.09 ∓ 2.54	2.04 ∓ 2.16	5.47 ∓ 3.28	0.35 ∓ 0.52
NCF	1.44 ∓ 0.77	1.47 ∓ 1.01	4.68 ∓ 2.03	0.21 ∓ 0.36
SNCF	4.36 ∓ 7.95	3.30 ∓ 3.85	9.02 ∓ 5.07	0.74 ∓ 0.78

Table 3.3: Average values of attractor features of Balcan *et al.* Model Networks. **RF**: Random Function, **CF**: Canalyzing Function, **NCF**: Nested Canalyzing Function, **SNCF**: Special Subclases of Nested Canalyzing Function. For details, see the captions of Figure 3.13.

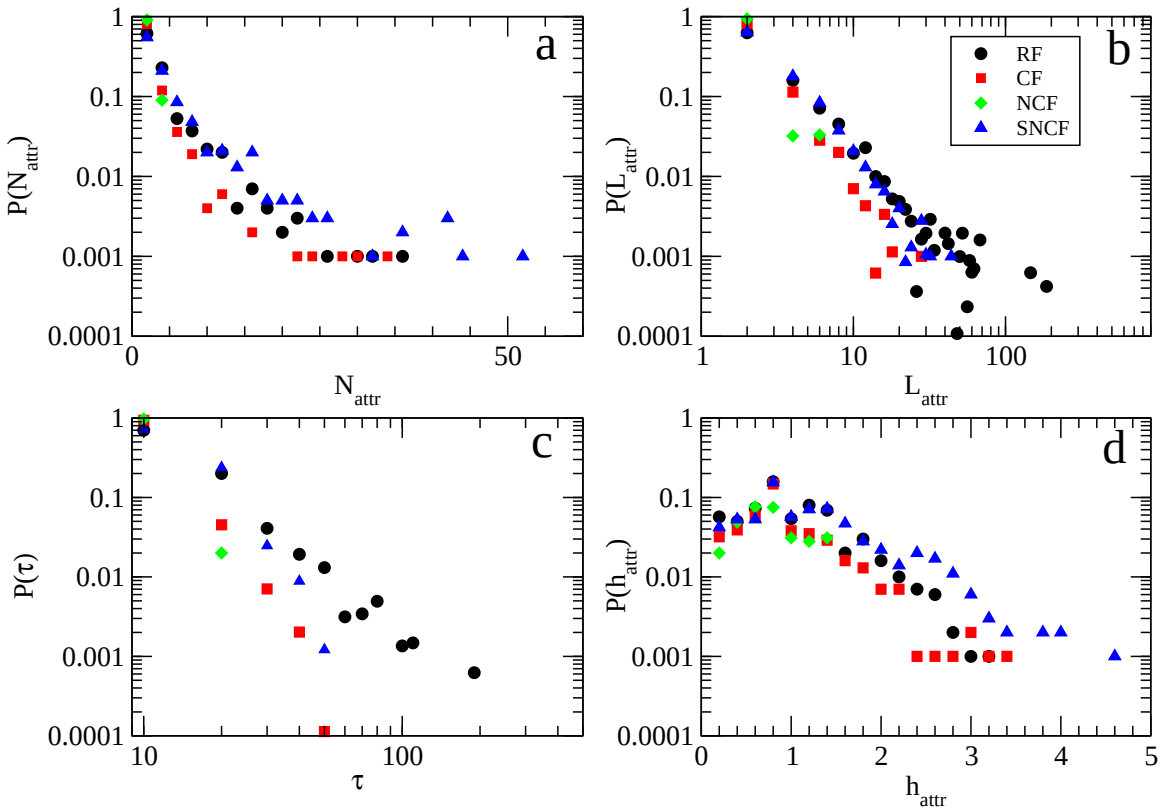

Figure 3.13: Probability distributions for attractor features of Balcan *et al.* Model Networks. **RF**: Random Function, **CF**: Canalyzing Function, **NCF**: Nested Canalyzing Function, **SNCF**: Special Subclasses of Nested Canalyzing Function. 100 model Balcan *et al.* networks and for each network, 10 realizations were created. Attractors were found by starting from 100 initials conditions for each realization where $p = 0.27$ was set for the functions.

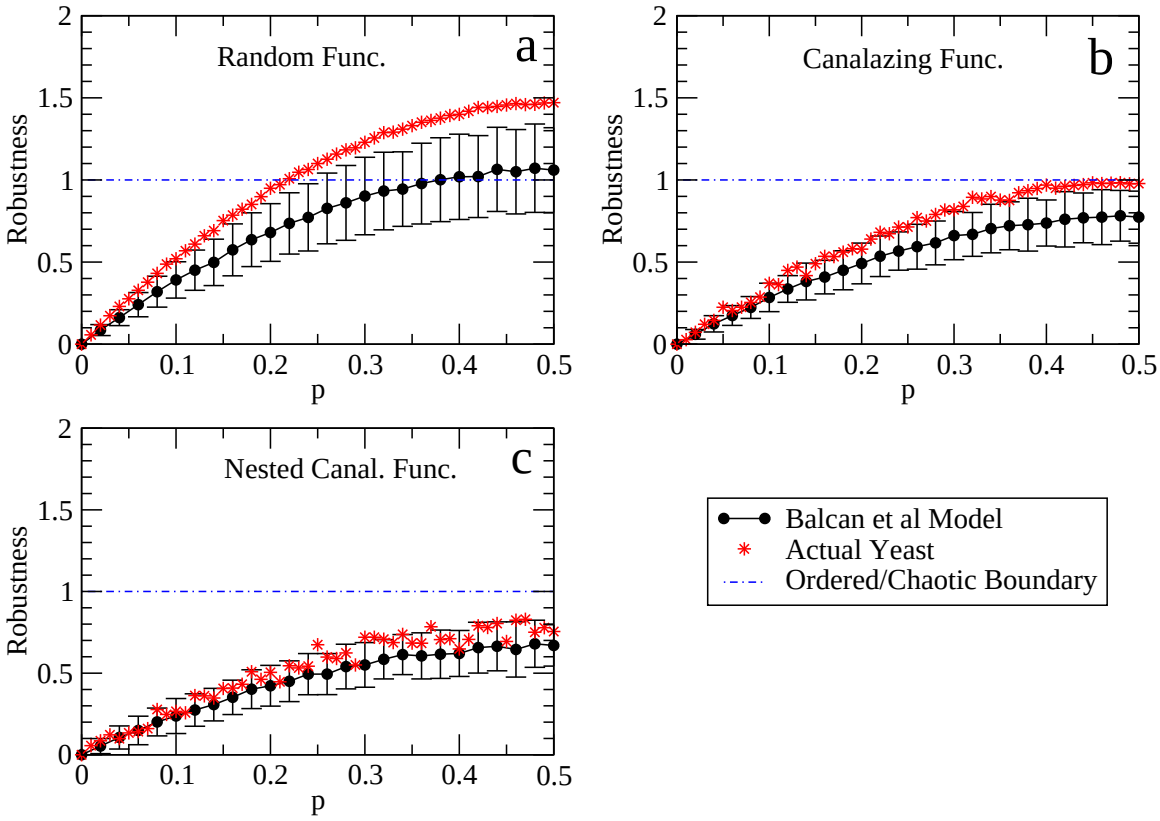

Figure 3.14: Robustness investigation of Model Yeast GRN with comparison to Actual Yeast; **a-)**, **b-)**, **c-)**. Also it should be noted that for Special subclasses of Nested Canalyzing Functions, robustness measure is 0.83 ∓ 0.08 where for Yeast it is 0.78.

Chapter 4

PROTEIN FOLDING

This chapter is organized as follows: Section 4.1 summarizes the protein folding problem in biology. Section 4.2 introduces a new approach to the problem by networks. Section 4.3 gives the investigations on protein *Serine Proteinase Inhibitor*. Section 4.4 emphasizes the results for other proteins.

4.1 Introduction

Living organisms consist of five types of organic compounds: carbohydrates, lipids, nucleic acids, vitamins and proteins. Among these organic compounds, the protein has a privileged place due to its functionality, in the cells [42]. Because of this importance, proteins have been studied highly for decades.

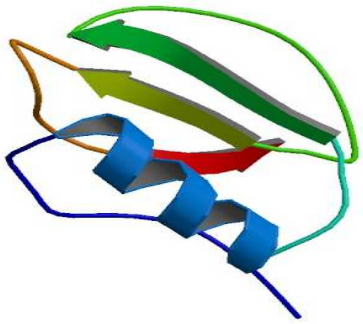

Figure 4.1: The structure of protein *Serine Proteinase Inhibitor* (PDB ID: 2CI2)

A protein is a chain of aminoacids which are bound to the neighbors in the chain by covalent interactions. The natural amino acids are 20 types and the number of amino acids in a protein can vary from 20 to 3000. The amino acid sequence is determined by responsible gene on the DNA. The DNA sends the necessary information to ribosomes via mRNA and they use mRNA to produce the proteins. After this process in ribosomes, this linear chain

immediately folds and becomes more compact in 3D shape named as the **native structure**. This structure is crucial for the functionality of proteins in the cell. The sequence of the amino acids in a protein is unique for this protein and is called the **primary structure** of the protein [42, 2]. Both the primary and native structures of proteins are available in a free-database: Protein Data Bank (PDB)[1] [43].

4.1.1 The Folding Problem

Since the native structure of a protein mainly affects its functionality in the cell, elucidating the mechanism from primary structure to native structure has been one of the leading tasks in protein studies [44]. Shortly, **the protein folding problem** is about finding out the "native" structure of protein from known and unique primary sequence (Figure 4.2). A more detailed review regarding to the protein folding can be found in Reference [44].

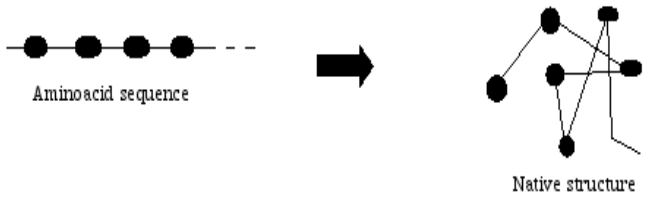

Figure 4.2: The Protein Folding Problem

4.1.2 The Quantifier for Protein Folding

ϕ-value, ϕ_i: To my knowledge, ϕ-value is the only experimental analysis of two-state folding proteins [44]. Two-state folding means that the folding occurs at first from unfolding state (U) to transition state (TS) and next, from transition state to folded state (F). In this analysis a particular aminoacid i is mutated to another aminoacid type which is generally *Alanine* [44]. Later, ϕ_i is determined by using the experimental Gibbs-Free energies of mutant and wild-type protein according to Eq. 4.1 [45]:

$$\phi_i = \frac{\Delta G_{TS-U}^{wild-type} - \Delta G_{TS-U}^{mutant}}{\Delta G_{N-U}^{wild-type} - \Delta G_{N-U}^{mutant}}. \tag{4.1}$$

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

4.2 A New Approach to the Protein Folding

Before getting into new approach, let me define the network of a protein, G(P). Each aminoacid of the protein is a node of G(P) and an edge is assigned to each pair of nodes in G(P) if and only if the real distance between aminoacids in native structure of the protein is less and equal to a certain threshold distance, r_{thr}. One should be aware that G(P) is also named as contact network (map) in literature [46, 47]

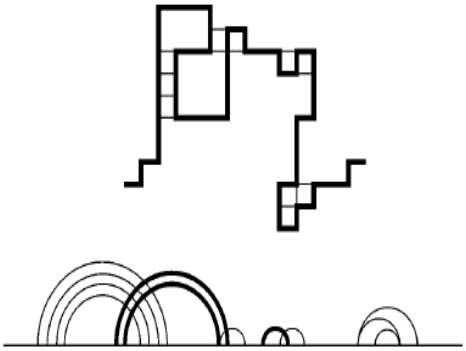

Figure 4.3: Definition of incompatibility networks

Consulting the literature [47, 46] and after some trials, I fixed the threshold distance as $r_{thr} = 6.5$ Å. Also, I prohibited the edges between the aminoacid i and j if $|i - j| \leq 3$.

New approach starts with a definition of another network from G(P) which is named as **incompatibility** network: IG(P) [48]. For definition of IG(P), let me refer to Figure 4.3. In this figure, a protein is extended like a linear chain and starting from v_i and finishing at v_k a half circle is drawn if an edge e_{ik} exists between v_i and v_k in G(P). In incompatibility network IG(P); e_{ik} in G(P) is defined as a node v_{ik} and an edge between v_{ik} and v_{jl} is attached if their half circles in the figure are crossed [48].

Such incompatibility networks have been studied as a tool for understanding mRNA [49] functionality [48] such as determining important regions of mRNA, etc. [50, 51, 52, 53]. Here my aim was to apply this concept to proteins.

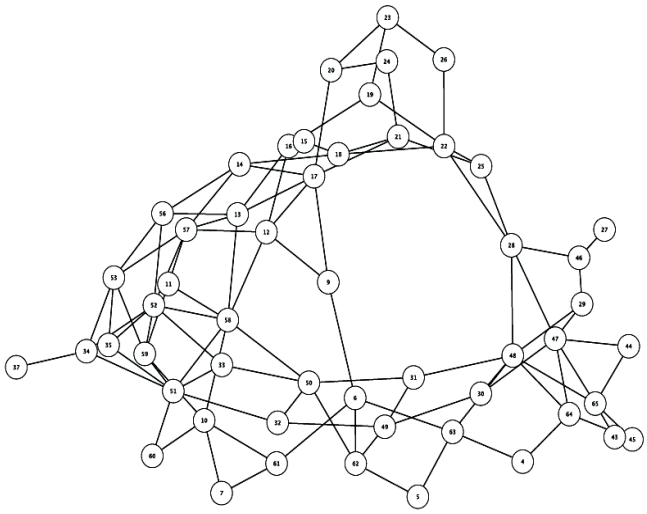

Figure 4.4: Network of protein 2CI2, i.e. G(CI2)

4.3 Example at hand: Serine Proteinase Inhibitor CI-2 (PDB ID:2CI2)

Serine Proteinase Inhibitor (PDB ID: 2CI2) [54] is a simple two-state folding protein with 65 aminoacids. It has been highly studied in literature and this is the main reason to be chosen in this thesis. A ribbon structure of protein 2CI2 is shown in Figure 4.1.

ϕ-value analysis of 2CI2 can be found in Reference [45]. However, I used an extended ϕ-value data of 2CI2 (Figure 4.6-b) and some other proteins (in Appendix C) found by a private communication with Prof. Michele Vendruscolo.

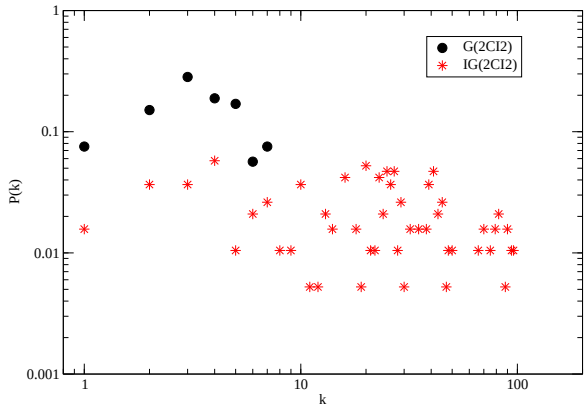

Figure 4.5: Degree probability distribution of normal (G) and incompatible (IG) network of protein 2CI2.

G and IG of 2CI2 are constructed as explained above. G(2CI2) can be seen in Figure 4.4. However, since the figure of IG(2CI2) was not clear due to high number of edges and nodes, it was not inserted here.

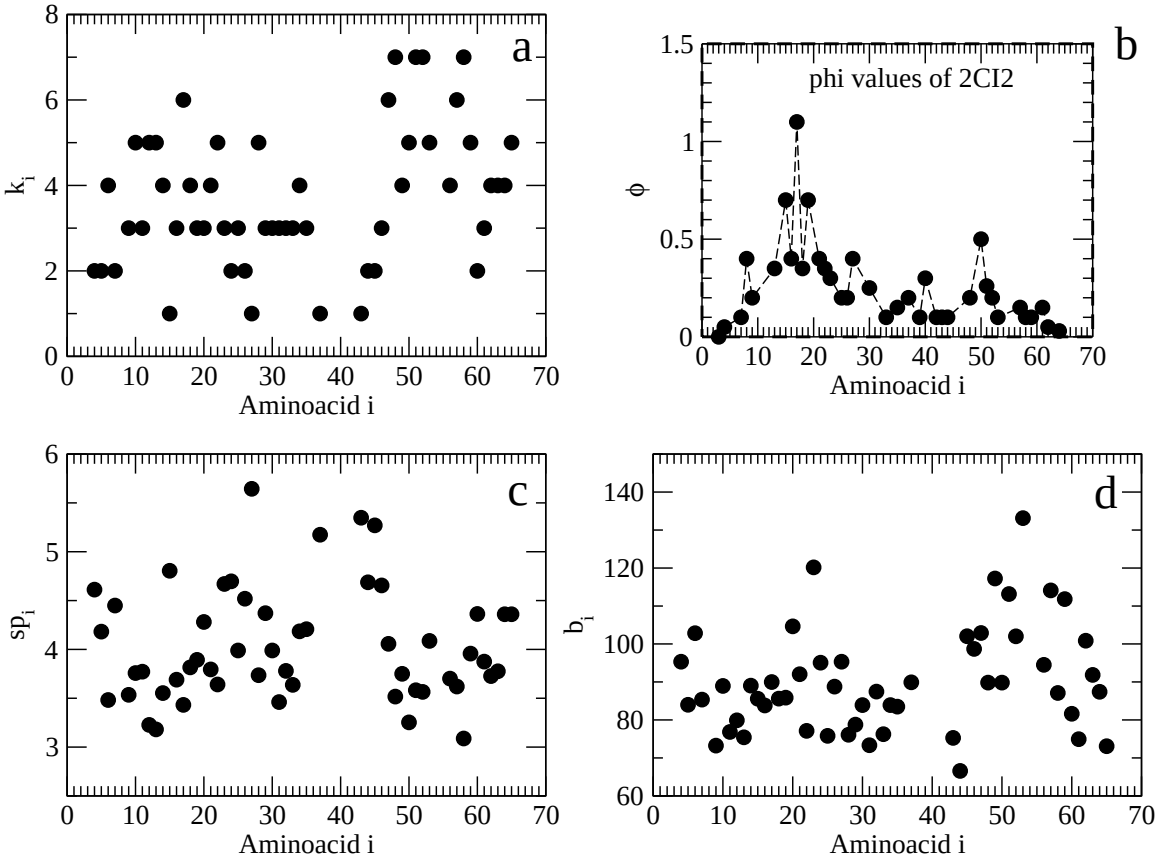

Figure 4.6: The comparison of ϕ-values and topological features of normal network for 2CI2, G(2CI2); **a-)** degree of node i, **b-)** ϕ-values of 2CI2, **c-)** node i's average shortest distance to other nodes, **d-)** betweenness of node i

The degree probability distributions of G(2CI2) and IG(2CI2) are given in Figure 4.5. Other topological features related to nodes are explored for the sake of seeking a correlation to corresponding ϕ-values (Figure 4.6 and Figure 4.7). When a similarity is seen by visual inspection, a statistical analysis is done in order to understand possible correlation. For the explanation of the statistical analysis, let me say x_is are the candidates for being signals to ϕ_i-values. I calculated the original root mean square deviation (RMSD) by $\sqrt{\sum_i (\phi_i - x_i)^2}$. Later, many times I shuffled the ϕ_i to have new ϕ-values sets and each time I calculated another RMSD with new set of ϕ_i which gave me a distribution of RMSD at the end. If the

original RMSD was in the first deviation part of this distribution, x_is lost its candidateship for a signal. With this method, it was concluded that no correlation between topological features and ϕ-values of 2CI2 exists.

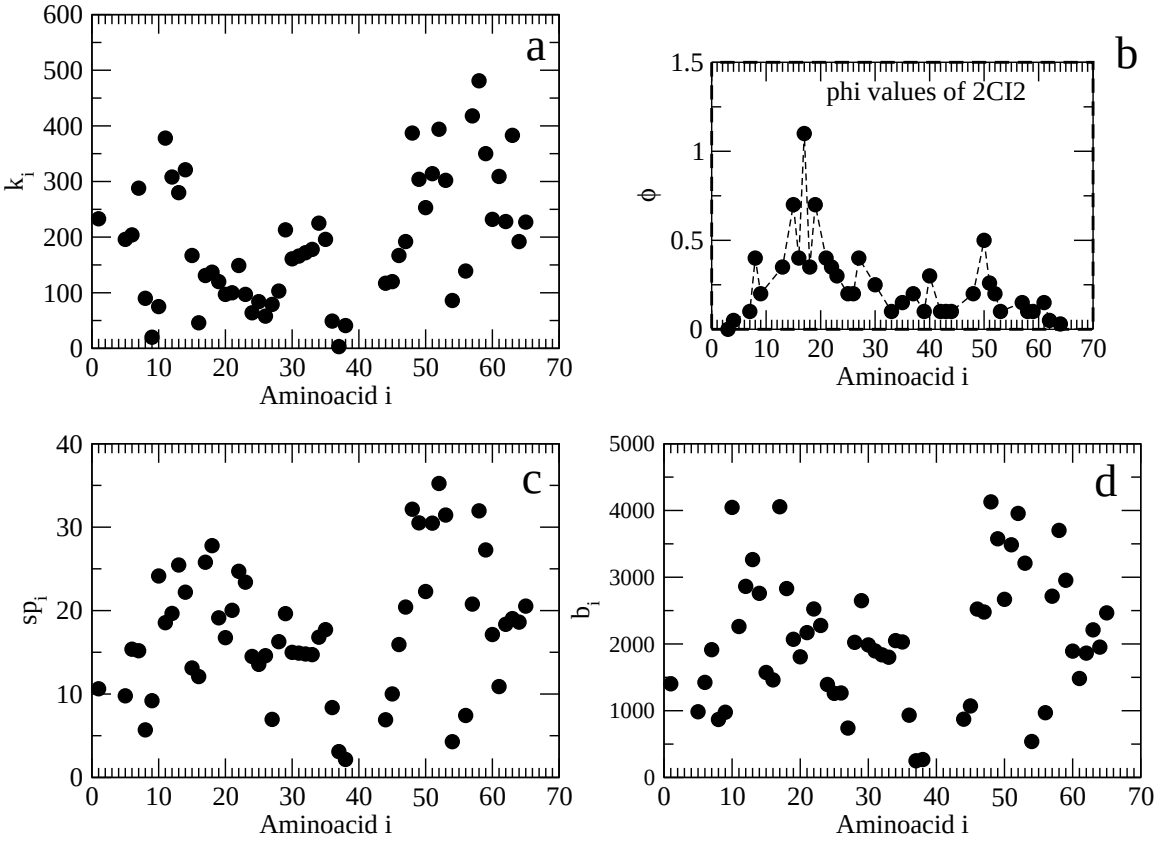

Figure 4.7: The comparison of ϕ-values and topological features of incompatibility network for 2CI2, IG(2CI2); **a-)** degree of node i, **b-)** ϕ-values of 2CI2, **c-)** node i's average shortest distance to other nodes, **d-)** betweenness of node i

4.4 Other Proteins

Apart from 2CI2, other proteins (2PTL, 1SHF, 1TEN and 1APS) are also studied, however, no correlation to ϕ-values was seen for the limited data that is available. The ϕ-value data of the proteins retrieved are given in Appendix C.

Chapter 5

CONCLUSIONS & FURTHER RESEARCH

Gene regulations are an important functional organization of the cells. Activation of the genes in eukaryotes has a dynamical structures. Because of the complexity of the systems, rough but powerful models: networks are used for both topological and dynamical investigations of the gene regulation. In particular, the boolean networks with synchronously updates are studied as generic models for the dynamical investigations. Although the individuals with interactions in the gene regulations are well established in literature, the functions that govern the activation of a gene are not fully understood. For this reason, different types of random functions are used in the studies, i.e. simple random, canalyzing, nested canalyzing and special subclasses of nested canalyzing functions.

Having fixed the network structures and the functions, the boolean dynamics possesses the state cycles called *attractors* which is the main feature of the dynamics to quantify. In particular, the number of attractors, the length of the attractors, the transient length to attractors and basin of attractions are studied for understanding the boolean dynamics. It has been argued that the attractors correspond to some cell cycles in living organisms. Moreover, the robustness of a system is studied as an important property of biological system as "Life at the edge of chaos" hypothesis argues.

In this thesis, the model networks were introduced and investigated both in topology and dynamics. It has been shown that the fraction of dynamically relevant nodes to system size only depends on the average indegree $\langle k_{in} \rangle$ not on the topology type. The attractor features were explored for the network realization parameters $\langle k_{in} \rangle = 2.0$ and $p = 0.5$ which are discussed widely in literature. The distributions show power law decays for model networks. Average values of simple random functions are considerably bigger than canalyzing and nested canalyzing functions. Another important investigation was the scaling of the attractor features with the system size. I have shown that $\langle N_{attr} \rangle$, $\langle L_{attr} \rangle$ and $\langle \tau_{attr} \rangle$ of in-NK networks ($50 < N < 1000, \langle k_{in} = 2.0 \rangle$) and simple random functions($p = 0.5$) scale

with $N^{0.53}$, $N^{0.87}$ and $N^{1.04}$, respectively. The result for the number of the attractor verifies $\sqrt{N}$ theorem while refutes for the length of the attractors. Also, $N^{0.53}$ scaling result refutes the recent study of Kauffman which argues that the scaling of the mean number of attractors is faster than linear. Apart from attractor studies, the robustness of the model networks was studied. It is shown that for all types of the topologies with simple random function, Derrida's expression $s = 2p(1-p)\langle k_{in}\rangle$ is valid with a finite-size effect.

The yeast gene regulation network was investigated dynamically with a boolean approach. For the sake of comparison in attractor features for all function types, p was set to 0.27 which is the output p value for the special subclasses of the nested canalyzing functions (SNCF) for the yeast GRN. The results show that SNCF are successful for the optimization of maximising the number of attractors and minimising the attractor lengths and transients which might be a desirable property for a biological system. Also, the distributions of the attractor features were observed with a nontypical distribution behavior for the number of attractors and the entropy. These distributions are not decreasing for all function types, especially for SNCF. As an important contribution, it was seen that SNCF type may be crucial to elucidate the actual dynamics of the yeast gene regulation. Moreover, the yeast GRN was compared with model networks whose indegree distributions decay similarly. The results show that yeast's attractor distributions and averages are not like model networks. But, it is observed that the robustness structures are similar. As a conclusion, it is seen that to know only the indegree distribution is not enough to produce attractor features of the yeast GRN while being very successful for the robustness behavior. As an another contribution, a recent model: *Balcan et al.* model was discussed and it has been shown that it produces the similar topological networks of the yeast. However, dynamical analyses of this model networks established that the model is not successful at producing neither the attractor nor robustness structures. The main reason of this might be due to that *Balcan et al.* model networks have low number of dynamically relevant nodes ($N = 36 \mp 15$) comparing to actual yeast's ($N = 82$).

At the last part of the thesis, a relation between the protein folding kinetics and a new network approach, incompatibility networks was asked. I have shown that no relation exists between some topological tools and known ϕ-values for the limited data of some proteins.

Further Research: Mean attractor features were studied and the results contradict with Kauffman *et al.*. An extended study which includes also SNCF and other p-values can be done a further work.

In the Appendix B, I discussed some finding attractor algorithms where I pointed a novel algorithm. I believe this algorithm would enhance the related research considerably and should be applied with a low-level programming language such as C++.

The SNCF type has been shown to give different results for the yeast GRN. It should be noted that this type of function is also very time-efficient since it uses a logical formalism (AND and OR functions). All these make me consider that only SNCF type can be used for a further the yeast attractor investigation.

I also saw the success of the Balcan *et al.* model for producing the similar topologies. I believe that this model can be enhanced for the dynamical successes also with a consideration of the dynamically relevant subnetworks used in this thesis and *motifs* in the literature.

Apart from the gene regulation part, I consider that the study I have done should be repeated when there is richer ϕ-value data of the other proteins.

Appendices

Appendix A

ANALYTICAL EXPRESSION FOR $\langle K_{IN} \rangle$

For in-NK networks we have exactly k_{in} edges going into each node, therefore, $\langle k_{in} \rangle = k_{in}$. However, for in-EXP and in-PL networks k_{in} values vary for different nodes. The approximate analytical calculation of $\langle k_{in} \rangle$ follows from substituting the sum with an integral:

in-PL:

$$1 = \int_{k_{in}^{min}}^{k_{in}^{max}} A(\alpha) k^{-\alpha} dk \;\; (Normalization\ cond.) \tag{A.1}$$

$$\begin{aligned} A(\alpha) &= \frac{1-\alpha}{(k_{in}^{max})^{1-\alpha} - (k_{in}^{min})^{1-\alpha}} \\ \langle k_{in} \rangle &= \int_{k_{in}^{min}}^{k_{in}^{max}} A(\alpha) k k^{-\alpha} dk \\ &= A(\alpha) \frac{({k_{in}^{max}}^{2-\alpha}) - ({k_{in}^{min}}^{2-\alpha})}{2-\alpha} \end{aligned} \tag{A.2}$$

in-EXP:

$$1 = \int_{k_{in}^{min}}^{k_{in}^{max}} B(\lambda) e^{-\lambda k} dk \;\; (Normalization\ cond.) \tag{A.3}$$

$$\begin{aligned} B(\lambda) &= \frac{\lambda}{e^{\lambda k_{in}^{max}} - e^{\lambda k_{in}^{min}}} \\ \langle k_{in} \rangle &= \int_{k_{in}^{min}}^{k_{in}^{max}} B(\lambda) k e^{-\lambda k} dk \\ &= B(\lambda) \frac{(k_{in}^{max} e^{\lambda k_{in}^{max}}) - (k_{in}^{min} e^{\lambda k_{in}^{min}})}{-\lambda} + \frac{1}{\lambda} \end{aligned} \tag{A.4}$$

However, since in real case we have quantized k values, i.e. k=1,2,3,..,N. $\langle k_{in} \rangle$ deviates from the Eqs. A.2 and A.4. Correspondence of exponents of PL and EXP networks to $\langle k_{in} \rangle$ for both analytical and actual cases is presented in Figure 2.4. This figure also discuss the main reason of this deviation.

Appendix B

FINDING ATTRACTOR ALGORITHMS

Finding all the attractors of a large network is a challenging task. There are mainly two types algorithms for finding the attractors: **exhaustive** and **heuristic**.

An exhaustive kind algorithm is desired solution which finds all attractors. One can consider two types of exhaustive algorithm. First one is the straightforward method which starts from each initial network states and finds the attractors. However, since the numbers of network states ($= 2^N$) grows exponentially with N, it is computationally infeasible with the available computers. After some trials I concluded this exact algorithm fails for $N > 18$. Second type is a novel algorithm which I have seen during surveying the literature and it claims to find all attractors of N around 100 [55]. Its main idea is to go back from the partial network states (a state description includes only $2-$, $3-$, etc. node sates) and to try to clarify which partial states are impossible to be found in any attractors of that network realization. I have scripted this algorithm with using **python**[1] [56] although I did not get the efficiency stated in the paper [55], mainly because of using a higher level script language rather than a low level programming language such as C++. Yet, it was more powerful than straightforward method approximately for $17 < N < 60$.

Second approach is the heuristic algorithms which sample from the random initial conditions and finds the algorithms. Since the biological networks in the dynamics of interest in this thesis are of size $N \cong 85$ or more I left the exhaustive algorithms and implemented a heuristic algorithm.

[1] Available at http://www.python.org

Appendix C

ϕ-VALUES OF SOME PROTEINS

ϕ-values which are found by private communication with Prof. Michele Vendruscolo are given in Table C, Table C and Table C.

1bf4 res. i	ϕ_i	1bk2 res. i	ϕ_i	1shf2 res. i	ϕ_i
3	0.01	3	0.16	4	0.28
14	0	6	0	6	0.18
16	0	18	0.32	18	0.06
26	1	19	0.29	20	0.22
29	0.44	24	0.22	24	0.41
30	0	31	0.25	26	0.15
31	0.43	38	0.26	28	0.71
34	0.3	39	0.48	39	0.86
36	0.25	41	1	41	1
40	0.22	47	0.58	44	0.74
42	0.21	48	0.61	50	0.37
44	0.59	50	0.53	55	0.01
45	0.09	53	0.16		
50	0.22				
54	0.21				
55	0.27				
58	0.6				

Table C.1: ϕ-values of 1bf4, 1bk2 and 1shf2.

2ci2 res. i	ϕ_i	2ci2 res. i	ϕ_i	2ptl res. i	ϕ_i	2ptl res. i	ϕ_i	1aps res. i	ϕ_i
3	0	42	0.1	4	0.58	31	0.4	11	0.93
4	0.05	43	0.1	5	0.27	32	0.21	13	0.37
7	0.1	44	0.1	6	0.41	33	0.3	17	0.1
8	0.4	48	0.2	7	0.64	34	0.09	20	0.18
9	0.2	50	0.5	8	0.56	35	0.29	22	0.09
13	0.35	51	0.26	9	0.14	36	0.29	29	0.15
15	0.7	52	0.2	10	0.41	37	0.12	30	0.42
16	0.4	53	0.1	11	0.59	38	0	36	0.22
17	1.1	57	0.15	12	0.18	40	0.16	39	0.14
18	0.35	58	0.1	13	0.55	44	0	42	0.37
19	0.7	59	0.1	14	0.79	45	0	45	0.58
21	0.4	61	0.15	15	0.68	48	0.26	47	0.54
22	0.35	62	0.05	17	0.4	49	0.33	51	0.39
23	0.3	64	0.03	19	0.25	51	0.24	54	0.98
25	0.2			20	0.59	52	0	61	0.21
26	0.2			21	0.85	55	0.12	64	0.34
27	0.4			22	0.5	56	0.25	65	0.27
30	0.25			23	0.45	57	0.14	71	0.09
33	0.1			24	0.3	58	0.28	75	0.02
35	0.15			25	0.45	59	0.17	78	0.02
37	0.2			26	0.8	60	0.19	83	0.04
39	0.1			29	0.27	61	0.09	86	0
40	0.3			30	0.08	62	0	89	0.07
								94	0.76

Table C.2: ϕ-values of 2ci2, 2ptl and 1aps.

1ten res. i	ϕ_i	1fmk res. i	ϕ_i	1imq res. i	ϕ_i
1	0.04	1	0	7	0.15
4	0.04	2	0.1	13	0.98
7	0.1	3	0.03	15	0.33
9	0.23	4	0.05	16	0.52
17	0.14	5	0	18	0.4
19	0.39	8	0.03	19	0.32
28	0.13	10	0.28	22	0.31
31	0.19	15	0.13	27	0.12
33	0.35	16	0.26	33	0.27
35	0.53	17	0.03	36	0.25
47	0.67	18	0.4	37	0.15
49	0.42	22	0.62	40	0.01
56	0.38	24	0.55	52	0.03
58	0.6	27	0.77	53	0.07
61	0.33	34	0.25	67	0.41
63	0.47	35	0.15	68	0.23
65	0.25	36	0.54	71	0.36
67	0.42	37	1	76	0.37
69	0.54	38	0.08	77	0.37
71	0.29	39	0.95	83	0.31
76	0.21	40	0.72		
80	0.03	42	0.86		
83	0.21	45	0.68		
85	0.08	47	0.56		
87	0.11	48	0.71		
89	0.11	49	0.24		
		53	0		

Table C.3: ϕ-values of 1ten, 1fmk and 1imq.

BIBLIOGRAPHY

[1] Anthony Serafini. *The Epic History of Biology.* Plenum Press, 1. edition, 1993.

[2] Karp. *The Cell Biology.* John Wiley, 3. edition, 2003.

[3] D.J. Lockhart and E.A. Winzeler. Genomics, gene expression and dna arrays. *Nature*, 405:827–835, June 2000.

[4] A.L. Barabasi and Z.N. Olvai. Network biology: Understanding the cell's functional organisation. *Nature Genetics*, 5:101–114, Feb. 2004.

[5] S. Bergmann, J. Ihmels, and N. Barkai. Similirities and differences in genome-wide expression data of six organisms. *PLOS Biology*, 2(1):1–9, 2003. www.plosone.org.

[6] Lee et.al. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–805, 10 2002.

[7] Douglas B. West. *Introduction to Graph Theory.* Prentice Hall, 2 edition, 2001.

[8] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97, Jan 2002.

[9] Bollobas B. *Modern Graph Theory.* New York: Springer Verlag., 1998.

[10] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, October 2002.

[11] Colizza V., Flammini A., Serrano M.A., and Vespignani A. Detecting rich-club ordering in complex networks. *Nature Physics*, 2:110–115, 2006.

[12] M.E.J. Newman. Scientific collaboration networks ii. shortest paths, wieghted networks, and centrality. *Physical Review E*, 64(0161132):1–6, 2001.

[13] N. Guelzim et al. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, 31:60–63, May 2002.

[14] M. Nicholas et al. Genomics analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312, September 2004.

[15] Gurbacha S. Miglani. *Advanced Genetics*. Alpha Science Int. Lmtd., 2. edition, 2002.

[16] Stuart A. Kauffman. Metabolic stability and epigenesis in randomly connected nets. *Journal Of Theoretical Biology*, (22):437, 1969.

[17] M. Aldana. Boolean dynamics of networks with scale-free topology. *Physica D*, 185:45–66, 2003.

[18] Johannes Norrell, Björn Samuelsson, and Joshua E. S. Socolar. Comparison of boolean and continuous attractors in small networks. *arXiv:q-bio*, 0701052v1, 2007.

[19] L. Mendoza, D. Thieffry, and E.R. Alvarez-Buyyla. genetic control of flower morphogenesis in arabidopsis thaliana: a logical analysis. *Bioinformatics*, 15(7/8):593–606, 1999.

[20] C. Espinosa-Soto, P Padillia-Longoria, and E.R. Alvarez-Buylla. A gene regulatory network model for cell-fate determianation durin arabidopsis thaliana flower development that is robust and recovers experimantal gene expressions profiles. *Plant Cell*, 16:2923–2939, 2004.

[21] Réka Albert and Hans G. Othmer. The topology of the regulatory interactions predicts the expressions pattern of the segment polarity genes in drosophila melanogaster. *Journal of Theoretical Biology*, 223:1–18, 2003.

[22] Harris et al. A model of transcriptional regulatory networks based on biases in the observed regulation rules. *Complexity*, 7(4):23–40, 1 2002.

[23] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein. Random boolean network models and the yeast transcriptional network. *PNAS*, 100(25):14796–14799, 12 2003.

[24] S. Nikolajewa, M. Friedel, and T. Wilhelm. Boolean networks with biologically relevant rules show ordered behavior. *Biosystems*, 2006.

[25] Stuart A. Kauffman. *The Origins of Order*. Oxford University Press, 1 edition, 1993.

[26] I. Shmulevich, M. Aldana, and S.A. Kauffman. Eukaryotic cells are dynamically ordered or critical but not chaotic. *PNAS*, 102(38):13439–13444, September 2005. www.pnas.org/cgi/doi/10.1073/pnas.0506771102.

[27] J.E.S. Socolar and S.A. Kauffman. Scaling in ordered aand critical random boolean networks. *Physical Review Letters*, 90(6):0687021–0687024, 2003.

[28] B. Derrida and Y.Pomeau. Random networks of automata: A simple annealed approximation. *Europhysics Letters*, 1(2):45–49, 1986.

[29] A. Bhattacharjya and S. Liang. Power-law distributions in some random boolean networks. *Physical Review Letters*, 77(8):1644–1647, 1996.

[30] U. Paul, V. Kaufman, and B. Drossel. The properties of attractors of canalyzing random boolean networks. *arXivv:cond-math/0511049*, 1:1–9, 2005.

[31] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. The yeast cell-cycle network is robustly designed. *PNAS*, 101(14):4781–4786, 2004.

[32] M. Mungan D. Balcan, A. Kabakcioglu and A. Erzan. The information coded in the yeast response elements accounts for most of the topological properties of its transcriptional regulation network. *PLOSone*, 6, 2007. www.plosone.org (freely available online).

[33] L.A.N. Amaral, A. Scala, M. Bartelemy, and H.E. Stanley. Classes of small-world networks. *PNAS*, 97(21):11149–11152, 2000.

[34] P. Krawitz and I. Shmulevich. Basin entropy in boolean newtork ensembles. *arXiv:cond-mat*, 07021144v1, Feb 2007.

[35] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. What is a gene, post-encode? history and updated definition. *Genome Res*, 17(6):669–681, June 2007.

[36] V. Filkov, S. Skiena, and J. Zhi. Analysis techniques for microarray time-series data. *Journal of Computational Biology*, 9(2):317–330, 2002.

[37] Miguel C. Teixeira, Pedro Monteiro, Pooja Jain, Sandra Tenreiro, Alexandra R. Fernandes, Nuno P. Mira, Marta Alenquer, Ana T. Freitas, Arlindo L. Oliveira, and Isabel Sá-Correia. The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. *Nucl. Acids Res.*, 34:D446–D451, 2006.

[38] D. Balcan and A. Erzan. Random model for rna interference yields scale free network. *The European Physical Journal B*, 38:253–260, 2004.

[39] M. Mungan et al. Analytical solution of a stochastic content-based network model. *Journal of Physics A: Mathematical and General*, 38:9599–9620, 2005.

[40] Harbison C.T., Gordon D.B., Lee T.I., Rinaldi N.J., and Macisaac K.D. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.

[41] Almirantis Y. and Provata A. Scaling properties of coding and non-coding dna sequences. *J. Stat. Phys.*, 97:233–262, 1999.

[42] Ricki. *Life.* McGrawHill Pub, 4. edition, 1998.

[43] H.M.Berman, J.Westbrook, Z.Feng, G.Gilliland, T.N.Bhat, H.Weissig, I.N.Shindyalov, and P.E.Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000. Available at http://www.pdb.org.

[44] A. Fersht and V. Daggett. Protein folding and unfolding at atomic resolution. *Cell*, 108:573–582, 2002.

[45] John M. Finke, Margaret S. Cheung, and Jose N. Onuchic. A structural model of polyglutamine determined from a host-guest method combining experiments and landscape theory. *Biophysical Journal*, 87:1900–1918, 2004.

[46] Eytan Domany. Protein folding in contact map space. Preprint submitted to Elsevier Preprint, 2000.

[47] Demirel M.C., Atılgan A.R., Jernigan R.L., Erman B., and Bahar I. Idetification of kinetically hot residues in proteins. *Protein Science*, 7:2522–2532, 1998.

[48] A. Kabakçıoğlu and A.L. Stella. Pseudolnots in a homopolymer. *Phys. Rev. E*, 70:011802, 2004.

[49] I. Tinoco Jr and C. Bustamante. How rna folds. *Journal of Molecular Biology*, 293:271–281, 1999.

[50] Z. Du, J. A. Holland, M. R. Hansen, D. P. Giedroc, and D. W. Hoffman. *J. Mol. Biol.*, 271:463, 1997.

[51] N. M. Wills, R. F. Gesteland, and J. F. Atkins. *Proc. Natl. Acad. Sci. U.S.A.*, 88:6991, 1991.

[52] David S. McPheeters, Gary D. Stormo, and Larry Gold. Autogenous regulatory site on the bacteriophage t4 gene 32 messenger rna. 1988.

[53] P. J. Farabaugh. *Cell*, 74:591, 1993.

[54] C.A. McPhalen and M.N. James. Crystal and molecular structure of the serine proteinase inhibitor ci-2 from barley seeds. *Biochemistry*, 26:261–269, 1987. PDB ID: 2CI2.

[55] IRONS David James. Improving the efficiency of attractor cycle identification in boolean networks. *Physica D*, 217:7–21, 2006.

[56] G. van Rossum and F.L. Drake (eds). Python reference manual. *PythonLabs, Virginia, USA*,, 2001. Available at http://www.python.org.

VITA

Murat Tuğrul was born in Tirebolu in April 1, 1983. After living in Tirebolu, İstanbul, Batman, Ankara, Anamur, Afyon and Eskişehir; in September, 2000 he arrived at Middle East Technical University, Ankara where he got his BSc degree in physics and minor degree in philosophy in June, 2005. With an intention to do research in the biological sciences he started the "Computational Sciences and Engineering" master program of Koç University in İstanbul with full scholarship in September 2005. He worked there as a teaching assistant of mathematics and physics courses. He graduates in December 2007 with a thesis title *The Structure and Dynamics of Gene Regulation Networks*. In February, 2008 he will join the Institute for Cross-Disciplinary Physics and Complex Systems in Palma de Mallorca for his PhD degree in physics with an intention to do research in ecology and evolution.