Topological properties of model gene regulatory networks under mutation-selection balance

Marcin Zagórski
Marian Smoluchowski Institute of Physics
Jagiellonian University

Supervisor: Prof. dr hab. Zdzisław Burda

Doctoral dissertation

Kraków 2013
Abstract

Gene regulatory networks arise in all living cells, allowing the control of gene expression patterns. Despite a diversity of living organisms, their regulatory mechanisms encoded at the molecular level exhibit a set of common properties. Gene regulatory networks tend to be sparse, with a low number of incoming interactions per gene and with a broad distribution of outgoing connections. Furthermore, regulatory connections are found to form network motifs, that is small subgraphs of interactions appearing at anomalously high frequencies. Why are these topological properties found in very distantly related organisms? To what extent does the structure of genetic network follow solely from its functional capabilities?

We address these questions in silico using a model that incorporates microscopic interactions between DNA and regulatory proteins. Particularly, Markov Chain Monte Carlo sampling is used to generate ensembles of regulatory networks constrained to have predefined gene expression patterns. By imposing idealized expression patterns, we find regulatory networks to have a low number of connections, narrow in-degree and broad out-degree. Furthermore, by constraining system to have different functional capabilities (multistability vs. time-periodic expression) we identify two disjointed classes of network motifs. By imposing experimentally derived conditions we find that for two yeast species almost all networks have from 50% to 70% of regulatory connections in common with the wild-type network.

The observed topological properties arise as a result of balance between mutation and selection. Due to mutations the system is evolvable, but it is the selection towards imposed constraints that drives the system to functional designs.
Acknowledgements

I am grateful to many people who supported me during my Ph.D. studies in various ways, and I would like to thank a few of them more personally.

I would like to thank my supervisor Prof. Zdzisław Burda who influenced my way of thinking and guided me along the scientific career path.

I would like to thank Prof. Andre Krzywicki for his hospitality and long discussions about physics and life in general.

I would like to thank Prof. Olivier C. Martin who engaged me in the problems laying on the border of physics and biology.

I would like to thank Prof. Wolfhard Janke for a very warm welcome in Leipzig.

I would like to thank Ph.D. Bartłomiej Waclaw for his hospitality, the ongoing collaboration and useful advice.

I would like to thank my colleagues Ph.D. Fakhteh Ghanbarnejad, Ph.D. Jamie Luo, and Ph.D. Areejit Samal who provided me feedback about my research.

I would like to especially thank my wife Ewa for always supporting me in the way of life I had chosen.

The project operated within the Foundation for Polish Science International Ph.D. Projects Programme cofinanced by the European Regional Development Fund, agreement no. MPD/2009/6.

This work was supported by the Polish Ministry of Science and Higher Education Grant No. N N202 229137 (2009 - 2012).

I acknowledge that during the work on my thesis I have been a scholarship fellow of the “Doctus - Małopolski fundusz stypendialny dla doktorantów” project cofunded by EU funds within European Social Fund.
Contents

1 Introduction 1
  1.1 Properties of regulatory networks 2
  1.2 Mutation-selection balance 4
  1.3 Motivation 5
  1.4 Author’s own contribution and thesis plan 6

2 Biological aspects of gene regulation 9
  2.1 Gene regulatory networks 11
  2.2 Topological properties of regulatory networks 12
    2.2.1 Sparsity 13
    2.2.2 Narrow in-degree 14
    2.2.3 Broad out-degree 14
    2.2.4 Network motifs 16
      2.2.4.1 Network motifs in developmental networks 21
      2.2.4.2 Network motifs in signaling networks 23
    2.2.5 Function determines motifs 23
  2.3 Transcriptional regulation 24
    2.3.1 Probability of transcription factor binding 26
    2.3.2 Michaelis-Menten kinetics 29

3 Gene regulatory network models 31
  3.1 Discrete time models 33
    3.1.1 Boolean networks 33
    3.1.2 General logical models 35
    3.1.3 Threshold models 36
### CONTENTS

3.2 Continuous time models ............................................. 37  
3.2.1 Nonlinear ordinary differential equations ................. 37  
3.3 Stochastic modelling .............................................. 40  
3.3.1 Stochastic simulation algorithm ......................... 41  

4 Essential network framework 43  
4.1 Genotype ...................................................... 43  
4.1.1 Probability of occupation of the binding site ............. 45  
4.2 Phenotype ..................................................... 46  
4.3 Transcriptional dynamics ....................................... 47  
4.3.1 Functional constraints .................................... 49  
4.4 Genotype fitness ............................................... 51  
4.5 Viable genotypes ............................................... 51  
4.6 Network of essential interactions ............................ 53  
4.7 Model limitations ............................................. 54  

5 Results for idealized gene expression patterns 55  
5.1 Single target expression ...................................... 56  
5.1.1 Target pattern specification ............................... 56  
5.1.2 Emergence of viable genotype .............................. 57  
5.1.3 Sparsity of essential interactions ......................... 61  
5.1.3.1 Narrow in-degree distribution ......................... 64  
5.1.4 Mutational robustness heterogeneity ..................... 64  
5.1.5 Broad out-degree distribution ............................ 65  
5.1.6 Evolvability of genotypes ................................ 67  
5.1.7 Biological range of parameters ........................... 67  
5.2 Multistability and cyclic conditions .......................... 69  
5.2.1 Target phenotype specification ........................... 69  
5.2.1.1 Multiple fixed points ................................ 69  
5.2.1.2 Cyclic gene expression ............................... 71  
5.2.2 Sparsity of essential interactions ....................... 72  
5.2.3 Basins of attraction for the regulatory dynamics ........ 73  
5.2.4 Topologies of essential networks and evolvability ....... 73  
5.2.5 Functionality leads to motif selection .................... 76  

viii
## Contents

5.2.5.1 Multiple fixed points constraint ........................................ 77  
5.2.5.2 Cyclic gene expression constraint ..................................... 77  

6 Results for experimental gene expression patterns ......................... 81  
   6.1 Model extension .............................................................. 82  
      6.1.1 Computational aspects ................................................. 83  
   6.2 Experimental target phenotypes .......................................... 84  
      6.2.1 Sparsity in biological ensembles ................................... 85  
      6.2.2 Edge usage ............................................................. 88  
      6.2.3 Overlaps with the wild-type networks ............................... 89  
      6.2.4 A multitude of different essential networks ...................... 90  
      6.2.5 Network motifs in biological ensembles ............................ 92  

7 Conclusions and outlooks ......................................................... 97  

8 Appendices ............................................................................. 101  
   8.1 Appendix A: Transcription factor search kinetics ...................... 101  
   8.2 Appendix B: Abundance of functional essential networks ............. 102  
   8.3 Appendix C: Results for mammalian cell cycle ........................... 102  

References ................................................................................. 105
Chapter 1

Introduction

Life on Earth is a very diverse phenomenon, yet all the living organisms are made of simple building blocks called cells. The single cell is a device that interprets environmental signals in order to enhance its survival prospects. In spite of being one of the smallest units of life, cells can deal with starvation, internal damage, or other hostile external conditions, by producing appropriate proteins coded by genes. The cell's response is obviously not decided on the level of conscious reasoning, but it is a consequence of various interactions encoded on the molecular level. Hence, the cell is a processing unit: based on the incoming information, e.g. toxin presence, high/low temperature, lack of oxygen, availability of nutrients, the cell's response follows, and the production of dedicated proteins is triggered. This mechanism of reacting appropriately to different stimuli becomes even more amazing when we realize how "crowded" the cell's interior is: it reminds a molecular soup composed of proteins floating in the cytoplasm rather than a meticulously designed man-made device.

Evolutionary forces have shaped living organisms since the beginning of life, so according to Darwin's natural selection, the living organisms are a result of a long distance race for increasing survival prospect. From time to time, a mutation occurs and the fate of an individual is decided by its adaptation to the environment. Starting from a very simple organism, by a process of tinkering, evolution has resulted in a variety of well-adapted unicellular and multicellular organisms. If we consider the history of these organisms by tracing their evolutionary trajectory, we can reveal the succession of changes that gradually increased their adaptation. But what does actually determine the form of living organisms? How is it possible that sometimes very different
organisms have come up with so similar design solutions? The expectation is that beside evolutionary forces there are also different design constraints that determine organism’s architecture. Specifically selection based on functional constraints may play a significant role. For instance, the design of an organism may be less related to its close-by relatives, than to functional capabilities required from the organism to prevail in its natural environment [1, 2, 3].

In this work we concentrate on the gene regulatory networks, a set of interactions between genes. These interactions along with the gene expression machinery allow living cells to control their output gene expression patterns. The output patterns correspond to production of proteins by certain genes, hence the gene expression profile can be interpreted as a cell’s functional capability or simply the cell’s function. To understand the cell’s behaviour it is crucial to incorporate physical knowledge about regulatory mechanism, and particularly the thermodynamical aspects of protein interactions with other proteins or the DNA strand. As a consequence, by working in a framework rooted in statistical physics, it should be possible to unveil design properties of regulatory networks.

1.1 Properties of regulatory networks

All biological processes should conform with physical laws and constraints. Molecular processes, like protein binding to DNA, conformal changes of molecules, creation of protein complexes (dimers, tetramers), and many others can be explained by writing down the Hamiltonian of the system and finding solutions that minimize energy. Processes like diffusion of gene products, directed transport of molecules or fluctuations in the gene expression level, can be modeled by resolving appropriate stochastic differential equations. The dynamical processes, like production rate of proteins, rate of gene transcription, rate of mRNA translation into protein, or cell growth, can be effectively described with differential calculus. On another level, one can use principles of thermodynamics by evoking canonical or grand-canonical ensembles to estimate probabilities at which certain bindings occur; tendency of the system to maximize its entropy can be used to quantify the information flow through the network of regulatory interactions.

But it should be also clear that biological processes are very complex, and it is not always clear how to reduce their behaviour to a few physical principles. Moreover, if
we find a solution that explains a certain biological mechanism, it is not evident that
the same rule will apply to other organisms. So, is a theoretical approach the correct
choice? Can it be of some use to experimentalists? We believe the answer to these ques-
tions is positive. Indeed, many experiments performed on the molecular level provided
evidence that biological systems, despite their evolutionary background, show similar
properties. In the context of gene regulatory networks, several qualitative properties
transpire: (i) a given gene is generally influenced by a small number of other genes;
the resulting distribution of incoming interactions is narrow \([4, 5]\); (ii) a few genes (so
called hubs) regulate a large number of other genes; hence the resulting distribution for
outgoing interactions is broad, possibly a power law \([5, 6]\); (iii) regulatory networks are
robust to fluctuations in internal (stochastic nature of protein production) or external
signals (variations in temperature, oxygen levels, nutrient abundance); this feature is
also found at many other levels of biological organization \([7, 8, 9, 10]\); (iv) regulatory
networks show modular structure, and particularly certain subgraphs (network motifs)
are overrepresented compared to randomized networks \([11, 12, 13, 14]\).

Because these properties are of rather general nature, there should exist an expla-
nation based on physical aspects of underlying molecular processes. For instance, one
possible explanation for the low number of incoming edges per gene can be related to
physical limitations. In simple organisms the promoter region (DNA sequence that ini-
tiates gene transcription) of a gene is simply too short to incorporate more than a few
binding sites where regulatory proteins (transcription factors) can bind. Yet, this kind
of explanation is of a posteriori type, that is, it does not provide an answer to why, in
the first place, the promoter region for simple organism is so short (~100 base pairs),
whereas we know the promoter length in higher organisms can be several (~1000 bp
for Yeast) or even hundreds times longer (~10000 bp for Human being) \([15]\).

Alternatively, we can try to understand this property through an evolutionary pro-
cess: every interaction between genes is encoded in the DNA\(^1\), so both emergence of
a new interaction and the deletion of the existing one is done through mutations in
the DNA. The selection acting on the level of organisms survival chances affects the
fates of interactions. Due to random mutation process it is more probable that the
interaction is lost (typically one mutation can have fatal effects) rather than produced

\(^1\) A regulatory sequence defining protein that mediates interaction between genes and the sequence
of its binding sites are encoded in DNA.
(only a few mutations are potentially advantageous [16, 17, 18]). As a result, only a few interactions that are specific for cell’s functioning are present in the regulatory network. Hence, if only a gene is functional under a small number of incoming interactions, a regulatory design with a small number of interactions will be preferred over the one with many of such interactions.

### 1.2 Mutation-selection balance

In this type of reasoning, the low number of interactions in regulatory networks (spar-
sity) is an effect of a mutation-selection balance. Random mutations destroy order in the regulatory network, but at the same time natural selection picks solutions which are the most fit regulatory designs. For a physicist this might remind the tradeoff between system entropy and energy: a system tends to be in configurations with higher entropy (more disordered) and simultaneously this system is dragged towards configurations with lower energies (typically more ordered). The tradeoff in the physical system is set by temperature at which these two properties balance out, the system finds itself in an equilibrium in which the tendency to increase entropy is counterweighted by its tendency towards lower energies. Hence, thermal fluctuations play a role similar to mutations, and the temperature corresponds to the mutation rate.

Getting back to biological setup intuitively we feel that, similar to the energy-
entropy tradeoff, mutation and selection should also balance out. The natural selection preserves organisms that are well-adapted to the environment, hence the organisms that perform an environment-specific functions most effectively. The function of an organism is implicitly related to its genetic information encoded in the DNA and all regulatory interactions between genes that make a given organism a successful competitor against other organisms under natural selection. By making a small alteration in the genetic material, the organism’s functionality might stay intact or it might change. If the change is advantageous the organism with this new mutation might invade the pop-
ulation. If the change provides an evolutionary disadvantage, the organism is removed from the population. However, the disadvantageous mutation does not necessarily mean that the modification caused by it is lethal. On the contrary, it might even happen that this new mutation is beneficial, but at the same time it is too costly: under current
1.3 Motivation

It is very tempting to ask, what properties of gene regulatory networks can be explained on the basis of mutation-selection balance? Particularly, if their topological features, like sparsity, narrow in-degree, pleiotropic effect (broad out-degree) or occurrence of certain network motifs, can be understood based on the tradeoff between mutation and selection. Further, to what extent are network topologies evolvable without losing their functional capabilities? Do different selection pressures (like different constraints on the network function) affect regulatory network topology? Specifically, whether a network constrained to produce cyclic expression pattern resembling cell-division cycle, has different network structure, than a network restricted to exhibit multiple fixed points as in cell differentiation process. Finally, to what extent the network structure of wild-type organisms can be predicted from experimentally derived gene expression patterns?

The goal of this work is to address these questions and identify generic properties that arise in regulatory networks due to mutation-selection balance. We address these questions within a relatively simple model of transcriptional regulation \cite{19, 20, 21}, that includes microscopic description of interactions between DNA binding site and transcription factors. Although this model does not include many known aspects of regulation, such as posttranslational modifications or chromatin remodelers, it gives access to questions arising when networks have complex gene expression patterns. Our approach differs from procedures that optimize structure through a design method \cite{22, 23, 24}: we want to get away from any dependence on the optimization algorithm and see how mutation and selection (functional capability) on their own constrain the possible network topologies.
1.4 Author’s own contribution and thesis plan

The author’s own contribution is distributed among chapters 4, 5, and 6. This material is mainly covered by joined publications with Prof. Olivier C. Martin, Prof. André Krzywicki and Prof. Zdzislaw Burda:


The author contributed in this collaboration by taking part in: designing research, performing research, analyzing data and writing papers. Author’s own perspective and insights into presented material were also covered in author’s own publications in peer-reviewed conference proceedings:


The thesis is divided into seven chapters (including Introduction) with additional appendices.

In Chapter 2, we give an overview of biological aspects of gene regulation, define the concept of gene regulatory network, elaborate on topological properties observed in biological regulatory networks, and derive probability of regulatory protein being bound to DNA using the concepts of statistical ensembles and Michaelis-Menten kinetics.

In Chapter 3, we present the most common theoretical frameworks used to model gene regulatory networks: we start with the most qualitative ones that neglect many
biological details (Logical models), next we move to more detailed frameworks that use quantitative knowledge about system (Ordinary differential equations calculus), and we briefly comment on the stochastic description on the level of single molecules (SSA algorithm by Gillespie).

In Chapter 4, we introduce our modelling framework and provide explanations of its components (genotype, phenotype, transcriptional dynamics, fitness, essential interactions), along with comments on computational aspects and methodology.

In Chapter 5, we present results obtained for gene regulatory networks constrained to produce idealized gene expression profiles (single and multiple fixed-points, cyclic expression). We identify and comment on the following generic properties: sparsity, narrow in-degree, broad out-degree, heterogeneous robustness, diversity of obtained topologies and presence of network motifs.

In Chapter 6, we investigate network topologies of three network ensembles generated respectively with two yeast species [25, 26] and the mammalian cell cycle [27] imposed on the network evolution. We show that for yeast species the number of interactions for in silico ensembles is in agreement with wild-type networks of two yeast species. Further, from 50% to 70% of interactions present in the generated networks are found in the experimentally reconstructed networks. Through mammalian cell cycle we illustrate limitations of our modelling.

In Chapter 7, we summarize the main findings and comment on possible future research.
1. INTRODUCTION
Chapter 2

Biological aspects of gene regulation

Genes are sequences of DNA which encode proteins, and proteins perform various functions in the cell. In some cases these proteins can be structural, they not only build cell membrane, but they also take part in cell motility, division, vesicle and organelle movement, and other non-regulatory processes. The gene products can be also enzymes which by catalyzing different reactions allow an organism to convert food into energy, remove toxins by breaking down harmful molecules, or process energy into production of new proteins. The third main group of gene products consists of transcription factors (TFs), that is molecules whose main role is to regulate the expression of other genes. How is this regulatory process realised?

In order for a protein to be produced by a gene, that gene needs to be expressed, the DNA sequence needs to be transcribed by RNA polymerase into mRNA, which is further translated into protein (gene product) by ribosomes. The crucial stage\(^1\) in this process is the localization of a gene promoter by RNA polymerase that binds itself to DNA and initiates transcription. If this regulatory region (promoter) is blocked by a protein (repressor), RNA polymerase will be unable to bind DNA, and hence the gene will not be expressed. Whereas, if the regulatory region is occupied by a protein (activator) that attracts RNA polymerase, the binding of RNA polymerase to DNA will be enhanced; as a consequence the gene production rate will be increased (see Fig. 2.1).

\(^1\)All stages of protein production are important, but without the initial recruitment of RNA polymerase and following gene transcription there will be no final products.
2. BIOLOGICAL ASPECTS OF GENE REGULATION

These repressors and activators are types of transcription factors, regulatory molecules that by binding to target sites at DNA regulate the flow of genetic information from DNA to mRNA [28], resulting in the modification of protein production in the cell.

It is worth to notice, that the whole process of gene expression is obviously more complex from a biological point of view: after transcription, the further processing of resulting mRNA can be obstructed by means of RNA binding proteins that control various steps that prepare transcripts for translation (these events include alternative splicing, nuclear degradation, processing, nuclear export, etc.). On the next step of protein production process, the effectiveness of translation can be modified by targeting the recruitment of ribosomes to the starting position at mRNA. Finally, even after a successful translation of mRNA into protein its activity can be modified by processes like phosphorylation (effectively it means turning on or off protein activity) or by proteolysis which is simply the breakdown of proteins into smaller parts (polypeptides or amino acids). In this sense also enzymes produced by cell, can indirectly regulate gene expression: enzyme activity can be affected by external signals (temperature, pH, availability of substrate), or internal signals (enzyme inhibitors/activators bind to enzyme molecule and decrease/increase its activity), and as a consequence metabolic pathways in the cell are affected leading to changes in concentration of gene products.

Although, the process of gene regulation is quite complicated, the essence of gene being expressed remains simple: if a gene is expressed its protein product is present in the cell\(^1\). Further, if this protein is a transcription factor molecule, it may affect expression of other genes. As a result transcription factors mediate interactions between genes: TF molecule by binding to regulatory region of target gene either enhances or inhibits its transcription rate (the change in transcription rate can be 10- to 100-fold [29, 30]). From here onwards we focus on the transcriptional regulation, since it is a very important player in terms of interactions between genes. Nevertheless, we should keep in mind that especially in eukaryotes, other modes of regulation (like the ones previously mentioned) can become relevant [29, 30, 31]. We discuss more closely biophysical aspects of transcriptional regulation in Sec. 2.3.

\(^1\)Formally, gene expression covers both the expression of its transcript (transcriptome) and of its protein (proteome).
2.1 Gene regulatory networks

Genetic information stored in DNA can be represented by a sequence of four letters (A - adenine, T - thymine, G - guanine, C - cytosine) that correspond to four bases. The information encoded in the DNA sequence defines an organism. Still, the length of DNA sequence or the total number of genes is not something that correlates with the organism complexity (see reviews [32, 33]). Interestingly, the number of protein-coding genes does not change a lot with increasing organisms complexity, which is known as “G-value paradox”\(^1\) [35]. For instance, let us compare the number of protein coding genes for different organisms, for \(E. \ coli\) there are roughly 4000 genes, in baker’s yeast about 6000, fruit fly has around 14000 genes, simple roundworm \(C. \ elegans\) about 19000, and we human beings have around 22000 genes. However, a small sea flea \(D. \ pulex\) has almost 31000 genes [36]! To explain this phenomenon more and more attention is given to regulatory interactions between genes [33, 37].

There is evidence that the number of regulatory genes grows faster than linearly in the total number of genes [2, 38]: for prokaryotes the growth is estimated to be roughly

\(^1\)The G-value paradox is an analogue of the “C-value paradox” [34] that pointed out the lack of correlations between genome size and complexity (already solved in 1970s with discovery of non-coding parts of DNA).
2. BIOLOGICAL ASPECTS OF GENE REGULATION

quadratic; the number of TFs is proportional to \((\text{total number of genes})^{\gamma}\), where \(\gamma\) is estimated to be approximately between 1.75 - 2.0. In eukaryotes this estimate results in \(\gamma \approx 1.2 \pm 0.1\). Furthermore, in a recent study [39] the \(\gamma \approx 2\) was shown to be universal for many prokaryotic clades (complete branches of phylogenetic tree) and lifestyle classes. This tendency might indicate that the newly added genes require more than a linearly increasing number of regulatory genes to be controlled. Further, the nonlinear relation might be caused by introducing more layers in the regulatory networks of more complex organisms or a need for individual regulation of specialized groups of genes [40].

In an idealized picture in which non-coding parts of DNA are omitted, we end up with a set of protein-coding genes. Among these genes there is a subset of regulatory genes that produce TF molecules. These TF molecules depending on the internal or external signals regulate the production rates of proteins (including TFs themselves), hence expression of genes is also regulated. By identifying regulatory interactions within a cell, one can aim for predicting the organism development, phenotypic traits or, ultimately, its cognitive capabilities.

Depending on the required level of faithfulness to biological details, different models of regulatory structures can be proposed (we discuss them in the next chapter), but among them a network representation of regulatory interactions is a very common and useful one.

2.2 Topological properties of regulatory networks

Regulatory network as a directed graph

If we consider only genes that code for TFs, we can easily think of a network in which nodes are genes producing TFs, and links in that network correspond to regulatory interactions between genes. Such a network (or a graph) naturally has directed edges, e.g. activation of gene Y by a regulatory product of gene X is a completely different process than activation of X by a product of Y. Moreover, influence of TFs on other genes is activatory, inhibiting, or there can be no interaction, thus edges in a regulatory network should store a character of interactions. The natural convention is to use the plus sign ("+"") to indicate an activatory link, and the minus sign ("-") for an inhibitory interaction.
2.2 Topological properties of regulatory networks

2.2.1 Sparsity

A network with N nodes can have up to $N^2$ directed links (including self edges), yet this maximal limit is never reached in biological networks. Particularly, the number of interactions found in biological networks with at least several genes is proportional to $N$ rather than $N^2$. For bacterium *E. coli* in [12] the regulatory network was reported to have 577 interactions and 419 nodes (operons - cluster of genes under control of the same regulatory signal or promoter). Hence the fraction of observed interactions to all possible ones is much smaller than 1, resulting in a *sparse* network. Similarly for *M. tuberculosis* in [5] the regulatory network is reported to have 937 interactions and 783 nodes that correspond to protein coding genes.

If we move from bacteria to eukaryotes, for baker’s yeast Lee et al. [11] observed nearly 4000 interactions and 2343 nodes (out of 6270 yeast genes). The nodes in this study indicated promoter regions that were bound by one or more of the 106 transcriptional regulators. By restricting this network to have only genes that produce these 106 regulators, the number of interactions drops to 320. In Chapter 6 we study the baker’s yeast cell cycle network constructed from the key regulators of the cell division cycle [25]: this is a small subnetwork that consists of 11 genes with 29 interactions (excluding 5 self-degradation added by the authors to reproduce idealized gene expression profile).

If we calculate the mean connectivity, given by $K = E/N$ where $E$ denotes the number of edges, for these three full networks and the subnetwork we get $K$ in the range 1.2 - 2.6. Even assuming that the number of regulatory interactions is underestimated (authors in [11], estimate that at the $p$-value = 0.001 threshold about one-third of regulatory interactions might be omitted), the mean connectivity of regulatory networks is typically between two and a few interactions per node.

---

1. Note that the total number of protein coding genes for *E. coli* equals to 4288 (that are organized into 2584 operons), but only a part of these genes play a role in the regulatory process. The latest summary of *E. coli* regulatory network can be found RegulonDB [41].

2. Small regulatory networks or modules in larger networks can have mean connectivity around 10 [42].
2. BIOLOGICAL ASPECTS OF GENE REGULATION

2.2.2 Narrow in-degree

In the regulatory network a given gene can be theoretically influenced by up to $N$ genes. However, in real networks typically the number of incoming interactions is concentrated around the mean connectivity \[4, 12, 43\].

In \[43\] authors report that for the baker’s yeast network the in-degree has exponential distribution, with 93% of genes being regulated by 1 to 4 transcription factor molecules. More specifically, the probability to find a node with $k$ incoming interactions is well fitted by exponential function $\sim e^{-\alpha k}$ where for yeast network $\alpha$ is estimated to be 0.45. For an analogous estimate in \textit{E. coli} \[4\] the resulting $\alpha$ is equal to 1.2. However, for an up-to-date network of \textit{E. coli} acquired from RegulonDB \[41\] we find $\alpha \approx 0.52$ (see Fig. 2.2). In regulatory network of \textit{M. tuberculosis} \[5\] the in-degree is also found to follow exponential distribution with $\alpha \approx 1.78$.

The interpretation of the coefficient $\alpha$ is simple: the higher $\alpha$ the fewer are chances of finding a gene regulated by a large number of other genes. Hence, this coefficient reflects molecular limitations on combinatorial regulation of target genes. As a consequence lower coefficients should be observed in organism with a more sophisticated regulation (\textit{e.g.} required to develop multicellular structure). Indeed, the regulatory regions of genes in simple organisms are not long enough (a few hundred of base pairs) to physically accommodate more than a few binding sites for transcription factors (around 10-15 base pairs per each binding). Although, even a presence of a few binding sites specific for different TFs allows for realization of complex logical functions. As a consequence, higher eukaryotes tend to have more binding sites per gene, allowing for complex computations needed during development \[43\].

2.2.3 Broad out-degree

Contrary to narrow distribution of incoming edges, the outgoing edges in biological regulatory networks tend to have a broad distribution. This means that most of the genes regulate only a few other genes, but there are a few genes that regulate tens of genes, and even in rare cases genes that affect hundreds of other genes. These last genes are called \textit{global regulators} and typically they respond to key environmental signals, for instance in \textit{E. coli} cyclic AMP receptor protein (CRP) can regulate the transcription...
2.2 Topological properties of regulatory networks

Figure 2.2: Left (semi-log plot): Histogram of incoming interactions for regulatory network of \textit{E. coli}: there are 1749 genes, 4164 interactions mediated by 189 regulatory proteins (transcription factors). The data was fitted with exponential function \( p_k = 1155 \exp(-0.52k) \), with a coefficient of determination \( R^2 \) equal to 0.997. Right (log-log plot): Histogram of outgoing interactions for regulatory network of \textit{E. coli}; the central part of data (from \( k = 2 \) to \( k = 64 \)) was fitted with power law decay \( p_k = 42k^{-0.84} \), \( R^2 = 0.81 \). The rightmost point corresponds to cyclic AMP receptor protein (CRP) with 534 outgoing interactions. The regulatory network of \textit{E. coli} was obtained from RegulonDB database [41] on 22th April, 2013.

of more than 100 genes [44] effectively responding to glucose starvation\(^1\). Thus, the out-degree distribution has a long tail of roughly a power law nature, at least over a certain range [6]: it is always bounded by the total number of genes.

For the yeast network [43] the probability of observing \( k \) outgoing edges can be approximated by \( k^{-\gamma} \) (times some normalizing constant), with \( \gamma \approx 1 \). Similarly for \textit{E. coli} the out-degree can be fitted with a power law with a \( \gamma \) also roughly equal to 1. According to [15] \( \gamma \) for regulatory networks can range from 1 to 2. In Fig. 2.2 we present the out-degree distribution for regulatory network of \textit{E. coli} [41], that we fitted with power law in the central part of data; clearly, the data has a broad distribution with a heavy tail.

In terms of complex systems theory, nodes with many more connections than the average value (the key regulators) are also called hubs [45, 46]. It is not straightforward to explain why biological networks have hubs. One option is that their presence makes the network robustness heterogeneously distributed among nodes: the removal of a random node should not affect network functioning, whereas removal of a hub could lead

\(^1\)In the latest RegulonDB [41] data we found CRP to have 543 outgoing interactions.
to loss of viability. Indeed, for knockout experiments performed on yeast *S. cerevisiae* showed that around 75% of genes when removed were not essential for organism’s functioning [47]. Furthermore, in [48, 49] it was showed that the likelihood that a gene is essential (its mutation is lethal or it is toxin sensitive) correlates positively with the number of outgoing interactions of that gene. Hence, some of regulatory networks seem to be vulnerable to loss of their hubs.

However, in the context of bacteria “robustness to node removal does not appear to be the function of the degree distribution in transcription networks” (quoted from [15]). Here, the presence of a long-tailed distribution can be explained by optimizing towards robust performance despite the uncertain environments [50]. The scale-free topology of regulatory networks can be also explained by non-adaptive processes (*e.g.* natural evolution) [51].

### 2.2.4 Network motifs

As we demonstrated regulatory networks are sparse with narrow in-degree and broad out-degree. These are topological properties that describe regulatory network on the global level. It is interesting to see, if topology of regulatory networks has also common features at the local level, *i.e.* restricted to small subsets of connected nodes. Particularly, let us term *network motif* [12, 15, 52] to be a subgraph of regulatory network that is overrepresented compared to a randomized version of this network\(^1\). For instance in Fig. 2.3 we see all possible three node motifs, and in Fig. 2.4 one of this motifs (so called coherent feed forward loop) is embedded in the larger network structure.

In the following subsections we discuss important examples of network motifs found in biological networks. Instead of quantitative description (including reaction specific constants) of regulatory mechanisms, we focus on functional aspects of presented motifs. We also give references to experimental studies in which these motifs were shown to take part in regulatory process.

**Feed forward loop (FFL)**

In order to detect network motifs, one needs to have a null hypothesis about what frequencies of network motifs are expected in random case scenario. This null hypothesis is typically based on an ensemble of networks that were rewired according to node

\(^1\)Generalizations of network motifs are also possible, see [53] for instance.
2.2 Topological properties of regulatory networks

Figure 2.3: All 13 possible three node motifs (with no distinction to activatory/inhibitory edges). Motif no. 4 is a feed forward loop (FFL), and this is the only three node motif out of 13 possible that is commonly found in biological regulatory networks (see text for details).

Figure 2.4: Directed network with edges representing activatory (solid) and inhibitory interactions (dashed). In colored subparts two motifs are embedded: (A) coherent feed forward loop (red background), and (B) mutually inhibitory interaction between two genes (blue background). The network motif highlighted in (B) is often found in developmental regulatory networks (see text for details).
degree preserving rule of Maslov-Sneppen [54]. Having the ensemble of randomized networks, the average frequencies along with standard deviations for network motifs are calculated, enabling discrimination of statistically significant from insignificant subgraphs of interactions in the original network. For instance, in *E. coli* the only 3-node network motif that is found to be overrepresented compared to randomized network is feed forward loop (FFL): in [12] 42 FFLs are reported, with $7 \pm 5$ expected in the degree-preserving random networks. In the case of a fully randomized network (node degrees are not preserved) the expected number of FFL is even lower and equals to $1.7 \pm 1.3$.

Amazingly, both in regulatory network of *E. coli* and yeast out of 13 possible three node motifs only the FFL is a significant motif [11, 52]. Further, each of interactions in FFL motif can be activatory or inhibitory, which gives 8 distinct variants of FFL. These 8 variants of FFL can be divided into two groups depending on whether the direct ($X \rightarrow Z$) and indirect ($X \rightarrow Y \rightarrow Z$) interactions have the same sign (“+” for activatory, “−” for inhibitory; odd number of arrows with “−” corresponds to overall minus sign) (see Fig. 2.5 for node labels). If the sign for direct and indirect interaction is the same the FFL is called *coherent*, and it is termed *incoherent* otherwise. Taking into account this distinction it was observed that mostly one type of coherent FFL (with all activatory edges; 1C-FFL) and one type of in-coherent FFL (1I-FFL) are overrepresented in transcriptional regulatory networks [15].

To understand this phenomenon one might try to decipher the functional characteristics of FFL motifs that are not embedded in the larger network. In Alon’s book [15] or his review paper [14] these functional capabilities are analysed quantitatively in different set-ups, but here we only give an intuitive explanation:

- the *coherent FFL (1C-FFL) with AND logic* ($Z$ requires activation from both $X$ and $Y$ in cooperative manner) works as an asymmetric filter that filters out short (temporary) periods of gene $X$ high (ON) expression. If $X$ responds to external stimuli (like food presence), the temporary turn ON of $X$ would not result in activating $Z$ (only a persistent ON signal activates $Z$). However, when $Z$ is already active, then turning OFF gene $X$ results in immediate deactivation of $Z$. This type of motif is found for instance in the arabinose system of *E. coli* [55],
2.2 Topological properties of regulatory networks

![Diagram of network motifs](image)

**Figure 2.5:** Graphical representation of network motifs: (A) coherent feed forward loop of type-1 (1C-FFL), (B) incoherent feed forward loop of type-1 (1I-FFL), (C) multi-output FFL with $k$ output nodes, (D) single input module (SIM) with $k$ output nodes, (E) coherent bi-fan, (F) coherent diamond, (G) double positive feedback loop, (H) double negative feedback loop.

- the **coherent FFL (1C-FFL) with OR logic** ($Z$ is being activated by at least one of $X$ and $Y$ genes) works as an asymmetric filter that filters out short periods of gene $X$ being OFF. Contrary to 1C-FFL with AND logic, now for activation of $Z$ a short period of $X$ being ON is enough. This type of coherent FFL is favoured if an accidental switching OFF gene $Z$ is costly; for instance in the flagella system of *E. coli* [56], after the construction of flagella motor is initiated the accidental response to fluctuations of environmental parameters should be avoided.

- the **incoherent FFL (1I-FFL) with activatory interactions** ($X \rightarrow Z$) and ($X \rightarrow Y$), and inhibitory ($Y \rightarrow Z$) works as a pulse generator of $Z$ activation. Qualitatively we see that after $X$ is turned ON the gene $Z$ is activated, the expression of protein encoded by gene $Z$ raises, and only after some time (depending on the reaction rates) $Z$ is turned OFF by $Y$. Provided that $Y$ acts on $Z$ as a strong repressor, this type of incoherent FFL might provide pulses in $Z$ expression: $Z$ level first increases and then declines to a low level [57, 58]. Alternatively this type of incoherent FFL can also speed up response times compared to simple regulations, see [59] in the context of galactose system of *E. coli*.

---

19
2. BIOLOGICAL ASPECTS OF GENE REGULATION

Multi-output feed forward loop

The generalization of FFL motif are multi-output FFLs that have a common $X \rightarrow Y$ interaction, but more output nodes ($Z_1, Z_2, \ldots$) that are influenced by $X$ both directly and indirectly through $Y$ (see Fig. 2.5). Multi-output FFLs can introduce temporal order in the expression of output genes: the order is decided by activation thresholds which in turn is fine-tuned by evolution (e.g. type of regulatory protein $Y$, its concentration, life-time, the affinity to target binding sites). Multi-output FFL can also act as asymmetric filters for each of its outputs. This design is also experimentally found in flagella system of *E. coli* [60].

Bi-fan and dense overlapping regulon

Concerning 4-node motifs present in transcriptional regulatory networks, apart from double-output FFL, we have bi-fan motif with two input genes $X_1$ and $X_2$ that jointly regulate the two output genes $Z_1$ and $Z_2$. This type of motif is a simple example of combinatorial decision making device. Depending on input functions of output genes that integrate signals from $X_1$ and $X_2$ (signals can be coherent or incoherent; different activation constants are possible) it is possible to get certain expression patterns as a response to certain ranges of input stimuli. The generalization of bi-fan motif are dense overlapping regulons¹ (DORs [12]; this is no longer a 4-node motif) which are composed of two layers of genes: a set of input genes (input layer) and a set of output genes (output layer) with dense connections between input and output genes, but without connections within layers. DORs integrate multiple inputs to compute the regulation of each output gene. In *E. coli* or yeast, several large DORs are present, controlling tens to hundreds of genes. Concerning functional aspects, DORs typically share common global functions as: response to stress, metabolisms of nutrients, or biosynthesis of crucial cellular components.

¹A regulon is a collection of genes or operons controlled by the same regulatory protein or by a common factor.
2.2 Topological properties of regulatory networks

Single-input module

To make the list of network motifs found in sensory transcriptional networks comprehensive, we discuss the single-input module (SIM) network motif that is formed by a master regulatory gene X (typically with autoregulation) that regulates a group of target genes. Importantly, these target genes are regulated only by X, and the regulation signs (activation/repression) of all interactions are the same (SIM can be also seen as hubs in the network, c.f. Sec. 2.2.3). The main function of SIM is the activation/repression of target genes with a defined time-order. Experimentally SIMs are found in metabolic pathways of arginine system [61] and SOS DNA repair system [62]. In these systems genes are expressed with delays of order 1 tenth of generation between genes (about 5 to 10 minutes). Another example of temporal system includes a genetic network that controls the bacterial cell cycle in Caulobacter crescentus [63, 64].

Can one provide an evolutionary motivation for emergence of temporal order found in SIMs? Indeed, a common denominator that unifies the experimental findings is found: “the earlier the protein functions in the pathway, the earlier its gene is activated” [quoted from [15]]. Hence, the proteins are produced exactly when they are needed. Further, this type of production strategy can be shown, using simplified mathematical models, to be optimal for rapidly reaching a production goal with minimal total number of protein enzymes [61, 65].

2.2.4.1 Network motifs in developmental networks

Up to now we discussed motifs present in the sensory transcription networks, that is networks which sense and respond to external changes. Another type of transcriptional networks are developmental transcription networks [66, 67, 68, 69] that decide on the fate of a cell during differentiation process. Particularly, these types of regulatory networks are responsible for expressing different sets of genes, that define different tissues (like nerves, muscles, blood). Note that these types of networks typically act on a scale of a couple or several cell generations (roughly on a timescale of hours) and their decisions are irreversible, and last even after the initial signal vanishes.

1We do not cover positive and negative one node autoregulatory motifs that are also found to be significant in biological networks (see review by Alon [14] and references therein).
2. BIOLOGICAL ASPECTS OF GENE REGULATION

All of the previously discussed motifs for the sensory transcription networks are also found in developmental networks (fruit fly, worms, and sea urchins [70], human embryonic stem cells [71], skeletal muscle differentiation [72]), but there are also motifs specific to developmental networks. The latter include double positive feedback loop and double negative feedback loop with possible autoregulation.

Double positive feedback loop

In the case of double positive feedback loop genes mutually activate each other, hence both genes are either ON or OFF. It is easy to prove that such a system remains [73, 74] in one of two steady states: X and Y both ON, or X and Y both OFF. This type of bistable switch is also termed a lock-on mechanism [66]. Biologically this bistable switch is mostly used in situations where both X and Y genes are in the same tissue.

Double negative feedback loop

Concerning double negative feedback loop genes mutually inhibit each other, hence they cannot be both ON at the same time. The two possible stable steady states are with either gene X being ON and Y being OFF or vice-versa. (We assumed that both genes would be active without expression, thus excluding the situation with both being OFF). This type of switch is typically present in situations when X and Y belong to different cell fates. A classic example of a double negative feedback loop (with additional autoregulation) comes from a bistable switch in λ phage (a virus infecting *E. coli*), that realizes changes between two operational modes of the virus: lytic (virus multiplies intensively killing the bacterium) and lysogenic (virus acts quietly integrating with host DNA) [29]. In [75] one can find other examples of organisms with regulatory mechanisms incorporating this motif.

Long transcriptional cascade

Another motif that is almost absent in sensory networks, but turns out to be significant in developmental regulatory networks is long transcriptional cascade [76, 77, 78]. This cascade pass information on time steps that correspond roughly to one cell division, which coincides with timing of assigning the cells fate. A cascade can be composed of

\[1\] Notice that this is also a positive feedback loop, since there is an even number of inhibitory interactions.
either activatory or inhibitory interactions. It was shown that cascades including only repressors are more robust with respect to fluctuations in protein production rates [79].

2.2.4.2 Network motifs in signaling networks

Signaling transduction networks are networks that sense the information from environment, process it through protein-protein interactions (not protein-DNA interaction), and afterward regulate the activity of transcription factors. These networks act on a timescale of minutes or even seconds. We are not going to discuss their properties. We only mention that they exhibit bi-fan (already discussed) and diamond motifs.

Diamond motif

Diamond motif is composed of four genes where gene X influences gene Z by two independent indirect pathways going through respectively $Y_1$ and $Y_2$. In signaling network diamond motifs are combined to form multilayer-perceptrons composed of three or more layers of signaling proteins [80]. Such structures are similar to patterns studied in the field of artificial intelligence and can carry out elaborate functions on multiple input signals [81].

2.2.5 Function determines motifs

On these simple examples of FFL functional capabilities, we see that overrepresentation of FFLs in regulatory networks compared to randomized networks might be a result of evolutionary forces that favour particular motifs due to their functional capabilities. Particularly, studies of yeast regulatory network and E. coli showed that network motifs have evolved independently multiple times [1, 2]. This converged evolution was further confirmed in a broad study by Babu et al. [3] that covered 1295 reference interactions across 175 microbial genomes (computationally analyzing over 500 thousands of protein sequences). Interestingly they found, that organisms with similar lifestyles tend to rediscover certain patterns of interactions. Furthermore, what might seem counter intuitive, closely related organisms (in a phylogenetic sense) that do not live in the same environment tend to have different network motifs regulating the same genes.

This convergent evolution of network motifs is possible thanks to rapid (on evolutionary timescales) rewiring of regulatory interactions [1, 2, 82]: only a few mutations
are enough to remove a regulatory binding site in a promoter region, and hence lose the corresponding interaction in the network [82, 83]. However, these few mutations in a DNA binding site or in TF binding domain could lead to evolving new target sites relatively easy, and hence new regulatory interactions [84, 85]. Even genes that are master regulators in one network can lose their functionality: if an organisms adapts to a new environment and the functionality connected with a particular master regulator is not required in this new environment, the loss of this key regulator does not affect fitness of the mutant [3]. In the study of different yeast species [86] the loss of gene activation was reported to be due to adaptation to different environmental conditions.

All these points lead to a conclusion that network motifs are reinvented by evolution due to their particular functional capabilities. Specific functions performed by network motifs can give selective advantage to organisms, as it was shown in a theoretical study of FFL motifs in different environments [87]. Further, network motifs are the most robust and they use the least number of possible interactions among many other possible subgraphs with the same functional capabilities [88].

2.3 Transcriptional regulation

In this section we concentrate on biophysical aspects of transcription factor (TF) binding to DNA. Particularly, we would like to understand the physics of this binding process and derive the likelihood of TF being bound to its target site. The presented derivation closely follows an excellent review by Lässig [89].

Transcription factor molecule has a special part called DNA binding domain, that through hydrogen bonds, allows TF to bind specifically to its DNA binding site. These DNA binding sites (or target sites, or functional sites) can have from 10 to 15 base pairs in prokaryotes and can be even shorter for eukaryotes. Further, these target sites are located upstream of the gene protein-coding part in the so called cis-regulatory region that includes a gene promoter. In the case of bacteria the promoter can extend for \( \sim 100 \text{ bp} \), for lower eukaryotes \( \sim 1000 \text{ bp} \), and in higher organism from \( \sim 10^4 \) to \( 10^5 \) bp [15].

In the case where the TF protein is not attached specifically to the DNA it can either be bound unspecifically to DNA or be unbound and float freely in the cell’s interior. The unspecific binding results in TF sliding along DNA due to electrostatic
2.3 Transcriptional regulation

interaction between the positively charged part of the TF molecule and the negatively charged DNA backbone. Hence, the TF protein can be in three thermodynamic states: (i) unbound, i.e. TF performs a three dimensional diffusion in cell’s cytoplasm, (ii) unspecifically bound, i.e. TF performs one dimensional diffusion along the DNA backbone, (iii) specifically bound.

The biophysics of transcription factor binding to DNA has been comprehensively explained in a series of seminal papers [90, 91, 92, 93]. Furthermore, for some bacterial transcription factors the characteristics of specific binding have been measured experimentally: Cro repressor [94], Mnt repressor [95, 96], c-myb protein [97], cAMP protein [98]. These can be summarized as follows:

(i) The contribution of each nucleotide in the binding sequence to the total binding energy is approximately independent and additive

\[ E(\bar{s}) = \sum_{i=1}^{L} \varepsilon_i(s_i), \]  

where \( \varepsilon_i(s_i) \) denotes the energy associated with the binding of nucleotide \( s_i \in \{A,T,G,C\} \) on the \( i \)th position in the binding site \( \bar{s} \equiv (s_1, \ldots, s_L) \) of length \( L \).

(ii) Typically at each position \( i \), there is one preferred nucleotide \( s_i^* \) that minimizes binding energy at that position \( \varepsilon_i(s_i^*) = \min_{s \in \{A,T,G,C\}} \varepsilon_i(s) \). As a consequence there is a unique best binding sequence \( \bar{s}^* \) with minimal binding energy \( E^* \equiv E(\bar{s}^*) \). This best binding corresponds to the strongest binding between TF protein and DNA, and is also termed the complementary binding.

(iii) The difference between the energy of the complementary sequence and alternative sequences involves an energy cost of \( \varepsilon_i(s_i) - \varepsilon_i(s_i^*) \) approximately equal 1 to 3 in units of \( k_B T \) per nucleotide [96, 99, 100]. This energy is also termed mismatch energy.

(iv) The energy difference between unspecific \( E_u \) and the strongest specific binding is about \( E_u - E^* \approx 15k_B T \).

Since we are interested in order-of-magnitude estimates, it is convenient to use the two state approximation [93], in which we do not distinguish between positions of nucleotides within binding site, and the energy cost for all non-complementary single bindings is the same and equal \( \varepsilon \). Formally, it means that \( \varepsilon_i(s_i) = \varepsilon_i(s_i^*) + \varepsilon \) for \( s_i \neq s_i^* \) and \( \varepsilon_i(s_i) = \varepsilon_i(s_i^*) \) otherwise, where \( \varepsilon \) is typically of order \( 2k_B T \). As a consequence, the difference between binding energy of a sequence \( \bar{s} \) and the perfect matching sequence...
2. BIOLOGICAL ASPECTS OF GENE REGULATION

\( \tilde{s}^* \) is simply \( \varepsilon \) multiplied by the number of positions in which the two sequences have different nucleotides, that is the Hamming distance \( d_H(\tilde{s}, \tilde{s}^*) \) between these two. Thus, the binding energy yields

\[
E(\tilde{s}) = E^* + \varepsilon d_H(\tilde{s}, \tilde{s}^*) .
\]

(2.2)

2.3.1 Probability of transcription factor binding

Let us start with an idealized situation of a single TF protein floating within a cell’s interior. We would like to find the probability of factor protein to get bound to its target site. The TF protein spends about half time on three dimensional diffusion in the surrounding medium, and half time being attached to DNA either in unspecific or specific conformation (the time required for conformational change is neglectable - it is of order 1 - 100 \( \mu \text{sec} \) [15]). This means that the probability of TF being unbound or bound to DNA is the same and roughly equals 1/2. Because we are interested only in the bound state of the protein we do not include the 1/2 factor in further calculations. We calculate probability under condition that TF is already bound to DNA. Furthermore, we use the fact that the TF molecule is at equilibrium between specific and unspecific binding.

Let \( \tilde{s}_j \) denotes the subsequence of the genomic sequence \( \{s_1, s_2, \ldots, s_{L_{\text{gen}}}\} \) from position \( j \) to \( j + L - 1 \), where \( L_{\text{gen}} \) is the length of genomic sequence equal to \( \sim 10^6 \) for bacteria and \( \sim 10^7 \) for simple eukaryotes. Hence, at each position \( j \) the probability of TF being bound in a specific way is given by the Boltzmann factor \( \exp[-E(\tilde{s}_j)/k_B T] \) and by \( \exp[-E_u/k_B T] \) for unspecific way. For a TF molecule the partition function takes a form

\[
Z = \sum_{j=1}^{L_{\text{gen}}} e^{-E(\tilde{s}_j)/k_B T} + L_{\text{gen}} e^{-E_u/k_B T} .
\]

(2.3)

If we assume that the target site of the TF molecule is at the position \( j = j_{\text{tr}} \), the corresponding specific binding energy \( E \) can be defined as \( E(\tilde{s}_j) \) with \( j = j_{\text{tr}} \). We can then single out this position in Eq. (2.3):

\[
Z = e^{-E/k_B T} + \sum_{j \neq j_{\text{tr}}} e^{-E(\tilde{s}_j)/k_B T} + L_{\text{gen}} e^{-E_u/k_B T} \approx e^{-E/k_B T} + Z_0 ,
\]

(2.4)

where \( Z_0 \) is the partition function of a completely random genome sequence. If we denote the free energy of a random sequence by \( F_0 = -k_B T \log Z_0 \), the probability of
2.3 Transcriptional regulation

the TF protein being attached to the functional site is given by

\[
P(E) = \frac{e^{-E/k_B T}}{Z} = \frac{1}{1 + e^{(E-E_0)/k_B T}}.
\]  

(2.5)

This probability has the form of the Fermi function or a sigmoidal function with \( E = F_0 \) being a threshold energy. As a result the probability of TF to be bound to a functional site is close to one for \( E \ll F_0 \) or it is close to zero for \( E \gg F_0 \).

Up to now, we calculated the probability of a single TF protein being attached to its target site, however it is rarely the case that only one TF is present in the cell. Typically, the cell’s interior contains roughly from 10 to even 10000 TF molecules. The lower value comes from some measurements of the multiplicity of transcripts [101, 102], while the higher value comes from other measurements of the numbers of transcription factor molecules [103]. Therefore, it is necessary to generalize the above formalism by using grand-canonical ensemble with chemical potential \( \mu \). After writing down the many-body partition function we can retrieve the average number of TFs from relation

\[
n = \frac{d}{d\mu} \log Z(\mu).
\]

If we neglect the effect from overlap between close sites (this can be done since DNA strand is much longer than binding sequence \( L_{gen} \gg L \)), each binding site \( j \) can be empty or be occupied by either specifically or unspecifically bound TF. The resulting many-body partition function is

\[
Z(\mu) = \prod_{j=1}^{L_{gen}} Z(\mu, j),
\]  

(2.6)

with

\[
Z(\mu, j) = 1 + e^{\mu - E(\vec{s})/k_B T} + e^{\mu - E_u/k_B T}
\]

(2.7)

that corresponds to sum over three possible thermodynamical states: (i) TF is not bound to DNA (freely floating), (ii) single TF is bound specifically, (iii) single TF is bound unspecifically. Following [93, 104] the chemical potential is approximately equal

\[
\mu \approx \frac{F_0}{k_B T} + \log n,
\]  

(2.8)

where it is assumed that the steric exclusion between TFs bound to the non-target sequences is negligible. With \( n \) transcriptions factors floating in the cell we get the probability that a single TF is bound to the target site:

\[
P(E) = \frac{e^{\mu - E/k_B T}}{Z(\mu, j_{tr})} = \frac{1}{1 + e^{E/k_B T - \mu} + e^{(E-E_u)/k_B T}}.
\]  

(2.9)
By using the estimate of the difference between unspecific and specific binding energy, $E_u - E^* \approx 15k_BT$, we can neglect the last term in the denominator. The resulting probability of occupation is then

$$P(E) = \frac{1}{1 + e^{(E - F_0)/k_BT - \log n}}.$$  

(2.10)

The calculation of the free energy $F_0$ of genomic background (random sequence of nucleotides) is a nontrivial task in principle [104]. However, it is possible to put conditions on $F_0$ based on genomic design principles that maximize the programmability of the binding threshold. To put it simply, for a given TF molecule the $F_0$ is always fixed, whereas the value of its target site energy $E$ can be modified by single point mutations. Further, for a TF involved in different regulatory processes (including cooperative bindings with other TFs) it might be desirable to have both weak and strong binding sites. Hence, one can seek characteristics of TF-DNA interactions, e.g. the binding energy $\varepsilon$, the length of binding sequence $L$, the non-specific energy $E_u$, that allow the cell to reach a high level of sensitivity. Examples of such considerations are following:

(i) Let us consider a random genome of length $L_{gen} \approx 10^7$ with a single target site of length $L$ with energy $E^*$ that corresponds to the perfect matching of the TF molecule. The biologically motivated condition that a single TF molecule should affect regulation requires $\exp(-E^*/k_BT) \gtrsim Z_0$, which reduces to

$$F_0 \gtrsim E^*.$$  

(2.11)

(ii) A similar condition can be established from the requirement that the entire genome cannot suppress the specific binding to the strongest binding site $\exp(-E^*/k_BT) \gtrsim L_{gen} \exp(-E_u/k_BT)$, that produces a lower boundary on the energy difference

$$(E_u - E^*)/k_BT \gtrsim \log L_{gen} \approx 15.$$  

(2.12)

(iii) For each target site, the $L$ suboptimal sites (differing by one mismatch from the strongest binding site) should not suppress the strongest binding site $\exp(-E^*/k_BT) \gtrsim L \exp(-(E^* + \varepsilon)/k_BT)$, that gives a lower bound on the binding energy per nucleotide $\varepsilon/k_BT \gtrsim \log L \approx 2 - 3$.  

(2.13)

Further, in [104] authors impose also dynamical conditions on TF-DNA interaction requiring that the effect of kinetic traps is negligible. Quite amazingly all these
2.3 Transcriptional regulation

Figure 2.6: Probability of finding transcription factor (TF) attached to its target site for different values of $n$; following Eq. (2.14) with $\varepsilon = 2$ and $k_B T = 1$ this probability is a sigmoidal (Fermi) function. A low number of mismatches corresponds to strong (specific) binding of TF molecule to DNA. For a high number of mismatches the binding of TF is weak (unspecific) resulting in a very low expression of gene regulated by this TF molecule.

2.3.2 Michaelis-Menten kinetics

The relation in Eq. (2.14) was obtained using statistical mechanics at equilibrium. Let us now get the same relation by incorporating dynamical description at chemical equilibrium.

Let us denote by $S$ the concentration of TF molecules in the cell, by $C$ the concentration of an empty (unbound) DNA target site specific to this TF, and by $[SC]$ the
concentration of this target site bound by TF. Because target site can be either empty or occupied we have a conservation law

\[ C + [SC] = C_T, \quad (2.15) \]

where \( C_T \) is the total concentration of the target site. For instance, if there is only one DNA binding site per cell it gives \( C_T = 1/\text{cell volume} \), which for bacteria equals to \(~1/\mu m^3\); for eukaryotes the volume of cell nucleus is on the order of 10-100 \( \mu m^3 \). Due to diffusion process (see Appendix A) TF and its DNA target site can collide, and by using mass-action kinetics the production rate of their complex is proportional to the collision rate \( k_{on} \), and concentrations \( S \) and \( C \). The complex of TF bound specifically to its target site can also fall apart with a rate proportional to \( k_{off} \) and concentration \([SC]\). Both, these creation and dissociation processes determine the rate of change in \([SC]\) through equation

\[ \frac{d[SC]}{dt} = k_{on}SC - k_{off}[SC]. \quad (2.16) \]

By solving the above equation for a steady-state \( d[SC]/dt = 0 \) we get an equation for chemical equilibrium

\[ K_d[SC] = SC, \quad (2.17) \]

where \( K_d = k_{off}/k_{on} \) is the dissociation constant (it has units of concentration). Combining conservation equation (2.15) with Eq. (2.17) yields an expression (also termed Hill function):

\[ \frac{[SC]}{C_T} = \frac{[SC]}{C + [SC]} = \frac{1}{1 + K_d/S}, \quad (2.18) \]

which is the probability to find DNA target site occupied by TF molecule. From this formula we see that the dissociation constant \( K_d \) can be also interpreted as the concentration at which the gene regulatory region is half-induced. Of course the above relation should match Eq. (2.14), and indeed it does since \( S \times (\text{cell volume}) \) is equal to the number of TF molecules \( n \), and \( K_d \) is proportional to \( \exp(\varepsilon \delta/k_B T) \).
Chapter 3

Gene regulatory network models

Gene regulation is a complex process mediated by gene products. This process can be represented by a directed graph where nodes are associated with genes, and edges represent molecular interactions (like TF binding to DNA or regulatory protein binding to another protein). These edges (links) can encode additional information about the interaction character (activatory/inhibitory), or the strength of interaction (by assigning weights to appropriate edges). But even if more biologically motivated constants (quantifying production rate, degradation rate, cooperativity level, etc.) are added, the network on its own can be an effective description of dynamical process that governs gene regulation. For instance, the presence of an edge between nodes X and Y in a graph, means that whether gene X is expressed it will influence gene Y in a certain way that is defined within dynamics. Depending on the required precision, we might consider more or less qualitative modelling schemes, starting from a very general one like “if X is on it will turn on Y in the next moment”, or being more specific “the time-dependent concentration of Y products is a function of the concentration level of X products bound to Y regulatory region”. In both cases an edge in a graph illustrates the regulatory process, but the edge interpretation depends on the network model that is used to define gene-gene interactions.

Here we cover the most common network models used for describing gene regulatory networks (see extensive review by Hidde de Jong [105] or a more recent one by Karlebach and Shamir [106]). The first distinction that we make is between static and dynamical network models.
In the case of static models there is no explicit time-dependence, so obviously there is no dynamics. Still the static modelling of regulatory interactions can be very useful in the analysis of high-throughput techniques (e.g. microarray techniques [107]), which generate vast amount of biological data. Briefly, the static models are based on finding correlations between gene expression profiles. For instance, in a gene expression profile often a certain group of genes is expressed together (positive correlation) whereas other genes can be rarely expressed together (negative correlation). Such an approach, in the simplest form, can be compared to taking snapshots of time-series of genes expressions and applying various statistical techniques to detect clusters of co-regulated genes [108, 109, 110, 111] (see [105] for additional references1).

On the contrary, dynamical network models predict an expression of genes at certain time point using information from the past time points. Network models incorporating time dependence can be divided into two classes based on their dynamics: with either continuous or discrete time. Network models that use continuous time, like Ordinary Differential Equations (ODEs), are typically better suited for quantitative reconstruction of gene expression patterns. Furthermore, in these kinds of models the real (experimental) time can be included explicitly, giving in turn predictions that can be quantitatively checked.

For network models with discrete time dynamics, the discretization of time can be justified by dividing gene expression pattern into consecutive states (snapshots) that represent different phases of gene expression. Specifically, the different phases of gene expression can be distinguished by setting a threshold value on a gene expression level: for instance, a gene is turned on when its expression (the concentration of its product) exceeds a predefined threshold, and the gene is off in the opposite case2. As a consequence, network models with discrete time dynamics are suitable for gaining a qualitative insight into network of inter-gene regulations. At the expense of reproducing an exact profile of gene expression, these models allow for better understanding of logic behind a gene regulatory system. A representative example of such a model are Boolean

---

1We do not comment static models in the following section, as these models are distantly related to the thesis scope.

2Obviously, in analyzing biological data, the selection of consecutive states in data tends to be difficult: thresholds to be set on gene expression levels are a priori unknown, and moreover the gene expression has a stochastic nature, so the noise level should be also taken into account.
networks which have discrete time dynamics and also discrete gene expression (gene can be either on or off).

In the following sections we present the most popular dynamical network models, starting from ones that have a low level of details (Boolean networks, threshold models), followed by ones that more faithfully represent the experimental data (ODEs), to the most precise models that aim at describing regulation at the level of single molecule events (Stochastic simulation algorithm (SSA)).

Alternatively the order that we choose for model presentation is justified in the following way. We start with models (logical frameworks) that can qualitatively deal with large networks without a lot of biological input, and still are helpful in inferring regulatory network connections. Then, we proceed to more detailed network models (ODEs) that quantitatively reproduce expression profiles, but are more computationally expensive and have lower prediction capabilities. Finally, the stochastic simulation algorithm is a very powerful tool in terms of reproducing regulatory dynamics, but requires feeding with reaction-probability constants for all the chemical processes involved in gene regulation.

3.1 Discrete time models

3.1.1 Boolean networks

Logic-based approach was first introduced by Stuart Kauffman [112, 113] (see also seminal Kauffman’s book [114] and recent review [115]) with its generalization by Thomas [116]. The purest form of logical models are Boolean networks in which genes can be either active (on, 1) or inactive (off, 0): gene products are assumed to be either present or absent in the system. The state of the network is determined separately for each node by a Boolean function that gives 0 or 1 output depending on the state of input genes. The form of a Boolean function is *a priori* not restricted, so various ways of gene regulation are covered.

Formally, let us consider a system with \( N \) genes such that the expression level \( S_i \) of gene \( i \) takes two values 0 or 1. The system state (phenotype) at time \( t \) is then a vector \( \mathbf{S}(t) = (S_1(t), \ldots, S_N(t)) \) of gene expressions; thus, the system state space consists of \( 2^N \) states. For a given gene \( i \) its expression at time \( t + 1 \) is determined through the
Boolean function $G_i(S(t))$ based on the system state in the preceding time point:

$$S_i(t+1) = G_i(S_1(t), \ldots, S_N(t)), \text{ for } i \in 1, \ldots, N,$$ (3.1)

where the Boolean function $G_i$ depends only on a subset of $k_i$ input genes. Notice that $k_i$ can differ for different genes, but in many situations (studying generic model properties) one assumes all genes have the same number of incoming interactions\textsuperscript{1} $k_i = K$. As $K$ grows the number of possible Boolean functions grows like $2^{2^K}$. This means that for $K = 2$ we have 16 possibilities, that include the most widely used AND, OR, NAND, NOR, XOR logic gates.

The transition from one phenotype $S(t)$ to the next one $S(t+1)$ is usually determined in a synchronous way, that is the system state at $t + 1$ is obtained by simultaneously updating all gene expressions at $t$ according to Eq. (3.1). As a consequence of the synchronous update procedure, the system phenotypes are determined in a deterministic fashion (for a given input phenotype there is exactly one corresponding output phenotype). Alternatively, in an asynchronous update procedure the expression of single genes are modified according to Eq. (3.1) in a sequential manner, so if the order of updates is random, it may result in a non-deterministic dynamics\textsuperscript{2}.

If we stay with the synchronous update procedure, the trajectories of the system are well defined, that is the system transits from one state to another in a unique way. Since the state space is finite the system will eventually reach a steady state (fixed point) or a cycle state (cyclic attractor). All states that are not attractors or part of attractors are called transient states. Further, all the transient states leading to an attractor constitute (together with this attractor) the basin of attraction.

Interestingly, Kauffmann showed [114] that for a network composed of $N$ genes with randomly assigned $K$ inputs and a certain choice of regulatory functions the system can exhibit different classes of dynamical behaviour. Specifically, in annealed approximation the critical connectivity $K_c$ can be estimated [117]. For $K$ below this value the system dynamics is highly ordered and for $K > K_c$ the system is chaotic. For Boolean functions that have roughly the same number of 1s and 0s in the output, $K_c$ is equal to 2 [115, 118]. Hence, for low $K$ values ($K < K_c$) a small perturbation

\textsuperscript{1}$K$ can be also identified with network connectivity or mean number of links, hence $K$ can have non-integer values.

\textsuperscript{2}The randomness in the system can also be instilled by the introduction of noise in the updating rules, but this is beyond the scope of the standard Boolean approach.
of input state (e.g. changing state of one gene from 0 to 1 or vice-versa) leads to no or only a small change in the output state. Whereas for moderate and high values of $K$ ($K > K_c$) a small perturbation of input state can cause large changes in the output state. The emergence of this chaotic behaviour is intuitive: when system connections become more dense (the connectivity parameter $K$ is increased) even change in one bit of input state can easily propagate with each step of dynamics.

If the connectivity $K$ is well-tuned ($K = K_c$) it is possible to observe the system at the edge of chaos, that is in the regime between highly ordered and disordered behaviour. At this border between ordered and chaotic regime Boolean networks exhibit properties typical for biological systems: (i) the stability towards perturbation and, at the same time, the ability of the system to evolutionary improvements is ensured [114], (ii) a small value of $K_c$ means that the network is sparse with a small number of incoming regulations per gene.

Boolean networks were also among the first frameworks used to infer the genetic network structure (see for example [119]). Because Boolean networks sacrifice a lot of biological details they are well suited to studies of global properties of large regulatory networks [114, 120, 121]. However, certain assumptions about Boolean networks, like on/off representation of gene activity, might result in omitting regulatory regulations that take place at the level of intermediate values of expression. Moreover, the synchronous update procedure may lead to predictions which are no longer valid with asynchronous update that is more biologically realistic.

3.1.2 General logical models

To overcome these drawbacks more general logic models have been proposed [116, 122, 123], allowing for more than two values of gene expression and enabling asynchronous transitions between states. The method proved its effectiveness by describing process of λ phage infection in E. coli [124], or the regulatory network controlling flower development in Arabidopsis thaliana [125, 126]. In the latter the four steady states exhibited by the network dynamics were found to correspond to plant floral organs: sepals, petals, stamens and carpels. More recently the logic-based formalism was used to highly detailed reconstruction of the regulatory network that controls the first several hours of sea urchin embryos development [66, 127].
3. GENE REGULATORY NETWORK MODELS

3.1.3 Threshold models

A particular example of a Boolean model can be obtained by having all regulatory functions to be threshold functions \[128\], such that if the threshold is exceeded the associated gene is activated. Using the same notation as for Boolean networks, the state of the system with \(N\) genes is represented by gene expression pattern \(S(t) = (S_1(t), \ldots, S_N(t))\). The regulatory interactions between genes are defined through \(N\) by \(N\) matrix \(W\) with elements \(W_{ij}\) representing the strength of interaction between gene \(j\) and gene \(i\). Particularly, for \(W_{ij} > 0\) the \(j\)th gene activates the \(i\)th gene, for \(W_{ij} < 0\) the interaction is inhibitory, and the influence is absent for \(W_{ij} = 0\). The matrix \(W\) can be considered as the regulatory genotype of this system, whereas \(S(t)\) is the system phenotype.

In the simplest case, one can restrict \(W_{ij}\) values to three possibilities: -1 (inhibition), 0 (absence of interaction), 1 (activation). Furthermore, if the number of ingoing interactions per gene is not restricted, the influence of all genes on gene \(i\) can be expressed as a sum of terms \(W_{ij}S_j\) for \(j = 1, \ldots, N\). Such additive approximation is a very crude one (it neglects cooperativity effects), yet for a sparse systems (like biological regulatory networks) in which only a few interactions per gene are meaningful, it can give qualitative insights into system dynamics. Using this approximation we can map system phenotype at time \(t\) to phenotype at the next time point through the following rule

\[
S_i(t + 1) = G_i(\sum_{i=1}^{N} W_{ij}S_j(t)) ,
\]

where

\[
G_i(\sum_{i=1}^{N} W_{ij}S_j(t)) = \begin{cases} 
1, & \text{if } \sum_{i=1}^{N} W_{ij}S_j(t) > 0 \\
0, & \text{if } \sum_{i=1}^{N} W_{ij}S_j(t) < 0 \\
S_i(t), & \text{if } \sum_{i=1}^{N} W_{ij}S_j(t) = 0 
\end{cases}
\]

Instead of 0 value as the threshold discriminating between activity or inactivity of a gene other gene-specific (experimentally motivated) values can be used \[26\]. Further, the step-like function from Eq. (3.2) can be replaced with a sigmoidal function \[129, 130\] (or other nonlinear relations, e.g. \[131\]) that allows also for intermediate values of expression, leading to models with a continuous gene expression.

Although the model might seem a bit too simple for explaining regulatory interactions, it has been successfully used in reconstruction of baker’s \[25\] and fission \[26\]
Moreover, within this framework it is possible to explicitly enumerate all functional networks that produce a given gene expression pattern, that is a trajectory in the state space:

\[ \mathbf{S}(0) \rightarrow \ldots \rightarrow \mathbf{S}(T) . \]  

The exact construction of the space of functional networks is possible due to property of Eq. (3.2) and Eq. (3.3) that allows to treat rows of \( \mathbf{W} \) separately, i.e. for the first row in \( \mathbf{W} \) determine all possible combinations of \( W_{1j} \) that give the target pattern \( S_1(0) \rightarrow \ldots \rightarrow S_1(T) \) for the first gene, then analogously work out all the combinations \( W_{2j} \) for the second row, etc. Hence the problem of finding functional networks in the space of all \( 3^N \) possible networks (for baker’s yeast \( N = 11 \), and the number of networks is roughly \( 5.39 \times 10^{57} \)) reduces to finding for any of \( N \) rows the functional combinations among \( 3^N \) possible rows (for \( N = 11 \) it gives 177,147 possibilities). By finding all the possible combinations for all rows, the space of all networks is just the Cartesian product of single row subspaces. Using this observation, in [132] it was showed that the number of networks that can realize the cell cycle process for baker’s yeast is \( 5.11 \times 10^{34} \); within this huge amount of possible network topologies there was one corresponding to the wild-type network (experimentally reconstructed baker’s yeast network [25]).

Apart from reconstructing the space of all functional networks, also for observables that can be calculated separately for rows one can get exact estimations for all the functional networks (like sparsity, in-, out- degree); for other observables the functional space can be sampled in an efficient way [132]. Within this threshold framework one can also easily study the effect of gene knockouts [133] or the connectivity of the space of functional networks [134]. As a consequence the system with moderate values of \( N \) and neglected strength of interactions can be exhaustively investigated.

### 3.2 Continuous time models

#### 3.2.1 Nonlinear ordinary differential equations

Although a discretized gene expression profile might be useful, by resorting to continuous timescale and real-valued gene expressions it is possible to compare model predictions directly with biological data. Particularly, if we keep the notation from
previous sections, the $i$th gene expression $S_i(t)$ is now a non-negative continuous variable that can be identified with concentration of the $i$th gene products (or alternatively the average number of the $i$th gene products). Hence, the rate equations for the set of all genes can be written as

$$\frac{dS_i}{dt} = g_i(S_1(t), \ldots, S_N(t)),$$  \hspace{1cm} (3.5)

where $g_i(S)$ is a nonlinear function that serves as an ansatz for the real molecular dynamics. Typically the $g_i$ function will be a sum of production terms that include production rates associated with different aspects of gene regulation, and degradation terms that correspond to different biophysical processes that decrease protein number.

Equation (3.5) can be extended to not only include dependence on gene products, but also to include other internal (e.g. cell mass) or external (e.g. concentration of supplied nutrients) signals. Furthermore, discrete time delays $\tau_{ij}$ associated with time required for products of the $j$th gene to undergo transcription, translation and diffusion process required to reach the $i$th gene regulatory region can be incorporated by replacing on the right hand side every $S_i(t)$ with $S_i(t - \tau_{ij})$.

For educational purposes let us solve a simple regulatory system made of two genes, such that the second gene activates the first gene. More specifically, the $S_2$ transcription factor molecules produced by the second gene are present, and each of this TF can be bound or unbound to a single DNA target site of the first gene. In case the TF occupies this DNA target site, the genetic material undergoes transcription, and then translation resulting in the production of $S_1$ molecules at the effective rate $k_{12}$. Apart from being produced, these molecules are degraded with a characteristic time $\tau$. Hence, the production rate of the first gene proteins is given by

$$\frac{dS_1}{dt} = k_{12}P_{12}(S_2(t)) - \frac{1}{\tau}S_1(t),$$  \hspace{1cm} (3.6)

where $P_{12}$ is a function that gives a probability of TF (encoded by the second) gene being attached to the regulatory region of the first gene; $P_{12}$ can be also interpreted as a mean occupancy of the first gene target site.

Since the process of TF binding/unbinding reaches equilibrium typically on time-scales shorter than $\tau$ (seconds vs. minutes), we can plug into Eq. (3.6) the form of $P_{12}$ derived in Sec. 2.3.2:

$$\frac{dS_1}{dt} = k_{12} \frac{S_2(t)}{S_2(t) + K_d} - \frac{1}{\tau}S_1(t),$$  \hspace{1cm} (3.7)
where we assumed that there is no feedback (regulation) of the second gene by the first one. Hence, for a fixed $S_2(t)$ the approach to equilibrium ($\frac{dS_1}{dt} = 0$) is exponential with the rate $\tau$, and the steady state level of gene expression $S_1$ is

$$S_1(t) = \tau k_{12} \frac{S_2(t)}{S_2(t) + K_d}.$$  (3.8)

If TFs considered in the above example would be repressors the derivation of Eq. (3.8) would be completely analogous, but with $P_{12}$ replaced with $1 - P_{12}$; thus the resulting expression of the first gene would be $S_1(t) = \tau k_{12} K_d / (S_2(t) + K_d)$.

Unfortunately, due to nonlinearity of $g_i$ functions the rate equations as given in Eq. (3.5) cannot be solved analytically for large systems. But, in special cases qualitative properties of solutions can be established: the number of steady states, or limit cell cycles. Particularly, the relation between single feedback loops and the dynamics of the system can be investigated [74, 123]. It was shown that the negative feedback can lead to oscillations around a single steady state, whereas loops with positive feedback lead to bistability (or more generally multistability; depending on the initial conditions system ends up in one of many steady state attractors). Furthermore, this relation was actually strengthened by showing that the presence of negative/positive feedback loops is actually a necessary condition for the occurrence of cyclic behaviour/multistability.

Apart from predictions of dynamical properties of gene regulation, ODE models can provide some quantitative insight into regulatory processes. Early works, addressed systems with only a few genes, in order to demonstrate that having a set of cytoplasmic reactions along with reaction constants is sufficient to explain observed regulatory phenomenon. In [135] the oscillations in early embrionic cell cycle of *Xenopus laevis* oocyte were quantitatively explained, and results were used to discriminate between two competitive hypotheses regarding activation of one of the proteins. Further, in [136] the experimentally well-studied *cro-cI* switch that controls growth of the $\lambda$ phage in *E. coli* was investigated by testing its functionality with certain genes being mutated. For more pioneering examples see reviews [105, 137].

More recently, the high level of details for a system consisting of several genes, was obtained in [138] a set of ODE that modeled the cell cycle regulatory network for baker’s yeast. The model included in total 36 equations, and 148 constants; the equations describe changes in protein concentrations, cell mass, DNA mass, the state of the mitotic division of the emerging bud that is going to become a daughter cell. Some
3. GENE REGULATORY NETWORK MODELS

of the constants have been taken from the literature and adjusted by trial and error procedure, other fitted manually. After fitting its parameters the model accurately predicted the duration of different cell cycle phases along with gene expression profiles (measured by concentration of gene products). Furthermore, roughly 90% out of 131 simulated mutant strains have characteristics (viability, growth rate, size at different stages of reproduction process) consistent with experiments\(^1\).

3.3 Stochastic modelling

Describing regulatory dynamics with differential equations allows for prediction of gene expression levels with great detail, yet the number of implicit assumptions underlying differential formalism are no longer valid at the molecular level. Both assumptions on continuous concentration of gene products and deterministic character of regulatory dynamics become questionable when a low number of regulatory molecules is present in a cell [139, 140, 141]. Due to fluctuations in the number of molecules and timing of cellular events, a regulatory network can manifest different behaviours even starting with the same initial conditions [141, 142, 143]. Moreover, it was explicitly shown in single-cell experimental setups that transcription [101, 144, 145] and translation [144, 146, 147] processes show stochastic behaviour.

Hence, to account for possible fluctuations in the system stochastic models of gene expressions were proposed [139, 148]. The time evolution of the system can be written as a master equation. Under certain conditions (in practice the system needs to be well stirred with large number of molecules) master equation can be approximated by Langevin equation (see for instance [149]), where the system dynamics can be well described in terms of mean values of gene expression with additional Gaussian fluctuations (random force) around the mean. In order to calculate higher order moments one can expand the master equation to the second order, obtaining Fokker-Planck equation (for instance see a recent review by Tkačik [150], for a concise presentation of this approaches with illustrative examples).

\(^1\)The interested reader can visit Tyson’s lab web page mpf.biol.vt.edu/lab_website/ (checked on 5th June, 2013) where over one hundred references are given to works oriented on modelling regulatory mechanisms.
3.3 Stochastic modelling

3.3.1 Stochastic simulation algorithm

Since solving the master equation by analytical or numerical means is typically far more difficult than deterministic rate equations, an alternative approach that basically disregards master equation and resorts to direct simulations on molecular resolution was proposed by Gillespi under term stochastic simulation algorithm (SSA) [151]. Based on the initial number of molecular ingredients of the system (e.g. proteins, mRNAs) and reaction-probability constants introduced, SSA simulates dynamics of the system reaction by reaction.

The algorithm is exact for well-stirred systems, that is each molecule has equal chance of being anywhere in the considered system volume. Such a condition implies that for systems with diffusion [152, 153, 154] or transport [155] the results might be incorrect. Besides these limitations the algorithm has proved to be useful for describing gene expression dynamics in several small regulatory networks: stochastic description of Cro - CI switch in λ phage [156], transition between vegetation and competence in *B. subtilis* [157], influence of noise on the robustness of circadian clocks [158].
Chapter 4

Essential network framework

4.1 Genotype

Let us consider gene regulatory network consisting of \( N \) genes where the \( i \)th gene produces transcription factor molecules of the \( i \)th type. As a result, the influence of gene \( i \) on gene \( j \) can be represented as a link \( i \rightarrow j \), where the activatory or inhibitory effect is transmitted through TF molecule produced by gene \( i \). Notice that the inverse relation, \( i.e. \ i \leftarrow j \), is a completely different process in which the product of the \( j \)th gene plays a regulatory role. Apart from defining the connection \( i \rightarrow j \) between two genes, it is useful to quantify the strength of the regulatory interaction between them. This can be done by introducing weight \( W_{ij} \). Values of \( W_{ij} \) can be \textit{a priori} quite arbitrary, but it is convenient to assume that \( W_{ij} \in [0, 1] \): \( W_{ij} = 1 \) corresponds to a very strong binding, so a very strong influence of TF \( j \) on the transcription level of gene \( i \), and \( W_{ij} = 0 \) corresponds to no interaction at all, hence no regulatory effect. Further, the character of the interaction is stored as a sign: “+” for activatory, \( i.e. \) the one that increases the gene production rate, and “-” for inhibitory, \( i.e. \) the one that decreases (or blocks completely) the gene production rate. As a consequence we can represent the regulatory network by a directed weighted graph with \( N \times N \) adjacency matrix \( W \equiv \{W_{ij}\} \) with an additional distinction between activating and repressory edges\(^1\).

Because, we do not want to exclude any possible interactions between genes, we allow any of TFs to attach to any of \( N \) considered genes. We assume that each gene

\(^1\)This representation is similar to the one used in threshold models (\textit{c.f.} Sec. 3.1.3).
can produce one type of TF, and has $N$ regulatory sites (each dedicated to one type of TF). We represent each TF as well as each binding site by a character string of length $L$ with characters belonging to a 4 letter alphabet (see Fig. 4.1). However, one should be aware that although double stranded DNA is composed of four base pairs A-T, T-A, C-G, G-C (TF binds to a double stranded DNA, and all the four canonical base pairs are distinguishable), the transcription factor protein can be composed of a subset of the 20 standard amino acids specified by the genetic code. In order to make the model simple, we do not consider all the possible binding energies between DNA and TF, but instead we assume the TF binding domain has 4 types of binding elements (similar to DNA), such that each of these binding elements can create a complementary binding with only one DNA base pair. The list of these strings (DNA binding sites along with DNA binding domains) defines the molecular genotype, that can be further used to determine the weights of interactions.

Following the standard practice [93] (c.f. Sec. 2.3.1), we assume that the free energy of one TF molecule bound to its target site is, up to an additive constant, equal to $\varepsilon d_{ij}$, where $d_{ij}$ is the number of mismatches between the strings representing the $j$th TF and its binding site in the regulatory region of gene $i$. The parameter $\varepsilon$ is the penalty for each mismatch, that is the contribution to binding free energy in units of $k_B T$ where $k_B$ is the Boltzmann constant and $T$ is the temperature in degrees Kelvin.

As discussed in Sec. 2.3, $\varepsilon$ have values between one and three in $k_B T$ units [96, 99, 100]. We can define the interaction strengths $W_{ij}$ arising in such a regulatory network via Boltzmann factor

$$W_{ij} = \frac{e^{-\varepsilon d_{ij}}}{Z},$$

where $Z$ is a normalization constant (actually a partition function). Following [104] (c.f. discussion in Sec. 2.3.1) the term $Z$ in Eq. (4.1) can be in practice approximated by 1. This way we are able to go from the molecular genotype, which consists of $N$ TFs and $N^2$ binding sites each of length $L$ to the weight matrix $W$.

\footnote{In the simulation, rather than storing two character strings for each such interaction, we simply store the binary difference string (also of length $L$) specifying which entries match (1) and which mismatch (0).}
4.1 Genotype

Figure 4.1: Schematic representation of the model. Each gene regulatory region contains \( N \) binding sites, one for each of the \( N \) transcription factors produced by the \( N \) genes. The transcription factor \( TF \) produced by gene \( j \) binds to its different binding sites with interaction strengths \( W_{ij} \) that depend on the number of mismatches \( d_{ij} \) between the associated character strings (see Eq. (4.1)). A red cross represents a mismatch, and a green line corresponds to a match; for two random strings the probability to have a match is \( 1/4 \), hence a mismatch is three times more likely to occur.

4.1.1 Probability of occupation of the binding site

In Sec. 2.3.1 we have derived a formula (Eq. (2.14)) for the probability of the functional site being occupied by one of \( n \) transcription factors. Suppose now, that there are \( n_j \) TF molecules of type \( j \) that can attach specifically to the \( i \)th gene binding site. Assuming that TFs bind to their regulatory sites independently from each other we can write Eq. (2.14) separately for each TF-gene pair (\( Z \) from Eq. (4.1) is set to 1):

\[
P_{ij} = \frac{1}{1 + 1/(n_j W_{ij})} \tag{4.2}
\]

that can be interpreted as the probability of finding \( j \)th TF molecule attached to the regulatory site of gene \( i \). In order to avoid a proliferation of free parameters, we denote by \( n \) the maximum number of TF molecules that can arise when a gene is fully ON, taking for simplicity this maximum to be independent of \( j \). Although setting the maximum number of TF proteins to be \( j \)-independent is a rather strong assumption, it does not affect the qualitative properties of the model. Considering situation with \( n \) being dependent on \( j \) could be useful for incorporating external control factors. For instance, it is known that the probability of a CAP molecule to attach to DNA and recruit a polymerase depends not only on the number of these molecules but on the
4. ESSENTIAL NETWORK FRAMEWORK

concentration of glucose \([159]\). Thus, glucose (an external signal) moderates the number of CAP molecules (activators) without any feedback from the regulatory network.

Biologically, the parameter \(n\) is known to take values in a wide range, going from of order unity to many thousands \([101, 102, 103]\). In our studies, we consider several values of \(n\), reporting on the range \(10 \leq n \leq 10^4\).

4.2 Phenotype

As we previously explained the \(j\)th gene can produce \(n_j \in [0, n]\) TF molecules, thus it is useful to define the normalized level of gene \(j\) products as \(S_j = n_j/n\). Such a quantity can be identified with the gene expression level that allows for intermediate values. Specifically, we define \(S_j\) as a continuous variable that lies in the interval \([0, 1]\), with 0 corresponding to the lack of TF molecules of a given type and equal 1 to the maximum number equal to \(n\).

In comparison with the Boolean models in which a gene can be either expressed (ON) or not (OFF) this a conceptually different framework, which in terms of real-valued gene expression levels is closer to the ODE approach to regulatory network modelling. More importantly, such an approach can be justified from the biological perspective, and the argument goes as follows. At a given time point, a gene is either being transcribed by RNA polymerase or not. Hence, concerning the instantaneous production a gene either produces or not a mRNA that needs to be further translated into fully functional TF molecule. However, these TF molecules will then have a certain half-life, so they will be present in the cell for some period during which other TFs are still produced. The number of these TFs can be roughly determined from the gene average production rate. Thus, it is necessary to distinguish the “effective” or average level of expression of gene \(j\) (here the number of TF molecules of type \(j\)) from the instantaneous transcription rate. In a model like this one, only the effective expression level is relevant.

In order to track the effective gene expression level at any given time, let us define \(S_j(t)\) as an average gene expression level at time \(t\). The system under consideration has \(N\) genes, so the vector of effective expression levels

\[
S(t) = (S_1(t), S_2(t), \ldots, S_N(t))
\]

constitutes the system phenotype.

46
4.3 Transcriptional dynamics

The dynamics of the $S_j(t)$ takes place on biochemical time scales (typically minutes). To model the dynamics of regulatory network, we consider the level of the $i$th gene expression $S_i(t + 1)$ to strongly depend on the transcription rate of gene $i$ at time $t$. This assumption is supported by biological argument, that the transcription rate is strongly associated with the degree to which the gene regulatory region is occupied by TFs. Therefore, the higher the probability of activatory TF being bound specifically to the gene functional site, the higher the production rate of the gene. In order to easier explain the logic behind the system let us start with defining dynamics only for the activatory interactions, and afterward generalize it by introducing inhibitory regulation.

Let us assume that gene $i$ transcription is ON whenever at least one TF is bound within its regulatory region and otherwise it is OFF. This is a reasonable assumption, although it does not take into account possible cooperative effects. The (normalized) average expression level of a gene is then identified with the probability (up to an overall scale) that transcription is ON. Hence, for gene $i$ we have

$$S_i(t + 1) = 1 - \prod_j (1 - P_{ij}(t)),$$

(4.4)

where $P_{ij}(t)$ is given by Eq. (4.2)\(^1\) with the term $n_j$ replaced by $nS_j(t)$:

$$P_{ij}(t) = \frac{1}{1 + \exp(\varepsilon d_{ij} - \log(nS_j(t)))}.$$

(4.5)

The above holds for TF occupancies in the regulatory regions being statistically independent. The form of Eq. (4.4) can be explained in a straightforward manner: the term $\prod_j (1 - P_{ij}(t))$ corresponds to the probability that none of the TF molecules is attached to the regulatory region of gene $i$ (under assumption that the TF occupancies of functional sites being statistically independent), hence r.h.s. of Eq. (4.4) is the probability that at least one TF protein is attached to gene $i$. It is a reminiscence of the OR logical gate whereby the output is ON if and only if at least one of the inputs is ON. Further, in the limit of low values of $P_{ij}$ the transcription becomes additive in these variables (the terms of order $O(P_{ij}^2)$ and higher are neglectable).

\(^1\)Compared to Eq. (2.14) we redefined $\varepsilon$ to be in $k_B T$ units.
For inhibitory interactions we assume that the transcription factor molecule (repressor) of type \( j \) stops the transcription of gene \( i \) if it is bound to this gene functional site. As a result, the binding of at least one TF acts as a *veto*, significantly decreasing the gene production rate. Such a mechanisms, can be justified by a simple biological argument, that the bound TF blocks access of the polymerase to the gene promoter, preventing the transcription to occur. Another possible biologically motivated justification, is based on an observation that some of TF make the DNA form a loop that conceals the other binding sites.

Within our framework, transcription proceeds as in Eq. (4.4) in the absence of any repressors, but as soon as any of the inhibitors are bound in the gene regulatory region, transcription is stopped. Again assuming that there are no cooperative effects, and repeating for repressors the same argument as for activators, we are led to modify Eq. (4.4) to

\[
S_i(t + 1) = \left[ 1 - \prod_{j \in A} (1 - P_{ij}(t)) \right] \prod_{j' \in R} (1 - P_{ij'}(t)), \quad (4.6)
\]

where \( j \) runs over the set of activating interactions \( A \) and \( j' \) over inhibitory interactions \( R \); the \( P_{ij}(t) \) remains unchanged and is given by Eq. (4.5)). In the case of at least one repressor being attached to the functional site of the gene \( i \), one of the \( P_{ij'} \) would be close to one, so the whole inhibitory term would be close to zero, resulting in a very low value of \( S_i \). Due to the separation of the activatory and inhibitory influence in Eq. (4.6), the activators cannot compensate for inhibitors. If only there are inhibitory interactions with high enough \( P_{ij} \), the \( i \)th gene remains suppressed.

Formula (4.6) embodies a simple logic for each \( S_i \): (i) each activator or repressor acts independently; (ii) the binding of at least one activator is required for the activation of the target gene; (iii) the binding of even a single repressor is sufficient to prohibit the activation and to turn off the expression of the target gene.

In the context of discrete dynamical systems Eq. (4.6) defines a map \( S(t + 1) = G(S(t), W) \), where \( G \) is the global transition function and \( S(t) \) denotes the configuration of the system at time \( t \); \( S(0) \) is termed *initial phenotype*. This discrete dynamics can be represented by a sequence of steps

\[
S(0) \rightarrow S(1) \rightarrow \ldots \rightarrow S(T) \rightarrow S(T + 1) \rightarrow \text{limiting behaviour}
\]

(4.7)
4.3 Transcriptional dynamics

leading to an attractor which is either a cycle or a fixed point; $T$ is the time at which system reaches limiting behaviour. More importantly, since $G(S(t), W)$ is a deterministic function, the gene expression profile, i.e. the sequence of $S(t)$ for $t = 0, 1, \ldots, T$, is well defined and depends only on the genotype $W$ and the initial phenotype $S(0)$.

4.3.1 Functional constraints

Within our framework we consider three qualitatively different functional constraints (target conditions) that can be imposed on the system: (i) single or multiple fixed points phenotype, (ii) time-periodic (cyclic) gene expression, (iii) gene expression trajectory leading to a fixed point phenotype. In the case of a single fixed point condition, there need to exist such a time $T$ that for all gene expressions the phenotype $S(t)$ remains unchanged for time $t \geq T$. By imposing single target phenotype $S^{(\text{target})}$, as soon as the system reaches fixed point$^1$ the distance between $S^{(\text{target})}$ and $S(T)$ is measured via the following expression:

$$D(S^{(\text{target})}, S(T)) = \sum_{i=1}^{N} \left| S_i^{(\text{target})} - S_i(T) \right|.$$  \hspace{1cm} (4.8)

For the multiple fixed points imposed on the system dynamics, let us denote the number of such fixed points by $k_{FP}$. The definition of target phenotype is now extended to encapsulate all the imposed attracting states $S^{(\text{target})} := \{S_k^{(\text{target})}\}_{k=1}^{k_{FP}}$. The resulting distance between imposed multiple fixed points and the system steady states is calculated according to

$$D(S^{(\text{target})}, S) = \sum_{k=1}^{k_{FP}} \sum_{i=1}^{N} \left| (S_k)_i^{(\text{target})} - S_i(T_k) \right|,$$  \hspace{1cm} (4.9)

where $T_k$ is the time required to reach the $k$th fixed point starting from the initial phenotype $S_k(0)$. Concerning the choice of appropriate initial phenotypes, the easiest way to ensure that they fall to the given attractor basin is to set $S_k(0) = S_k^{(\text{target})}$ (see Fig. 4.2 for an illustrative example).

Concerning time-periodic expression pattern, the attractor is a cycle of phenotypes that are repeated in regular intervals. In general, a system initiated in $S(0)$ goes

$^1$If instead of the fixed point the system reaches a cyclic attractor, the condition for the fixed point is not fulfilled, hence no distance to $S^{(\text{target})}$ is measured.
4. ESSENTIAL NETWORK FRAMEWORK

Figure 4.2: A schematic representation of the distance calculation. Left: Steady state behaviour and $k_{FP} = 3$: crosses (heavy dots) stand for the target (fixed point) states, while the line is the system trajectory. The total distance (via Eq. (4.9)) is $D_T = D_1 + D_2 + D_3$. Right: Similar as before, but for a cycle. Grey dots stand for successive states obtained by iterating Eq. (4.10). Here $D_T = D_1 + \cdots + D_8$.

along a sequence of phenotypes, until the time $t_0$, when its evolution starts to repeat a cycle of period $T$. For the purpose of this argument, we can always shift $t_0$ to 0, and hence consider $S(0)$ as the cycle initial state. The imposed cyclic phenotype has the form of target gene expression profile $(S^{(target)}(0), S^{(target)}(1), \ldots, S^{(target)}(T))$ where $S^{(target)}(0)$ is equal to $S^{(target)}(T)$.

We define the distance between this target profile and the system gene expression as

$$D(S^{(target)}, S) = \sum_{t=0}^{T} \sum_{i=1}^{N} |S_i^{(target)}(t) - S(t)|.$$  \hspace{1cm} (4.10)

This definition can also be used for a non-cyclic path of length $T$. As for the gene expression trajectory leading to a fixed point phenotype, the distance is calculated according to Eq. (4.10) with an additional test for the last phenotype in the trajectory to be a fixed point of system dynamics.

Since we are dealing with real-valued expressions the condition for a phenotype to “remain unchanged” is equivalent to $\sum_{i=1}^{N} |S_i(t+1) - S_i(t)| < \delta$. The value of $\delta$ is set to be a small number corresponding to the required precision of numerical calculations (by default we use $\delta$ in the range $10^{-8}$ to $10^{-4}$; the results presented in the following chapters are not affected by changing $\delta$ in this range).
4.4 Genotype fitness

Having defined the distance between target phenotypes and genotype expression profiles, for different types of functional constraints, we would like to quantify how well a genotype is adapted to the imposed condition. Because $S$ is uniquely defined by the genotype and the initial phenotype\(^1\), we propose the fitness $F$ of the genotype $W$ to be defined through the fitness on the phenotypic level

$$F(W) = e^{-fD(S^{(target)}, S)},$$

(4.11)

where $f$ acts as a control parameter allowing one to be more or less stringent on the fidelity. As it might not be straightforward from the form of Eq. (4.11), the dependence of the r.h.s of this equation from $W$ is hidden in $S$ that is given by Eq. (4.6) (with dependence on $W$ in $P_{ij}$). As a consequence, in a set of all genotypes (with the same initial phenotypes) one can assign a relative weight $F(W)$ to each genotype. This forms an ensemble of genotypes, and by increasing $f$ we can focus on those genotypes that are the most functional. In our simulations we use $f = 20$; the results depend only very weakly on this choice provided $f$ is in the range 10 to 100. (These values of $f$ might seem quite large, but during simulations typical changes in the distance $D(S^{(target)}, S)$ are of $10^{-2}$ order, hence the product of these two is of order of unity.)

4.5 Viable genotypes

Given this definition of fitness, we would like to find genotypes that have high fitness. These genotypes produce gene expression patterns which are close to the predefined functional constraint.

In a naive way one could try to produce random genotypes by drawing randomly all letters defining $N^2$ DNA regulatory sites and $N$ DNA binding domains of length $L$, calculating interaction weights $W_{ij}$ via Eq. (4.1), and then the corresponding fitness, but such a procedure will almost always result in a very poorly adapted genotype. The reason is that the space of functional genotypes occupies only a tiny fraction of the space of all genotypes. The number of all possible $W_{ij}$ can be readily estimated as $(2(L + 1))^{N^2}$ (each $W_{ij}$ entry can take $L + 1$ values, corresponding to the number

\(^1\)Under deterministic transcriptional dynamics.
of mismatches $d_{ij} \in \{0,1,\ldots,L\}$, each entry can be either activatory or inhibitory, and there are $N^2$ such entries) which behaves roughly like $L^{N^2}$. Even for $L = 10$ and $N = 10$ we have $10^{100}$ possibilities, and if we consider only the presence of links neglecting the strength and types of interactions ($W_{ij} \in \{0,1\}$) we get $2^{100}$, which is still not a computationally feasible number. As it could be anticipated the number of functional genotypes is much smaller, yet to find them we need to introduce some effective sampling method.

Since the space of all genotypes is endowed with a fitness function, one can apply Markov Chain Monte Carlo (MCMC) techniques\textsuperscript{1}. The goal is to bypass genotypes with a low fitness and to focus on those whose phenotype is close to the target one. A standard approach is to take the fitness as a probability measure, so that each genotype appears in the ensemble of genotypes with the probability proportional to its fitness. We do this by using MCMC with the Metropolis rule \textsuperscript{160}: a computer algorithm produces a (biased) random walk in the fitness landscape that visits at long times different genotypes according to their fitness as given in Eq. (4.11).

At each step of the algorithm, we go from a genotype $W$ to a neighboring genotype $W'$ by introducing a single mutation: a random character in the regulatory sites defining the molecular genotype is transformed\textsuperscript{2} or the sign of an interaction is switched. Then the phenotype $S'$ and fitness of genotype $W'$ are determined. This modified genotype is accepted or not, based on its fitness value, according to the following Metropolis acceptance probability

$$P(W \rightarrow W') = \min\left(1, \frac{F(W')}{F(W)}\right)$$

(4.12)

that has a simple interpretation: if a mutation is beneficial (results in higher fitness) the modified genotype is always accepted, and if it is disadvantageous the new genotype is accepted with the probability given by the ratio of the new and old fitnesses. It is also easy to check that point mutations (single steps in the algorithm) allow for visiting all the possible genotypes.

After the initial relaxation time, the algorithm samples the genotype space with the specified measure, resulting in the ensemble of genotypes with high fitness values. Such

---

\textsuperscript{1}In general it is very difficult to find an inverse relation leading from a target phenotype to the associated genotype $W$. Hence, a sampling procedure is required.

\textsuperscript{2}In simulation it corresponds to transforming a match into a mismatch, and transforming a mismatch into a match with probability 1/3.
genotypes, due to their functional capabilities (producing approximately the target pattern), can be considered as viable, i.e. being able to fulfill imposed conditions. In biologically motivated scenario, a cell with a genotype that is unable to properly reproduce the required gene expression profile has typically a significantly decreased chance of survival.

4.6 Network of essential interactions

As already mentioned, genotypes can be represented by an N by N matrix of $W_{ij}$ interaction strengths along with the $N^2$ signs specifying the activating vs. inhibitory nature of each interaction. Since $W_{ij}$ are never zero (see Eq. (4.1)), one cannot say that an interaction is completely absent. Nevertheless, one may expect some interactions to be more important than others, for instance when the $W_{ij}$ are larger than average. An arbitrary cut-off could be introduced for separating small and large values, but it is better to base such a classification on functionality. We thus consider what happens when an interaction $W_{ij}$ is removed by setting it to zero.

Starting with one of viable genotypes, we determine the change in its fitness resulting from setting $W_{ij}$ to zero: if the change is rejected by the Metropolis rule in five successive attempts, we say that this interaction is essential, motivated by the corresponding biological definition. A very similar result is obtained by defining the essentiality as the sensitivity to a single mutation that increases the mismatch by 1.

Note that both these definitions include the Metropolis test Eq. (4.12), a random event, that in general can give different outcomes when repeated several times. Particularly, if the removal of a certain link $j \rightarrow i$ (putting $W_{ij}$ to zero) results in decreasing of genotype fitness by the factor $e^{-1}$, we can expect that by performing the test five times in a row, the probability to have false essential interaction is of order $e^{-5} \approx 0.007$. In practice this probability is much lower, since the reduction of fitness by a factor $e^{-1}$ corresponds to increasing the distance of output pattern from $S^{(\text{target})}$ by $1/f$, which results in a modification of gene expression by the factor $1/f \approx 0.05$ (assuming the default $f = 20$); typically the removal of an essential interaction results in switching OFF (or turning ON in the case of repressors) that leads to change $\Delta S \sim 1$ in gene expression. This change in expression results in a single drop of genotype fitness by $e^{-f}$, and as a consequence the probability of a false interaction when Metropolis test
is done five times in a row can be estimated as $e^{-5f}$ which number is negligible for the used $f$ values.

This definition leads to a summary description of a genotype through a list of ordered pairs $(i, j)$ specifying the essential interactions as well as their nature (activating or inhibitory). This list of pairs can then be represented by a directed graph, with the plus signs on the edges that are activating and the minus signs on the edges that are inhibitory. In all our work we refer to this oriented and signed graphs as to essential networks of the genotypes or as gene regulatory networks. Working with just essential networks provides a great deal of intuition about topologies and network features that are relevant for the system functionality. The price to pay for this simplicity is some loss of information: in principle, by assigning different interaction strengths ($W_{ij}$) the same network structure can lead to different target phenotypes. As a result, for some network structures it might be difficult to figure out the network behaviour only from the topology of essential interactions.

### 4.7 Model limitations

One can also ask whether our model is structurally robust. Any real expression dynamics includes fluctuations in the number of TF molecules [102, 161, 162] whereas our $S_i(t)$ dynamics can be considered as a mean-field model that neglects fluctuations. (On the right-hand side of Eq. (4.6) we have probabilities and on the left-hand side there is the normalized level of gene products.) In general such an approximation can be justified when the number of molecules involved is large, \textit{i.e.}, when $n$ is large. In practice it is \textit{a priori} not clear what is “large enough”. We checked that our model gives qualitatively the same results in the biologically relevant range of $n$ from 10 to 10000.

In order to reproduce experimental gene expression patterns we need to introduce a phenomenological modification. Without this modification the sampling procedure could not find a viable genotype with expression pattern fully matching the experimentally derived expression profile. The modification imposes cut-off on expression levels corresponding to single TF molecules. This model generalisation is explained in Sec. 6.1 at the beginning of Chapter 6. Still, all the results from Chapter 5 can be reproduced within the extended model.
Chapter 5

Results for idealized gene expression patterns

In this chapter we study regulatory networks with idealized gene expression patterns (phenotypes) imposed as functional constraints. We concentrate on two qualitatively distinct classes of constraints that were inspired by cell’s biology: (i) network dynamics that has single or multiple fixed points resembling process of cell differentiation, or (ii) a time-periodic (cyclic) phenotype as in the case of cell-division cycle. Importantly these two classes are also the only two possible limiting behaviours for a discrete system governed by deterministic dynamics. It is thus interesting to ask: how does each of these two types of general constraints affect the regulatory network topology? Specifically, within our modelling scheme we can impose idealized (i.e. obeying different symmetries) fixed point or cyclic phenotypes on regulatory networks. By doing so we identify which topological properties result from the selection under a particular functional constraint.

Concerning the multiple fixed points constraint, we investigate separately the case with one fixed point imposed on the regulatory network [19] and the case with regulatory network constrained to have two, three or four predefined fixed points [20]. Although, both these cases show similarities (e.g. interaction sparsity, narrow in-degree distribution, heterogeneous robustness) there are also important differences (e.g. lack of repressors or network motifs for single fixed point constraint). Furthermore, in the case of the single fixed point condition we give analytic justifications for some of the observed phenomena, namely: the emergence of viable networks, the distribution of
5. RESULTS FOR IDEALIZED GENE EXPRESSION PATTERNS

mismatches in viable genotypes or estimates of mutational robustness. In the case of multiple fixed points it is difficult to provide analytical results, so numerical estimates are given.

Results for regulatory networks under selection towards cyclic phenotypes are presented in the same section as for multiple fixed points. This is mostly to illustrate the dissimilarities between these two cases: especially for the network motif statistics, where differences are of great importance [20].

Throughout this and the next chapter, if not stated otherwise, we assume transcription factor (TF) encodings (strings representing DNA binding domains) to be given and fixed. Hence mutations are introduced only in sequences representing DNA binding sites. The justification for this comes from the fact that TFs are typically pleiotropic (they regulate several genes), so they are subject to strong stabilizing selection (though this is not always the case, c.f. discussion in Sec. 2.2.3). It is easier to permanently modify binding domain/site which interacts with only one gene, rather than with several genes. Thus TFs are generally thought to evolve slowly, while DNA regulatory regions typically have a high level of polymorphism and evolve more quickly [163].

5.1 Single target expression

In the case of a single target expression pattern, we sample a set of all viable genotypes leading from $S^{\text{initial}}$ to $S^{\text{target}}$. This choice [128] is motivated from cases arising in early embryo development, where the initial expression levels are inherited during the formation of the egg. A network performs the desired function if and only if, for a given initial phenotype the vector of gene expressions converges to a fixed point (steady state) that is close\textsuperscript{1} to the target pattern.

5.1.1 Target pattern specification

For simplicity, we fix the $S_i^{\text{initial}}$ and $S_i^{\text{target}}$ for $i = 1, \ldots, N$ to be 1 (ON) or 0 (OFF). If the initial and target phenotype are drawn at random, the number of components set to 1 will be approximately equal to that set to 0 for large $N$. Hence, to reduce finite size effects in $N$, these numbers are set exactly to $N/2$. Without loss of generality, to accommodate for the permutation symmetry of the model we work

\textsuperscript{1}In terms of distance defined in Eq. (4.8).
with \( S^{(\text{initial})}_i = 1 \) for \( i \leq N/2 \) and 0 otherwise. We also impose \( S^{(\text{target})}_i = 1 \) for \( N/4 < i \leq 3N/4 \) and 0 otherwise. Notice that \( \sum_i S^{(\text{initial})}_i = \sum_i S^{(\text{target})}_i = N/2 \) and \( \sum_i S^{(\text{initial})}_i S^{(\text{target})}_i = N/4 \).

For \( N = 20 \) genes the resulting phenotypes are:

\[
S^{(\text{initial})} = \begin{array}{cccccccccccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array} \quad (5.1)
\]

\[
S^{(\text{target})} = \begin{array}{cccccccccccccccc}
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0
\end{array} \quad (5.2)
\]

Using the permutation symmetry, it is easy to check that the model properties depend not on the positions of bits in the initial and target patterns, but only on the number of components of \((S^{(\text{initial})}, S^{(\text{target})})\) falling in each class \((0,0), (0,1), (1,0)\) and \((1,1)\). (We have examined how the network connectivity is affected when we take \(S^{(\text{target})}\) to have \(N/4\) or \(3N/4\) components set to 1 instead of the \(N/2\), and found that the same qualitative properties emerge.) Both the initial and target patterns are idealized, \textit{i.e.} in the biological setup they would have been determined by evolution to enhance survival prospects.

### 5.1.2 Emergence of viable genotype

The first problem we address is the emergence of a viable genotype starting from an entirely random genotype. Although it is difficult to derive properties of viable genotypes, general genotypes (or random genotypes) are relatively easy to understand because they have no functional constraints imposed. The statistical properties of the interaction strengths \(W_{ij}\) follow from the distribution of the mismatch between the character strings for DNA binding sites and TF binding domains. Each of the \(L\) characters of a binding site gives a mismatch with the probability \(3/4\) (there is only one complementary binding in four possible binding options). As a consequence, the total number of mismatches \(d\) is governed by the binomial distribution with mean \(3L/4\):

\[
p(d) = \binom{L}{d} \left( \frac{1}{4} \right)^{L-d} \left( \frac{3}{4} \right)^d . \quad (5.3)
\]

To the biologically realistic values of \(L\) (from 10 to 15), using for instance Stirling’s formula, we have

\[
p(d) \sim \frac{3^d L^d}{d! 4^L} \ll 1 \quad \text{when} \quad d \ll L . \quad (5.4)
\]
5. RESULTS FOR IDEALIZED GENE EXPRESSION PATTERNS

Figure 5.1: Shown is the distance between stationary expression pattern \( S \) and \( S^{(\text{target})} \) for two trajectories representing the emergence of viable genotype (functional network) from a random genotype.

Hence, the above binomial distribution gives a very low probability of observing a small number of mismatches. According to Eq. (4.1) small \( d \) corresponds to very weak interactions, hence unfit genotypes. Typically, these random genotypes lead to trivial target pattern which is composed only from zeros (see Fig. 5.1). If the system is evolved under a selection pressure on the fitness, then it is possible that the genotypes will undergo a gradual improvement from a very poor (unfit) to viable (highly fit) genotypes.

By default we start in a random genotype and do a series of single point mutations, that successively alter the underlying genotype. In effect, we examine the change in the fitness with increasing number of mutations. The result of this simulation is shown in Fig. 5.1, where we plot the distance \( D(S, S^{(\text{target})}) \) versus the number of mutational sweeps for two typical realizations (the sweep is defined as a succession of \( LN^2 \) attempted modifications).

It is instructive to follow such a trajectory in more detail. Since we do not observe repressors for the single target expression pattern (we will explain it later on, why there are no repressor), let us consider dynamics given by Eq. (4.4), in which \( P_{ij}(t) \)
5.1 Single target expression

from Eq. (4.5) can be rewritten as

$$P_{ij}(t) = S_j(t)/(S_j(t) + x_{ij}) ,$$

(5.5)

where $x_{ij} = 1/(W_{ij}n)$. For a four letter code, assuming that the binding energy is additive in mismatches and that every mismatch costs the same, one gets the average value of $W_{ij}$ for a random system to be

$$\langle W_{ij} \rangle = (0.25 + 0.75 e^{-\varepsilon})^L .$$

(5.6)

When $L = 12$ and $\varepsilon = 2$ this quantity is fairly small, about $3.6 \times 10^{-6}$. Hence, for $n < 10^4$, all $x_{ij}$ have large values leading to $P_{ij}$ of order (roughly) $n/10^6$. Consequently, for the random (initial) genotypes, the dynamical Eq. (4.4) yields $S_i(t) \approx 0$ after only a few steps. However the target pattern $S^{(\text{target})}$, by construction, has $N/2$ elements equal to 1 and all other being 0, so the distance $D$ between these two phenotypes is approximately equal to $N/2$.

Such a situation may persist for a rather long time, until as a result of continuing mutations an unlikely event occurs: a diagonal element of the array $x_{ij}$, say $x_{kk}$, becomes small enough by chance. Then, the dynamical equations for the $k$th gene expression reduces to

$$S_k(t+1) \approx S_k(t)/(S_k(t) + x_{kk}) ,$$

(5.7)

with a non-trivial stationary solution $S_k = 1 - x_{kk}$. The domain restriction for $S_k$ being non-negative implies that $x_{kk} = \exp(\varepsilon d_{kk})/n \leq 1$, thus $d_{kk}$ is not greater than an integer threshold value

$$d_n = \lfloor \log n/\varepsilon \rfloor$$

(5.8)

that we call critical mismatch hereafter.

Particularly, any mismatch, let us call it subcritical, that is less or equal to $d_n$ corresponds to a strong interaction $W_{ij}$. For the used values of $n$ one can easily derive the corresponding values of subcritical mismatches, and also from Eq.(5.3) the probability of emerging subcritical mismatch on the diagonal of the genotype (see Tab. 5.1).

Equivalently, the condition for the interaction $W_{ij}$ to be strong\(^1\) can be derived from Eq. (4.5) by imposing $P_{ij} > 1/2$ for $S_j = 1$.

---

\(^1\)For the single fixed point target phenotype a strong interaction is also an essential interaction as defined in Sec. (4.6). In more complicated cases with multiple fixed points or cyclic phenotypes not all essential interactions are strong.
5. RESULTS FOR IDEALIZED GENE EXPRESSION PATTERNS

<table>
<thead>
<tr>
<th>$n$</th>
<th>$d_n$</th>
<th>$p(d_n) \times 10$</th>
<th>$S_k$ : $d_{kk} = 4$</th>
<th>$d_{kk} = 3$</th>
<th>$d_{kk} = 2$</th>
<th>$d_{kk} = 1$</th>
<th>$d_{kk} = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>4</td>
<td>$2.4 \times 10^{-2}$</td>
<td>0.7019</td>
<td>0.9597</td>
<td>0.9945</td>
<td>0.9993</td>
<td>0.9999</td>
</tr>
<tr>
<td>1000</td>
<td>3</td>
<td>$3.5 \times 10^{-3}$</td>
<td>0.5966</td>
<td>0.9454</td>
<td>0.9926</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>$3.5 \times 10^{-4}$</td>
<td>0.454</td>
<td>0.9261</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>$2.1 \times 10^{-5}$</td>
<td></td>
<td>0.2611</td>
<td>0.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: For the range of $n$ values (1st column) presented are: values of critical mismatch $d_n$ (2nd column), probability of finding $d_n$ in one in ten positions ($= N/2$ positions that correspond to $S_k = 1$ in the target pattern) on the diagonal that results in the $k$th gene activation (3rd column), the $k$th gene expression $S_k$ for subcritical mismatches (from 4th to 8th column).

When a strong interaction appears on the genotype diagonal element $W_{kk}$ with $k$ such that $S_k^{(target)} = 1$, the distance $D$ to the target phenotype drops by an amount of order unity ($D$ is reduced by $S_k$ from Tab. 5.1). A subsequent significant increase of this distance is highly improbable because of the selection pressure: one unit of increase is counter-selected by a factor $\exp(-f)$.

As the process continues with additional mutations, the increase in fitness dramatically accelerates. The presence of every essential interaction has a double effect: it renders more probable the formation of new essential interactions which no longer need to be located on the diagonal, and it tends to lift up almost simultaneously several $S_i$. The distance to the target phenotype then rapidly drops until the system reaches a regime where $D$ fluctuates around a plateau value. For the choice of parameter values used in Fig. 5.1, this leads to $S_i \approx 0.94$ for each $i$ satisfying $S_i^{(target)} = 1$, while the remaining expression levels are negligibly small.

Hence, within our framework the system can evolve from a completely random genotype towards a well-adapted regulatory network under a realistic selection pressure if given enough time. Furthermore, when the system reaches regime in which genotypes are viable the evolution of genotypes does not stop, but instead the space of functional genotypes is sampled through a random walk. Because of this evolvability the large autoregulation $W_{kk}$ that has appeared as the first noticeable evolutionary event is not further necessary. The system already reached evolutionary plateau. We verified this expectation by measuring the fraction of large auto-regulatory interactions at long times (when the Monte Carlo provides equilibrium samples of viable), and found that
the fraction of auto-regulatory interactions is no larger than expected by chance.

### 5.1.3 Sparsity of essential interactions

Monte Carlo evolutionary dynamics with the Metropolis rule described in Sec. 4.5 generates at large times an equilibrium distribution of fit (viable) genotypes. Viable genotypes define in essence a null hypothesis since this space determines the distribution of any gene regulatory network property arising from the sole constraint of function, \( i.e., \), the only constraint imposed on the regulatory network is they to be fit. We will now see that the vast majority of these gene regulatory networks are in fact sparse, a feature that does not at all arise in regulatory network models that ignore the microscopic origin of the \( W_{ij} \).

For our computational study, as suggested in Fig. 5.1, it is best to perform measurements every few hundred sweeps so that the data collected are not correlated. We typically used 10000 measurements separated by 100 sweeps. In Fig. 5.2 we show the distribution of the mismatch \( d \) when there are \( N = 20 \) genes, for decreasing values of the number of TF molecules \( n \). At large values of \( d \) we observe a binomial distribution with a peak near \( d = 3L/4 \) as expected from Eq. (5.3); this bulk represents all the large mismatches that correspond to weak interactions. But for small \( d \) one observes a significant deviation from the binomial distribution that corresponds to strong interactions. In fact as \( n \) decreases the viability constraint becomes more selective and the distribution becomes bimodal: a second peak appears at low mismatch values.

Note that this peak shifts as \( n \) decreases, indicating that there are nearly perfect matches that appear in that regime. Indeed the position of this peak could be anticipated from Tab. 5.1 in which the critical mismatch values \( d_n \) along with gene expression levels are listed. We see that the peak is centered at \( d_n - 1 \) mismatch for different \( n \) values. The reason for the peak not to be centered exactly at \( d_n \) can be understood from Tab. 5.1 where there are given expression levels \( S_k \) for \( d_n \) and \( d_n - 1 \). The position of this peak is settled by mutation-selection balance: mutations increase mismatch values, but at the same time smaller mismatches give higher fitness levels favoured by selection.

If one significantly increases the \( f \) parameter (at least by an order of magnitude) that tunes the strength of selection in our model, one would get the peak placed at \( d_n - 2 \) (for \( d_n \geq 2 \)). It may also happen that for intermediate values of \( n \) that the peak
5. RESULTS FOR IDEALIZED GENE EXPRESSION PATTERNS

Figure 5.2: Distribution of the Hamming distance between a TF and the receiving DNA site for \( N = 20, L = 12, \varepsilon = 2.0 \) and for various values of \( n \). Random case corresponds to binomial distribution. The lines are to guide the eye.

at small \( d \) extends over two neighbour values of \( d \); the essential interactions then do not necessarily have the same strength. Hence, even under perturbation of the fitness strength and the number of TFs the overall picture remains the same.

In Fig. 5.3 we show an illustrative case of the mismatches for a toy viable network with \( N = 8, \varepsilon = 2.0 \) and \( n = 1000 \). The bimodal nature of the mismatch distribution is clearly apparent from this figure: the “pale” background corresponds to weak interactions whereas there are only a few “dark” plaquettes that represent large \( d \) values. Recalling that \( W_{ij} = \exp(-\varepsilon d_{ij}) \), the interaction matrix can be considered to be sparse if we focus only on strong interactions. However doing so requires defining rather arbitrarily a distinction between strong and not strong \( W_{ij} \). This difficulty is inherent to our framework where \( W_{ij} \) is never 0, in contrast to the situation in other gene network models where an interaction is present or absent. By using the concept of essential interactions we gain a mathematical notion of sparsity in our context without too much of arbitrariness. With this definition, one can ask whether viable genotypes have sparse essential interactions, and in particular whether there are few essential interactions per row in the genotype matrix \( W \).

For a viable genotype the number of active rows (i.e. having not too small expression
levels) must be equal to the number of genes that are ON in the target phenotype, otherwise the fitness is far too low. We find that as soon as \( n \) is not too large, there is almost always just one essential interaction per row as shown in Fig. 5.4 for \( N = 20 \) and \( L = 12 \). The same result holds for other relevant values of \( N, n \) and \( L \), suggesting that within our model, the drive towards sparse interactions arises in regimes of biological relevance.

As a consequence, the mismatch distribution for an active row can be approximated by the following ansatz:

\[
\tilde{p}(d) \approx \frac{N - 1}{N} p(d) + \frac{1}{N} \frac{p(d) \Theta(d_n - d)}{\sum_{m<d_n} p(m)}, \tag{5.9}
\]

where \( \Theta(x) \) is a heavyside step function with \( \Theta(x) = 1 \) for \( x > 0 \) and 0 otherwise, and \( p(d) \) is given by Eq. (5.3). The above formula has a simple interpretation: the first term corresponds to the distribution of mismatches in the genotype background (weak interactions; there are on average \( N - 1 \) non-essential links per one gene), and the second term is the renormalized binomial probability resulting in a narrow peak with maximum at \( d_n - 1 \). There is thus one “leading” mismatch taking care of most

---

**Figure 5.3:** Left: Genotype with associated weight matrix \( W \) for \( N = 8 \), \( L = 12 \), \( \varepsilon = 2.0 \) and \( n = 1000 \). The dark (brownish) squares represent strong interactions. The target phenotype for this genotype has all eight genes turned ON. Right: Gene regulatory network extracted from this genotype with only essential interactions indicated.
5. RESULTS FOR IDEALIZED GENE EXPRESSION PATTERNS

Figure 5.4: Left: Probability distribution of the number of essential interactions per row of the matrix specifying a viable network for $N = 20, L = 12, \varepsilon = 2.0$ and a range of values of $n$. Right: The same statistics can be also interpreted as the in-degree distribution. Data is presented for $n = 10000$ with the dashed line being an exponential fit.

of the viability constraint, while the other mismatches behave approximately as if they were unconstrained. Interestingly, qualitatively the same distribution is observed in studies of transcription factors binding energies [164] where the peak at low energy values corresponds to good matching of TF to DNA target sequence.

5.1.3.1 Narrow in-degree distribution

One can also interpret the normalized histogram of essential interactions per row as the in-degree distribution for the associated gene regulatory network. In Fig. 5.4 we see that the in-degree distribution of interactions between genes is very narrow. From the right part of Fig. 5.4 where the distribution of ingoing links for $n = 10000$ is presented in the log-linear scale, we get that the essential connections qualitatively follow the exponential distribution.

5.1.4 Mutational robustness heterogeneity

In the context of biological systems, a natural question to address is: how do such systems respond to mutations? Or in other words what is the system robustness against mutations. In our framework, the answer is easily derived: most of the mutations do not affect the genotype viability, but a few of them can be deleterious. Specifically, all the mutations that take place in the genotype background (in non-essential interactions)
are found to be innocuous, and the ones that occur within DNA binding sites related to essential interactions are typically deleterious. This concept is closely related to our notion of interaction essentiality: instead of testing the viability constraint against complete removal interaction (putting $W_{ij} = 0$), we ask that viability be lost when the number of mismatches associated with the interaction is increased by one (modifying one of the characters that codes for $d_{ij}$).

As a result of testing genotype robustness against point mutations we found that the rule one essential incoming interaction per gene generally hold here too. More precisely, the average robustness of genotype with respect to binding site mutations, $R_{bs}$, can be estimated as a fraction of non-sensitive interactions ($N^2 - N/2$) in the set of all $N^2$ interactions, hence one expects $R_{bs} \approx 1 - 1/(2N)$. For $N = 20$ and the single target pattern with 10 active genes, the expected value of $R_{bs}$ is 0.975, and indeed we find $R_{bs} = 0.977(2)$ (with some weak dependence on $n$ in the third decimal). Thus vast majority of mutations taking place in DNA binding sites have no consequence on the fitness, while mutations in the essential interactions are typically deleterious.

### 5.1.5 Broad out-degree distribution

Up to now we have investigated the distribution of interactions among the rows of genotype $W$. It is equally interesting to ask how the essential interactions are distributed among the columns of genotype. As the former corresponds to the in-degree distribution the latter is the out-degree distribution for a given gene. First of all, in our example (single target pattern) the essential interactions appear only in $N/2$ columns. This is an immediate consequence of the target phenotype being a fixed point for the genotype dynamics. Since, in the final state of the dynamics none of these $N/2$ columns is preferred over others, the essential interactions are distributed completely randomly among the $N/2$ active columns. Hence, the distribution of essential interactions among columns is binomial with the probability parameter equal to $2/N$.

More generally, if $K_{tot}$ is the total number of essential interactions, the viable genotype will typically have $K_{tot}$ equal to the number of active genes in $S^{(target)}$ (leading to $K_{tot} = N/2$). These essential interactions define a regulatory network with in-degree almost always equal to 1 while the out-degree has a binomial distribution that is well approximated by a Poisson law with mean 1 ($= K_{tot} \times 2/N$).
Up to now we assumed that the strings of characters associated with TFs were fixed, and the mutations were only introduced in DNA binding sites. Of course, the genes coding for TFs can mutate. These mutations have very specific consequences for the genotype: when the $j$th TF is modified, the whole $j$th column of the matrix $\{W_{ij}\}$ is affected at once. Hence, one can expect a selective pressure on regulatory networks that will affect the out-degree distribution. The selective pressure can be understood on the level of population dynamics: if we consider two regulatory networks with different robustness against mutations in TF coding sequences, $R_{tf}$, the network with higher robustness would have a higher chance of surviving in the population of networks exposed to mutations.

Qualitatively, one expects broad distributions to be favoured; indeed, consider mutating a TF coding gene. If the TF is associated with at least one essential interaction, then such a mutation is typically highly deleterious. Oppositely, if the TF has no essential interactions, the mutations will be harmless since no essential interactions are affected. Intuitively, broad distributions should be favoured as few of the TF mutations affect essential interactions. In an extreme situation, when one wants to maximize the $R_{tf}$ without losing viability, the perfect solution would consist of a genotype with all the essential interactions in a single column (only mutation in the TF associated with this column would be deleterious). Hence, in the presence of TF mutations, there is a strong correlation between the robustness $R_{tf}$ to mutations in different TFs and the column occupancy by essential interactions. Furthermore, it is useful to quantify the column occupancy by the average out-degree of essential network, denoted by $\langle k_{out} \rangle$.

Particularly, if $N_{occupied}$ is the number of columns with essential interactions, the $\langle k_{out} \rangle = K_{tot}/N_{occupied}$. Since mutating essential interactions is highly deleterious, the average robustness with respect to mutating a TF coding gene is accurately given by $R_{tf} = N_{empty}/N$, where $N_{empty} = N - N_{occupied}$ is the number of columns in a genotype without essential interactions. After elementary algebra one gets

$$R_{tf} = 1 - \frac{K_{tot}}{N \langle k_{out} \rangle}.$$  \hspace{1cm} (5.10)

For target phenotype defined in Eq. (5.2) total number of interactions is equal to $N/2$, and therefore

$$R_{tf} = 1 - \frac{1}{2 \langle k_{out} \rangle},$$  \hspace{1cm} (5.11)
a result fully confirmed by an explicit numerical simulation. Thus, gene regulatory
network with smaller $k_{out}$ are disfavoured as soon as we allow TFs to mutate. Provided
the population size is not too small, the distribution of the out-degree may depart
strongly from the Poisson distribution found in the absence of TF mutations.

In order to study the effects of selection pressure on $R_{tj}$ one needs to consider a
population of genotypes. By simulating many viable genotypes the networks with broad
out-degree distributions should outperform those with more narrow distributions. Large
scale simulations of such a population would provide the distribution of $k_{out}$. Since, it
is computationally not feasible to carry out such a study within our full framework, we
propose an effective model for genotypes (see [19]), where the evolution of population
composed of effective genotypes is modelled with the Moran process [165].

5.1.6 Evolvability of genotypes

A subject that is closely linked to effects of mutations on a biological system, is the
ability of the system to evolve under mutations. After some initial relaxation period the
genotype reaches high fitness levels. It keeps on evolving. The mechanism responsible
for this evolution is a long succession of binding site mutations. These mutations
move an essential interaction from one column to another. A few hundred sweeps
are sufficient for that. Measuring the moments of the out-degree distribution one can
calculate the autocorrelation time of this process. It turns out to be roughly of the
order of $10^3$ sweeps for $n = 100$ and about $10^2$ sweeps for $n = 1000$. This implies that
very different genotypes can arise under mutation-selection balance if given enough
time; our gene regulatory network model thus allows for gradual change of genotypes
while maintaining phenotypes, a characteristics of evolvability. We elaborate more on
the evolvability and structure of viable genotype space for more sophisticated target
phenotypes (see Sec. 5.2.4; for a single fixed point the situation is somewhat trivial).

5.1.7 Biological range of parameters

This overall picture corresponds to having all factors $P_{ij}$ in Eq. (4.4) be rather small
except for the one which is associated with the incoming essential interaction.
5. RESULTS FOR IDEALIZED GENE EXPRESSION PATTERNS

Figure 5.5: The curves show the lower limit of the region where $h \equiv 1/n = 10N[0.25 + 0.75\exp(-\varepsilon)]^L$ for $N = 20$; this region depicts the relevance domain of our framework. Rectangle corresponds to experimental range of parameters.

Assuming all $P_{ij} \ll 1$ (from Eq. (4.2) it is equivalent to $W_{ij}n_j \ll 1$ for $j \neq k$ and $P_{ik} \approx 1$, one can expand Eq. (4.4)\(^1\) to

$$S_i \approx P_{ik} + \sum_{j \neq k} W_{ij}n_j$$  \hspace{1cm} (5.12)

by neglecting quadratic terms in $P_{ij}$. Hence, for a strong essential interaction not to be overwhelmed by many weak (random) interactions we have a condition

$$1/n \gg N \langle W_{ij} \rangle = N(0.25 + 0.75e^{-\varepsilon})^L$$  \hspace{1cm} (5.13)

where we set all $n_j = n$. Further, we replaced $W_{ij}$ with its expected value provided that DNA target sites have a four letter code, the binding energy is additive in mismatches and that every mismatch costs $\varepsilon$.

Within a two letter code, the condition forces one to larger values of $L$ (close to 20) and thus beyond what is realistic biologically. According to Eq. (4.2) the probability that a TF occupies a DNA site is controlled by $n$, which is the number of TF molecules. We considered the range $10 < n < 10000$. Interpreting the sign $\gg$ in Eq. (5.13) as larger by one order of magnitude, that is by a factor 10, one gets an allowed region in parameter space as illustrated in Fig. 5.5.

This figure can be used to obtain the predicted domain of relevance: it is above the corresponding curves (taken at $N = 20$ and illustrative values of $n$). We see that $\varepsilon$ and

\(^1\)As we consider a steady state there is no time dependence.
5.2 Multistability and cyclic conditions

5.2.1 Target phenotype specification

5.2.1.1 Multiple fixed points

In the previous section we studied consequences of the single fixed point constraint imposed on the genotypes evolution. Now, we impose more than one fixed point to be a stable phenotype. Although the change might seem only quantitative, the resulting genotypes have a significantly different structure, that is the repressors are necessary for fulfilling viability condition, which was not the case for the single target phenotype. The choice of 2, 3, or more target phenotypes is motivated by the process of cell differentiation during organisms development; each target phenotype is associated with a different tissue. These type of phenotypes involving up to a dozen of genes have been reported in various organisms [126, 167]. Here, we are going to use idealized target phenotypes which, similar to the single fixed point phenotype, consist of $N/2$ genes being OFF ($S_i = 0$) and $N/2$ genes ON ($S_i = 1$). We set these target phenotypes to be orthogonal (for the 0/1 coding for $S_i$ this means that the scalar product of target expression vectors is $N/4$).

For $N = 16$, we define four mutually orthogonal target phenotypes as follows:

$$S_1^{\text{(target)}} = 1111111100000000$$
$$S_2^{\text{(target)}} = 1111000011110000$$
$$S_3^{\text{(target)}} = 1100110011001100$$
$$S_4^{\text{(target)}} = 1010101010101010$$
This choice is justified by the fact that at large $N$, random binary vectors are typically nearly orthogonal. Moreover, these target phenotypes can be transformed one into another by an index permutation. Due to this symmetry the attraction basins leading to these target phenotypes are on average of equal size. Concerning the choice of $N$ it is not important as long as it has a moderate value (we have not investigated what happens at large $N$).

As for the initial states there is not so much freedom left, as it was in the case of single target phenotype. Because we use deterministic dynamics, there are at least as many initial phenotypes required as there are imposed target fixed points. In order to provide initial conditions that guarantee the convergence to corresponding target pattern, we make the simplest possible construction by setting $S_k^{(\text{initial})} = S_k^{(\text{target})}$ where $k$ runs from 1 to the number of fixed points. In the MCMC simulation, for a given genotype we separately check convergence to imposed targets for every initial phenotype. The distances resulting from the difference between reached fixed point and the target phenotype are then added (see Fig. 4.2); the total distance calculated according to Eq. (4.9) is then used to get the genotype fitness.

As a last point, there is a small technicality that one should be aware of: in the initial phenotypes, instead of null values we use a very small non-zero value (e.g. $10^{-3}$). This helps to avoid genotypes for which the desired attractors are actually not stable fixed points. To give some intuition, this problem arises due to the existence of a degenerated solutions of our dynamical system: activatory interactions are present only on the genotype diagonal, and all the other entries in $W$ are inhibitory interactions; such a peculiar solution disentangles activatory interactions on the diagonal, resulting in a genotype behaving like the identity operator. Hence, all the genes that are initially turned ON/OFF remain ON/OFF after a long time. However, it is enough to only slightly perturb the expression of an initially inactive gene to end up in a fixed point phenotype with that particular gene being activated. As we show in the next chapter (see Sec. 6.1) this instability has actually a deeper meaning (it is outside of the applicability domain of the mean-field approach that we use), and we also give a “cure” to overcome this unwanted behaviour.
5.2 Multistability and cyclic conditions

5.2.1.2 Cyclic gene expression

For the cyclic target phenotypes the gene expression is periodic in time. Thus, the cyclic expression pattern is qualitatively different than the multiple fixed point condition. The biological motivation for periodic patterns comes from the cell division cycle, during which cell divides into two daughter cells. Such cases of cycling patterns are available for various organisms for which the gene expression levels where acquired experimentally (see for instance two yeast species [25, 26] and mammalian cell cycle [27]). Typically, the structure of such a cyclic pattern can be represented by a block of active genes which are shifted with each time step. As a consequence, we propose an idealized cell cycle where the genes ($N = 16$) are taken to lie on a ring:

\[
\begin{align*}
S_{\text{target}}(0) &= 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 \\
S_{\text{target}}(1) &= 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 \\
S_{\text{target}}(2) &= 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 \\
S_{\text{target}}(3) &= 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 \\
S_{\text{target}}(4) &= 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 \\
S_{\text{target}}(5) &= 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 \\
S_{\text{target}}(6) &= 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 \\
S_{\text{target}}(7) &= 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 \\
S_{\text{target}}(T = 8) &= 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
\end{align*}
\]

This type of functional constraint requires the gene expressions to follow tightly step by step the target pattern. Particularly, the initial phenotype is equal to the target phenotype at time $t = 0$ and system is forced to do the full cycle at least once. For the MCMC we just calculate the distance between the genotype output pattern and the imposed target sequence by using Eq. (4.10) (see Fig. 4.2 for schematic representation). The resulting distance is then used to calculate genotype fitness, and after a long sequence of consecutive mutations the space of viable networks is inferred. We have also considered a cycle where the shift is not by two, but by 1 or 3 genes (imposing the same number of steps in the target cycle). The results are nearly the same in all these cases.
5. RESULTS FOR IDEALIZED GENE EXPRESSION PATTERNS

5.2.2 Sparsity of essential interactions

Concerning the multistable phenotype, we impose $k_{FP} = 2, 3$ or 4 target fixed points. For the single fixed point case $k_{FP} = 1$, we found that the great majority of genotypes had just one essential interaction for each gene. In the present case this sparseness property also holds, but as we impose more fixed points, the mean number of essential interactions grows resulting in $1.2, 1.5, 1.9$ interactions for $k_{FP} = 2, 3$, and $4$ fixed points at $N = 16$. These average values are quite stable if one increases $N$.

In the case of using cyclic target phenotypes one gets analogous results. Again, we find that only a small fraction of interactions are essential (see Tab. 5.2). Furthermore, the number of interactions per gene that are essential hardly changes as one increases the number of genes.

Interestingly, both in the multistable case and in the cyclic scenario, for a genotype to be viable there are always a few essential inhibitory interactions present in the regulatory network. Particularly, in the case of phenotypes associated with $k_{FP} = 2$ the ratio between activatory and inhibitory interactions is like 6 to 1, and gets weaker as $k_{FP}$ increases, resulting in the ratio of 2 to 1 for four fixed points (see Tab. 5.2). This bias towards the higher number of repressors in regulatory networks conditioned to have more fixed points, can be explained by a need for more complex logic embedded within networks realizing the higher number of constraints. We get back to this topic in Sec. 5.2.4 devoted to network topologies.

For the cyclic target pattern imposed the analogous ratio is about 3 to 1, so still the activators are dominant. This ratio is not exactly what one expects for biological networks, where the proportion is more balanced $[25, 26, 27]$.

Qualitatively, the observed sparseness can be understood in the same way as in the single fixed point scenario: any additional strong interaction that emerges due to mutations is very likely to be lost (also due to mutations), unless it is kept in the regulatory network by selection pressure. Hence, the resulting average number of both activators and repressors is determined by the tradeoff between mutations and selection towards imposed functional constraints. This result is in a sharp contrast with what would arise in a model without molecular modelling of the interactions.
5.2 Multistability and cyclic conditions

<table>
<thead>
<tr>
<th>Observable</th>
<th>2FP</th>
<th>3FP</th>
<th>4FP</th>
<th>Cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle E_{ess} \rangle$</td>
<td>14.70(1)</td>
<td>21.08(2)</td>
<td>29.39(1)</td>
<td>23.42(1)</td>
</tr>
<tr>
<td>$\langle E_{rep}^{ess} \rangle$</td>
<td>2.500(12)</td>
<td>6.38(2)</td>
<td>10.11(1)</td>
<td>7.38(1)</td>
</tr>
<tr>
<td>$\langle R_{bs} \rangle$</td>
<td>0.9494(2)</td>
<td>0.9259(2)</td>
<td>0.8998(3)</td>
<td>0.9099(3)</td>
</tr>
<tr>
<td>$\langle D_T \rangle$</td>
<td>0.966(3)</td>
<td>1.390(4)</td>
<td>1.961(4)</td>
<td>3.583(4)</td>
</tr>
</tbody>
</table>

Table 5.2: A sample of results from our simulations. $k$FP stands for phenotypes with $k$ fixed points, while Cycle stands for the cyclic phenotype; for both, the number of genes is $N = 16$. The number of essential interactions is denoted by $E_{ess}$ and $E_{rep}^{ess}$ stands for the number of inhibitory interactions among all essential ones. The robustness $R_{bs}$ is defined as the frequency of genotypes surviving a random mutation according to the Metropolis rule. Notice that it is fairly well reproduced by $1 - k/2N$, a result generalizing an analogous result of Sec. 5.1.4. The distance to the target $D_T$ is the distance entering the fitness. It is nearly constant ($\approx 0.5$) when divided by the number of target phenotypes. Further division by the number of active genes in a phenotype, i.e. $N/2$, yields approximately 6%, which measures the average deviation of their activity from the maximum $S_i = 1$.

5.2.3 Basins of attraction for the regulatory dynamics

We have checked that the basin of attraction of the target phenotype represents a large fraction of all possible initial phenotypes, so that the gene regulatory network property of performing the imposed function is not merely a consequence of starting with the right expression vector. For instance, in the case of the fixed point phenotypes, these basins constitute approximately 99.8(9)% for $k_{FP} = 2$, 52.9(9)% for $k_{FP} = 3$, and 49.6(8)% for $k_{FP} = 4$, respectively. With our choice of target phenotypes and the permutation symmetry of the model, the basins associated with individual targets are equal (after averaging over functional regulatory networks).

5.2.4 Topologies of essential networks and evolvability

Given the large number of essential networks obtained from the ensemble of viable genotypes sampled by the MCMC, we have determined which essential network topologies arise most often. Particularly, two essential networks have the same topology if they are isomorphic: the corresponding networks with removed node labels are indistinguishable. Hence, the permutation symmetry exhibited by the imposed target expression patterns is taken into account. In Fig. 5.6 we display the most frequent topologies when imposing $k_{FP} = 2$ and $k_{FP} = 4$ fixed point target phenotypes. As $k_{FP}$ increases, the
connections become more complex as expected. At small $k_{FP}$ much of the topology is tree-like. The general structure of these networks can be explained in the following way: a few key (master) genes cooperate to switch network dynamics into one of the imposed target phenotypes, and then these few genes regulate in a downstream manner the other ones.

In Fig. 5.7 we show the most frequent topology arising when imposing the cyclic target phenotype on the gene regulatory network. The logic behind the network topology is immediately revealed: the clockwise feed-forward interactions drive the shifting of the active gene pattern.

Another interesting feature that emerges is that there are many different essential networks produced by our MCMC sampling. Indeed, it turns out that for $k_{FP} = 2$ and 3 there are a few most frequent topologies that are dominant in the space of all viable networks. In contrast, for $k_{FP} = 4$ the network topologies are genuinely different, and there is no single dominant topology. To get more insight into this problem, we have carried out a complementary investigation, counting the number of distinct topologies in subsamples of the large sample with 5000 independent genotypes (separately for each $k_{FP}$). Specifically, we analysed the number of distinct topologies $M$ as a function of subsample size: 125, 250, ..., 4000; for subsamples smaller than 4000 we averaged $M$ over as many non-overlapping subsamples as it could fit within the whole data set. As a result we found that the number of distinct topologies increases like a power of $M$, with the exponent increase with $k_{FP}$. The resulting exponents are approximately equal to 0.69, 0.78 and 0.97 for $k_{FP} = 2$, 3 and 4, respectively. Similar results are obtained when networks are grouped with cluster algorithm (see Appendix B for details).

This demonstrates that the same functional capability can be obtained from a very large number of distinct essential networks. We saw before that there are many gene regulatory networks that have the same phenotype, but extending this to essential networks is non trivial. Thus for a given phenotype, we have many genotypes, many essential networks, and even many topologies of essential networks! This is in a close analogy with the high degeneracy of the genotype to phenotype map found in a number of other models of gene regulation [130]. Furthermore, by construction of our MCMC, our networks are connected by a succession of point mutations, showing that gradual changes that are nearly neutral (i.e. keep the phenotype nearly constant) can lead to large evolutionary changes at long times.
5.2 Multistability and cyclic conditions

**Figure 5.6:** The most common essential network topologies when $k_{FP}$ fixed point expression patterns are imposed. In each case we see the presence of the motif with two mutually inhibitory and self-activating genes. Interactions shown are essential, and those genes whose target expression is zero in all the fixed points are omitted since they provide no information. Solid arrows correspond to activatory and dashed arrows to inhibitory interactions, respectively. Data for $N = 16$; subfigures left and right are for $k_{FP} = 2$ and 4. These topologies arise for over 20% and nearly 1% of the networks, respectively.

**Figure 5.7:** The most common essential network topology when one imposes a particular cyclic expression pattern. That pattern corresponds to a group of active genes that shifts clockwise by 2 genes at each time step. One sees very clearly the activating interactions acting forward and the inhibitory interactions acting backward. Data for $N = 16$; this topology arises for over 15% of the networks.
5. RESULTS FOR IDEALIZED GENE EXPRESSION PATTERNS

5.2.5 Functionality leads to motif selection

To obtain insights into network structure, one can search for network motifs. The fact that a complex network can be constructed from small standard subelements is by itself not surprising: this property is at the root of electronics and is based on the mathematical structure of logical functions. However, the fact that nature also uses this strategy is not obvious, and that some motifs and not others are employed in different network functions is even less obvious.

This presence of motifs is revealed through detailed studies of (rather rare, for technical reasons) biological networks reconstructed from data, and it has been partly explained by arguments borrowed from communication systems techniques. This brings us to inquire what happens in a model where the transcriptional rules are known and where thousands of networks can be generated with given network functional capabilities. Do the same motifs emerge when the functional capabilities are modified?

To answer this question, we determine the motifs in our different ensembles. The results presented here concern the most prominent motifs; others have frequencies that are either very small or at least roughly comparable to those of randomized networks. We discard motifs with leaves (degree-one nodes), which are somewhat trivial. We select only motifs that are not a subgraph of a larger motif with the same number of nodes. However, our motifs can partly overlap. The randomization used is that proposed by Maslov and Sneppen [54]: edges are interchanged so that both the in- and out-degrees of network nodes remain unchanged. Our results\footnote{Concerning network motif search software we have implemented our own realization dedicated to estimating motif frequencies for large ensembles with moderate network sizes (around 20 genes). Alternatively, a few excellent tools for searching network motifs can be found at Alon’s research group web page \url{www.weizmann.ac.il/mcb/UriAlon/} (checked on 5th June, 2013). The “mfinder” software from this web page is not quite adapted to our needs since it does not distinguish between activators and repressors, and it does not accept self-interactions. However, it was helpful in this work, enabling us to single out the relevant motif topologies and provided a benchmark for testing our own algorithm.} are summarized in Tab. 5.3 and the motifs are listed in Fig. 5.8. We see right away a very strong dichotomy: the motifs are very different for our two classes of functional capabilities (imposing multistability vs. cycling).
5.2 Multistability and cyclic conditions

5.2.5.1 Multiple fixed points constraint

In the case of multistability, *double negative feedback loop with autoregulation* (two genes are mutually inhibitory and self-activating) stands out to be extremely important. Interestingly, this simple motif is found in developmental regulatory networks (see Sec. 2.2.4.1), and in particular in the genetic switch between lysogeny and lysis of the phage $\lambda$ [29]. Clearly such a pair of genes can act as a bistable switch that influences downstream genes according to the expression pattern that is required for the considered fixed point.

For more than two target fixed points, multiple copies of this motif are found. That is illustrated in Fig. 5.6 which displays the most represented essential network (ignoring permutations of indices) for 2 and 4 imposed fixed points. Not surprisingly, the same trend also emerges for the less frequent essential networks. Roughly, the networks display a core of central genes that belong to one or more motifs of the double negative feedback type (labeled A in Fig. 5.8) and these genes then influence other genes by a simple downstream effect along the associated tree-like graph of activating interactions.

5.2.5.2 Cyclic gene expression constraint

Now consider motifs present when imposing cyclic expression targets. The motif A - *double negative feedback loop with autoregulation* - is absent and instead we have several four gene motifs that are strongly overrepresented as displayed in Fig. 5.8. In the nomenclature of Alon [14, 15], motifs B and C are *incoherent diamonds*, while motif F is the *incoherent bi-fan*; the others, motifs D and E, involve a regulatory loop, and in fact these loops are “frustrated” in that they have an odd number of inhibitory interactions. Again, biological gene networks have been found containing some of these motifs, the bi-fan motif being perhaps the most prominent (see Sec. 2.2.4). Furthermore, the presence of “frustrated” loops (negative circuits) is expected for the systems that exhibit time-periodic behaviour [74].

None of the motifs B to F were overrepresented in the networks satisfying multistability. Their presence here can be understood by looking at the most frequent essential network displayed in Fig. 5.7. The phenotype used for that case corresponds to having a block of ON genes that shifts by two units at each time step inside a background of OFF genes. One way to implement this shifting is to have genes activate the two genes...
5. RESULTS FOR IDEALIZED GENE EXPRESSION PATTERNS

<table>
<thead>
<tr>
<th>Motif</th>
<th>2FP</th>
<th>3FP</th>
<th>4FP</th>
<th>Cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>motif A: model</td>
<td>0.706(16)</td>
<td>2.358(39)</td>
<td>2.984(4)</td>
<td>0.000</td>
</tr>
<tr>
<td>randomized</td>
<td>0.002(1)</td>
<td>0.002(1)</td>
<td>0.002(2)</td>
<td>0.000</td>
</tr>
<tr>
<td>motif B: model</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>5.451(41)</td>
</tr>
<tr>
<td>randomized</td>
<td>0.001(1)</td>
<td>0.008(3)</td>
<td>0.071(9)</td>
<td>0.023(5)</td>
</tr>
<tr>
<td>motif C: model</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>5.170(40)</td>
</tr>
<tr>
<td>randomized</td>
<td>0.000</td>
<td>0.008(3)</td>
<td>0.078(9)</td>
<td>0.029(1)</td>
</tr>
<tr>
<td>motif D: model</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>4.533(42)</td>
</tr>
<tr>
<td>randomized</td>
<td>0.000</td>
<td>0.014(4)</td>
<td>0.0180(6)</td>
<td>0.030(7)</td>
</tr>
<tr>
<td>motif E: model</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>6.676(22)</td>
</tr>
<tr>
<td>randomized</td>
<td>0.017(4)</td>
<td>0.102(16)</td>
<td>0.057(8)</td>
<td>0.173(14)</td>
</tr>
<tr>
<td>motif F:</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>2.296(29)</td>
</tr>
<tr>
<td>randomized</td>
<td>0.000</td>
<td>0.001(1)</td>
<td>0.050(6)</td>
<td>0.003(2)</td>
</tr>
</tbody>
</table>

Table 5.3: The frequencies of the most important network motifs found for multistability and cyclic constraints. kFP means “k fixed point phenotypes”. Cycle refers to our 8-step cyclic phenotype. We typically used 1000 gene regulatory networks in our motif search. Motifs corresponding to the symbols are shown in Fig. 5.8.

ahead of them and to inhibit genes sufficiently far behind. As can be seen in Fig. 5.7, the actual strategy used in the most frequent essential network is very close to that: if we consider the two genes at the front of the ON block, the front-most one activates the two ahead of it while the second has no excitatory action. And each inhibits one gene at the back of the block that is to be turned OFF at the next time step. The combination of these excitatory and inhibitory actions ensures that the block preserves the correct size as it shifts.

Interestingly, and in contrast to the situation for the phenotypes consisting of steady states, the operation of the network is not readily understood from looking at the motifs in isolation. Indeed, the genes in these motifs do not provide oscillatory behaviour on their own. In effect, the motifs here are just like parts in a larger machine, and it is necessary to consider how they cooperate within the overall network to reveal their function. Only for much smaller networks the corresponding motifs reveal functions on their own1.

---

1In the Supporting Information of [20], we illustrate this by considering a three gene network which exhibits periodic oscillations [168, 169].
Figure 5.8: The most prominent motifs found for our two classes of functionality constraints. Case of multistability (more than one fixed point): (A) double negative feedback loop with auto regulation, which has two mutually inhibitory and self-activating genes. Case of target expression patterns that are cyclic in time: (B,C) incoherent diamond, (D,E) frustrated four-node loop, (F) incoherent bi-fan.
5. RESULTS FOR IDEALIZED GENE EXPRESSION PATTERNS
Chapter 6

Results for experimental gene expression patterns

Up to now we have used idealized expression patterns as constraints imposed on the network functionality. In this chapter, we use experimentally reconstructed gene expression patterns found in three organisms: baker’s yeast (S. cerevisiae) [25], fission yeast (S. pombe) [26] and mammals [27]. Although these experimental patterns resemble our toy sequence for the cyclic expression, they also have significant differences, like varying number of active genes with each time step. As a result, the imposed pattern is irregular and in effect also the obtained networks are specific for each of the three considered functional ensembles. However, the structure of these networks can be compared with the reconstructed wild-type networks. By studying topological properties of networks we can quantify how well our relatively simple thermodynamical model can reproduce connections found in real organisms.

Interestingly, we find that the actual number of effective genetic interactions becomes an emergent feature in our model. Furthermore, for the two yeast species on average roughly 60% of interactions are well reproduced, and in the whole ensemble almost all networks have from 50% to 70% of links in common with the wild-type network. As for the mammalian cell cycle the results are less impressive, showing the model limitations. Actually, it would be suspicious if the thermodynamic considerations on their own explained the complex regulatory mechanism that is present in mammals.
6.1 Model extension

In Chapter 4 we developed a biophysical framework for modelling the regulatory interactions between a transcription factor (TF) and its DNA binding site. For biological cell cycling systems, a number of molecules, such as cyclins, are not TFs, so our framework must be generalized. The starting point is a network whose nodes refer to molecules and edges correspond to activation/repression processes of transcriptional, post-transcriptional, translational or post-translational nature. Hence, we build an effective description for a more general class of processes.

Particularly, all interactions involve specific binding sites or surfaces that allow two molecules to physically establish contact and bind through the addition of small forces. Almost all of the facing elements (atoms, bases, amino acids, ...) have to “match” for the two molecules to bind. Similar to [170], we consider that each of the facing regions is specified by an ordered lists of symbols that can be thought of as a string or a table characterizing that type of molecule and its site or surface dedicated to that binding partner; in the previous context, these two strings were respectively DNA binding site and DNA binding domain that is present on TF molecule. Then the mutual binding energy of two molecules is taken to be additive in the number of mismatches between their strings, each mismatch contributing a penalty $\varepsilon$.

Such a framework is easily justified for TF-DNA binding energies and leads to a thermodynamic formula for the probability of occupation of a given DNA binding site in the presence of $n_j$ TFs of type $j$. The derivation of Eq. (4.2) in Sec. 2.3.2 shows that it is quite general (does not explicitly depend on the type of regulatory protein), so we shall apply it to all of the molecular species in our cell cycle system.

Notice that Eq. (4.2) holds provided the system is in equilibrium. Thus, in writing Eq. (4.6) it is assumed that the binding-unbinding reaction is fast compared to that for full the activation/repression process. However, the rate of a binding-unbinding reaction is proportional to the concentration of reactants and the equilibration time becomes very large for small concentrations. Hence for very small concentrations the unphysical assumption that there is always enough time to reach equilibrium sometimes leads to pathologies. In particular, the system suffers from an instability: even a relatively weak interaction has an a priori capacity of generating a spurious progressive amplification of physically insignificant (because very low) expression levels. To avoid
6.1 Model extension

Figure 6.1: Input to output relation for (left) the model with just thermodynamics; (right) the model with the heuristic threshold added. The lines correspond to the expression level of gene $i$ activated by gene $j$, where the interaction strength has the form $W_{ij} = \exp(-\varepsilon d_{ij})$ with $\varepsilon = 1.75$. The crossing point with the diagonal represents the fixed point expression level when the system is restricted only to the $i$th and $j$th genes (except for the crossing with vertical part in the right subfigure which is related to the discontinuity in $S_i$).

this bad behaviour and improve the robustness of the model, we have recourse to a phenomenological correction, introducing a threshold $H$ to modify Eq. (4.6) whenever the expression level is too small:

$$S_i(t + 1) = S^{\text{min}} \quad \text{when} \quad S_i(t + 1) < H.$$  \hfill (6.1)

In effect, $S^{\text{min}}$ can be interpreted as a basal rate or some level of leakiness on the transcription, whereas the threshold $H$ ensures that expression levels stay low unless the input signal is sufficiently activatory (see Fig. 6.1). In practice we set $S^{\text{min}} = 1/n = 0.001$ as if there were just one molecule of that species and $H$ is set to $H = 0.01$. Then, $S_i(t)$ can leave the minimal level only when at least one $d_{ij}$ is small enough to have a dynamical significance.

6.1.1 Computational aspects

Similar to TF-DNA binding we keep the modelling simple, we consider that the original strings use a four letter code. For each $i$ and each $j$, $(i,j)$, the interaction strength $W_{ij}$ is given by Eq.(4.2) where $d_{ij}$ can be considered to be the number of mismatches between two strings of length $L$ [170]. Then the probability that a mutation replaces
a match by a mismatch is 3/4, while the probability that a mismatch is converted into
a match is 1/4, embodying the fact that it is easier to have a mismatch than a match.
Note that even for protein-protein interactions, this rule can be motivated by the fact
that molecular changes are often the result of a point mutation at the DNA level.

Values of the parameters used in our computations are \( L = 12, \varepsilon = 1.75, f = 20, \)
\( H = 0.01 \) and \( n = 1000 \). Compared to previous Chapter 5, \( H \) is a new parameter, \( \varepsilon \)
is slightly smaller while the remaining are the same. Interestingly, it is essential that
\( L \) and \( \varepsilon \) be such that the \textit{a priori} probability of a small mismatch is very low. Once
that constraint is taken into account, the model results are robust with respect to the
variation of the parameters.

The Metropolis test makes the total distance between the expression pattern and
target pattern to be relatively small if \( f \) is large, but the MCMC is inefficient in that
regime, forcing us to work with not too large values of \( f \). Then in the \( T \times N \) table of
numbers \( S_i(t) \), there may appear a few that are “ambiguous”, that is that are away from
the target value which in this study is either 0 or 1. For example an expression level
may have a value 0.3 in a place where the target value is 0. Forcing a better agreement
by increasing \( f \) is not computationally feasible. To overcome this difficulty, from our
large MCMC sample, we have selected genotypes where such ambiguities are absent,
imposing that expressed (respectively unexpressed) molecular types have in fact levels
above 0.6 (respectively below 0.2). This selection is done mostly for esthetic reasons,
the properties of the full sample are essentially the same because of a low frequency of
these ambiguities.

### 6.2 Experimental target phenotypes

Compared to the idealized cyclic case, now we are interested in the process with a
sequence of target patterns that ends up in a fixed point phenotype. We use three target
gene expression patterns that were derived experimentally for cell division cycles in: (i)
baker’s yeast [25] (see Tab. 6.1), (ii) fission yeast [26] (see Tab. 6.2), (iii) mammals [27]
(see Appendix C, Tab. 8.1).

Note that the number of time steps between target phenotypes does not correspond
to real time that elapses between two stages in a cell division process. Instead these dic-
scretized (each gene is either ON or OFF) expression profiles indicate consecutive stages
6.2 Experimental target phenotypes

<table>
<thead>
<tr>
<th>Cln3</th>
<th>MBF</th>
<th>SBF</th>
<th>Cln1,2</th>
<th>Clb5,6</th>
<th>Clb1,2</th>
<th>Mcm1</th>
<th>Cdc20</th>
<th>Swi5</th>
<th>Sic1</th>
<th>Cdh1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.1: The cell cycle expression profile of baker’s yeast (from [25]). There are \( N = 11 \) genes. Time runs from the top to the bottom. Successive, discretized expression levels, are either 0 or 1, at each time step. The initial state expression \( S^{(\text{initial})} \) is given by the first line of the table, and the consecutive states form a target pattern \( S^{(\text{target})} \). The last state is a fixed point associated with the cell size check point. When a critical cell size is reached, the Cln3 is activated, and the cycle is triggered again.

of the division process in which different genes are activated. Particularly, Tab. 6.1 shows the sequence of \( T = 13 \) states with 11 genes for baker’s yeast [25], fission yeast has \( T = 10 \) states with 10 genes [26] and for mammals the modelling [27] specifies \( T = 7 \) steps with 7 genes.

6.2.1 Sparsity in biological ensembles

Our model naturally generates sparse essential networks because relevant mismatches are \( a \text{ priori} \) unprobable and arise only under a strong selection pressure. Then simply because of mutation-selection balance, the model strongly favours low numbers of essential interactions. It is interesting to see, if the same principle leads to quantitative results that can be directly compared to biological data as inferred by the different authors [25, 26, 27].

In Tab. 6.3 we give the average number of essential interactions in the considered
6. RESULTS FOR EXPERIMENTAL GENE EXPRESSION PATTERNS

<table>
<thead>
<tr>
<th></th>
<th>Start</th>
<th>SK</th>
<th>Cdc2</th>
<th>Cdc25</th>
<th>Cdc2*</th>
<th>Slp1</th>
<th>PP</th>
<th>Ste9</th>
<th>Rum1</th>
<th>Wee1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.2: The cell cycle expression profile of fission yeast (from [26]). There are \( N = 10 \) genes. The rest of conventions are analogous to Tab. 6.1. The cell cycle is triggered by a so-called “Start” node, when a critical cell size is reached. The Cdc2 (Cdc2*) stands for Cdc2/Cdc13 (Cdc2/Cdc13*), where the * sign indicates the highly activated form of the complex [26]. Similarly Wee1 stands for the Wee1/Mik1 kinases.

ensembles. Furthermore, in Fig. 6.2 the distributions of essential interactions can be compared to the number of interactions present in the wild-type networks: 29 for baker’s yeast, 23 for fission yeast and 22 for mammals. The mammalian biological network thus appears to use significantly more interactions than our model, but the two yeast species seem to be in a good agreement with our model\(^1\). To quantify this, we have measured the fluctuations in each ensemble: we find means \( \mu \) = 27.0, 21.0 and 13.9 and standard deviations \( \sigma \) = 1.8, 1.3 and 0.9 respectively for baker’s yeast, fission yeast and mammals. Comparing to the maximal possible number of interactions, we see that the gene regulatory networks produced by the model are sparse and have only small fluctuations in the number of essential interactions.

The fact that our model does much better for yeasts than for mammals may not be a coincidence. Indeed, although baker’s yeast and fission yeast are highly divergent evolutionarily, much more so than any mammals amongst themselves, they are both unicellular organisms. It is well known that the regulatory control of gene expression is intrinsically more complex in multicellular organisms, so a posteriori one may draw

\(^1\)We do not count the “self-degradation” interactions added by hand by some of these authors which are specific to their modelling of the discrete time dynamics of the expression levels.

86
6.2 Experimental target phenotypes

<table>
<thead>
<tr>
<th>Observable</th>
<th>Baker’s wt</th>
<th>Fission</th>
<th>Mammalian</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{wt}$</td>
<td>29</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>$\langle E_{ess} \rangle$</td>
<td>27.98(2)</td>
<td>20.96(2)</td>
<td>13.92(1)</td>
</tr>
<tr>
<td>$\langle E_{rep} \rangle$</td>
<td>12.20(2)</td>
<td>9.02(2)</td>
<td>5.30(2)</td>
</tr>
<tr>
<td>$\langle D_T \rangle / T$</td>
<td>0.534(1)</td>
<td>0.493(1)</td>
<td>0.491(1)</td>
</tr>
<tr>
<td>$\langle S_{ON} \rangle$</td>
<td>0.881(1)</td>
<td>0.861(1)</td>
<td>0.879(1)</td>
</tr>
<tr>
<td>$\langle S_{OFF} \rangle$</td>
<td>0.011(1)</td>
<td>0.014(1)</td>
<td>0.026(1)</td>
</tr>
</tbody>
</table>

Table 6.3: Statistics from networks associated with the cell cycle expression patterns of the baker’s yeast, fission yeast and mammalian expression pattern. $E_{wt}$ stands for number of interactions (edges) in the wild-type network. The other symbols refer to properties of in silico networks generated by MCMC: $E_{ess}$ is the number of essential interactions; $E_{rep}$ is the number of inhibitory interactions; $D_T$ is the total distance between the actual expression patterns and target one (nearly constant and $\approx 0.5$ when divided by the number of steps of the cycle); $S_{ON}$ ($S_{OFF}$) is the expression level of genes that are ON (OFF) in the target phenotype.

Figure 6.2: Number of essential interactions in regulatory networks generated within the baker’s yeast (left), fission yeast (middle) and mammalian cell cycle (right) ensembles. For both yeast species the number of interactions measured in the wild-type networks [25, 26] is close to the position of the peak in the ensemble distribution. In the case of the mammalian cell cycle [27] a large discrepancy between the distribution peak and number of interactions in wild-type network is present.
### 6. RESULTS FOR EXPERIMENTAL GENE EXPRESSION PATTERNS

#### Activatory interactions

<table>
<thead>
<tr>
<th></th>
<th>Cln3</th>
<th>MBF</th>
<th>SBF</th>
<th>Clb5,6</th>
<th>Clb1,2</th>
<th>Mcm1</th>
<th>Cdc20</th>
<th>Swi5</th>
<th>Sic1</th>
<th>Cdh1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cln3</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MBF</td>
<td>1</td>
<td>0.5</td>
<td>0.54</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SBF</td>
<td>0.01</td>
<td>0.51</td>
<td>0.52</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clb5,6</td>
<td>0</td>
<td>0.07</td>
<td>0.06</td>
<td>0.58</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clb1,2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.99</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mcm1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cdc20</td>
<td>0</td>
<td>0.02</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Swi5</td>
<td>0.05</td>
<td>0</td>
<td>0.01</td>
<td>0.05</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sic1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cdh1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Inhibitory interactions

<table>
<thead>
<tr>
<th></th>
<th>Cln3</th>
<th>MBF</th>
<th>SBF</th>
<th>Clb5,6</th>
<th>Clb1,2</th>
<th>Mcm1</th>
<th>Cdc20</th>
<th>Swi5</th>
<th>Sic1</th>
<th>Cdh1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cln3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MBF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SBF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clb5,6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clb1,2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mcm1</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cdc20</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Swi5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sic1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cdh1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 6.3:** Frequencies of activators (left) and inhibitors (right) for networks produced in the ensemble relevant for the cell cycle in baker’s yeast. Array element \((i,j)\) corresponds to the probability of finding a link \(i \leftarrow j\) in the analysed ensemble of gene regulatory networks. Dark grey entries have frequencies higher than light grey entries; a thick solid frame indicates that the link is present in the yeast wild-type network [25] (see also Fig. 6.6).

some satisfaction from the fact that our simple model seems to do well when regulation is simple but not that well when regulation involves very complex mechanisms.

#### 6.2.2 Edge usage

Now consider the edge usage in essential networks. The frequency with which there is an essential interaction from node \(j\) to node \(i\) can be represented by a square matrix with entries between 0 and 1. We have computed these matrices for the three cell cycles, and in Fig. 6.3 we display the results separately for activating and inhibitory interactions in baker’s yeast. Particularly, there are 9 activators (7 common with the wild-type) and 7 repressors (4 common with the wild-type) almost always present in the ensemble for baker’s yeast (c.f. dark grey entries in Fig. 6.3).

Further, the results can be compared with frequencies of activatory/inhibitory interactions from [132], where 6 activators and 4 repressors (all common with the wild-type) are found to be absolutely required for a network to produce the cell cycle process. Specifically, among these absolutely required interactions, 8 edges are almost always present and two edges are present in more than 80% of networks in our ensemble.

On a more general level there is a manifest dichotomy: certain interactions are very rare or absent, while others are very frequent or always present. Furthermore, there is a clear structure in the matrix: in the line just below the main diagonal, activating
### 6.2 Experimental target phenotypes

#### Activatory interactions

<table>
<thead>
<tr>
<th>Gene</th>
<th>Start</th>
<th>SK</th>
<th>Cdc2</th>
<th>Cdc25</th>
<th>Cdc2/Star</th>
<th>Slp1</th>
<th>PP</th>
<th>Ste9</th>
<th>Rum1</th>
<th>Wee1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>0.47</td>
<td>0.08</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>SK</td>
<td>1</td>
<td>0.53</td>
<td>0.14</td>
<td>0</td>
<td>0.01</td>
<td>0.14</td>
<td>0.52</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cdc2</td>
<td>0.1</td>
<td>0.16</td>
<td>0.38</td>
<td>0</td>
<td>0.01</td>
<td>0.14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cdc25</td>
<td>0.02</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cdc2/</td>
<td>0.01</td>
<td>0.17</td>
<td>0.82</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Slp1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PP</td>
<td>0</td>
<td>0.01</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ste9</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.47</td>
<td>0.02</td>
<td>0.24</td>
<td>0.68</td>
<td>0.34</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rum1</td>
<td>0</td>
<td>0.02</td>
<td>0.12</td>
<td>0.08</td>
<td>0.01</td>
<td>0.12</td>
<td>0.89</td>
<td>0.12</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>Wee1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
<td>0</td>
<td>0.14</td>
<td>0.64</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Inhibitory interactions

<table>
<thead>
<tr>
<th>Gene</th>
<th>Start</th>
<th>SK</th>
<th>Cdc2</th>
<th>Cdc25</th>
<th>Cdc2/Star</th>
<th>Slp1</th>
<th>PP</th>
<th>Ste9</th>
<th>Rum1</th>
<th>Wee1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SK</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cdc2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cdc25</td>
<td>0.06</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cdc2/</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Slp1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ste9</td>
<td>1</td>
<td>0.47</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rum1</td>
<td>1</td>
<td>0.32</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Wee1</td>
<td>1</td>
<td>0.35</td>
<td>0</td>
<td>0.08</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 6.4:** Frequencies of activators (left) and inhibitors (right) for networks produced in the ensemble relevant for the cell cycle in fission yeast. The conventions for highlighting the values are the same as in Fig. 6.3 but here we refer to the wild-type network of fission yeast as specified in [26] (see also Fig. 6.7).

Interactions are very frequent. A different pattern arises for the inhibitory interactions where all the entries near the diagonal are always absent, and elements 4 or 5 steps shifted from the diagonal are frequent or always present.

Qualitatively the same pattern arises in fission yeast but it is noisier (Fig. 6.4), and in the case of the mammalian cell cycle which involves only 7 genes, there are simply remnants of this pattern (see Appendix C, Fig. 8.1).

#### 6.2.3 Overlaps with the wild-type networks

If an essential interaction of an *in silico* gene regulatory network is present in the wild-type network, we shall call this interaction a *hit*. What fraction of the essential interactions of a regulatory network are hits? If the fraction is close to 1, there is a very strong overlap between that regulatory network and the wild-type network. As can be seen in Fig. 6.5, in practice the fraction varies from network to network but typically takes values in the range from 50% to 70% in both yeast species. The mean fractions are 57% and 65% for baker’s and fission yeast so clearly there is a strong overlap between the *in silico* and wild-type networks. The highest fraction of hits found in our MCMC is 73% for baker’s yeast and 76% for fission yeast. Notice that this fraction is close to the overlap of hits as compared to all links in the wild-type network, since in our model $E_{wt}$ and $E_{ess}$ are similar for both yeast species (see Tab. 6.3). To compare visually the associated networks to the corresponding wild-type networks, we display...
6. RESULTS FOR EXPERIMENTAL GENE EXPRESSION PATTERNS

Figure 6.5: Histograms for the fraction of essential interactions that are “hits” in the baker’s yeast (left), fission yeast (middle) and mammalian cell cycle (right) ensembles. For the two yeast species we see that networks in our ensembles have roughly 50% to 70% of interactions in common with the wild-type networks. Notice that this fraction, can significantly differ from the fraction of hits with respect to all links in the wild-type networks, when the average number of links in MCMC ensemble differs significantly from number of edges in the wild-type; for both yeast species this is not an issue, but it is relevant for mammalian cell cycle (c.f. Fig. 6.2).

them in Fig. 6.6 for baker’s yeast and in Fig. 6.7 for fission yeast. A comparison of the number of overlapping interactions with the distinction between activatory and inhibitory interactions is given in Tab. 6.4.

The main differences arise for the Sic1 and Clb1,2 genes which in the wild-type have higher degree and more mutually inhibitory interactions than in the regulatory networks generated by MCMC. Such interaction pairs are quite natural in the context of protein-protein interactions and thus probably reveal a shortcoming of our modelling framework.

6.2.4 A multitude of different essential networks

The MCMC generates a very large number of genotypes from which we construct essential networks. In effect, each essential network corresponds to a different way interactions can be specified so as to reproduce the target gene expression dynamics. Many genotypes give rise to the same essential network, but are there many different essential networks? The answer of course is yes, we find hundreds of different such networks.

One may then ask whether some arise much more frequently than others. On the right hand side of Fig. 6.6 (respectively Fig. 6.7) we show the most frequently
6.2 Experimental target phenotypes

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>$E_{wt}^{act}$</th>
<th>$\langle E_{ess}^{act} \rangle$</th>
<th>$\langle E_{hits}^{act} \rangle$</th>
<th>$E_{wt}^{rep}$</th>
<th>$\langle E_{ess}^{rep} \rangle$</th>
<th>$\langle E_{hits}^{rep} \rangle$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baker’s</td>
<td>15</td>
<td>14.78(2)</td>
<td>9.04(1)</td>
<td>14</td>
<td>12.20(2)</td>
<td>6.30(1)</td>
</tr>
<tr>
<td>Fission</td>
<td>8</td>
<td>11.94(2)</td>
<td>5.19(1)</td>
<td>15</td>
<td>9.02(2)</td>
<td>8.46(2)</td>
</tr>
<tr>
<td>Mammalian</td>
<td>9</td>
<td>8.61(1)</td>
<td>3.86(1)</td>
<td>13</td>
<td>5.30(2)</td>
<td>4.04(1)</td>
</tr>
</tbody>
</table>

**Table 6.4:** Comparison of the average number of hits and essential interactions separately for activators and repressors for the baker’s yeast, fission yeast and mammalian cell-cycle ensembles. Symbols correspond to: $E_{wt}^{act}$ is the number of activators in the wild-type network, $E_{ess}^{act}$ is the number of essential interactions that are activators, $E_{hits}^{act}$ is the number of these activators that are also present in the wild-type network; similarly symbols with $rep$ instead of $act$ refer to inhibitors instead of activators.

**Figure 6.6:** Left: The wild-type regulatory network for the baker’s yeast cell cycle has 29 links, of which 15 are activating (solid) and 14 are inhibitory (dashed). Compared to ref. [25] we do not include the added self-degradation links. Middle: Regulatory network in the MCMC ensemble with the highest fraction of hits (73% of the essential interactions are present in the wild-type). This network has 19 hits (black) and 7 non-hits (light grey) for a total of 26 essential interactions. Right: The most frequent network in the baker’s yeast ensemble (2.6% of all gene regulatory networks, 16 hits (black) and 10 non-hits (light grey) for a total of 26 essential interactions).
6. RESULTS FOR EXPERIMENTAL GENE EXPRESSION PATTERNS

Figure 6.7: Left: The wild-type regulatory network for the fission yeast cell cycle has 23 links, of which 8 are activating (solid) and 15 inhibitory (dashed). Compared to [26] we do not include the added self-degradation links. Middle: Regulatory network in the MCMC ensemble with the highest fraction of hits (76% of the essential interactions are present in the wild-type). The network has 16 hits (black) and 5 non-hits (light grey) for a total of 21 essential interactions. Right: Most frequent network in the fission yeast ensemble (2.1% of all gene regulatory networks, 14 hits (black) and 6 non-hits (light grey) for a total of 20 essential interactions).

found essential network for baker’s yeast (respectively fission yeast); these arise with frequencies between 2 and 3%. Their characteristics are similar to those of the essential networks having the highest overlaps with the wild-type networks. These most frequent essential networks arise hundreds of times more frequently than the rarest ones. In Fig. 6.8 we show the associated rank histogram for the concurrence frequencies on a log-log plot. The fat tail in this distribution is roughly compatible with a power law. Such a shape for a rank histogram arises in a number of other systems, in particular in neutral networks where many genotypes give rise to the same phenotype.

6.2.5 Network motifs in biological ensembles

Given the genotypes in each of our ensembles, produced by the MCMC, we have extracted network motifs present in the associated essential networks. The different motifs we find are represented graphically in Fig. 6.9. For two nodes, we find no significantly overexpressed motifs. For three nodes, we find that for the baker’s yeast ensemble three coherent feed-forward loops (FFLs) are strongly (by a factor close to 10) overrepresented. In case of the fission yeast ensemble, two of these FFLs are overrepresented,
6.2 Experimental target phenotypes

Figure 6.8: Frequency in the MCMC of different gene regulatory networks (GRNs) for the three ensembles: baker’s yeast (squares), fission yeast (circles) and mammalian cell cycle (triangles). In all cases there are a few most frequent networks and the distribution of frequencies for the rarer networks roughly follows a power law.

Figure 6.9: Most prominent network motifs found in our gene regulatory network ensembles. The corresponding frequencies are given in Tab. 6.5.
6. RESULTS FOR EXPERIMENTAL GENE EXPRESSION PATTERNS

but they are present only in roughly 20% of generated regulatory networks. Finally, for four nodes, we find the presence of numerous motifs: two kinds of frustrated four-point loops, two kinds of incoherent diamonds and incoherent bi-fans. In the case of the mammalian cycle we find no significant network motifs. The detailed frequencies of each of the different motifs in these ensembles are given in Tab. 6.5.

Finally, instead of working within the \textit{in silico} gene regulatory networks produced by our ensembles, one can check for motifs in the wild-type networks. We find that the motifs are not all the same as in the associated ensemble. Among the strongest differences, the double negative feedback loop motif (RR) consisting of two mutually inhibitory nodes is strongly overrepresented in the wild-type networks, but not in the \textit{in silico} networks. Particularly, in the baker’s yeast wild-type network this motif is present 3 times whereas in our ensemble the frequency is 0.33(1) (for randomized networks the frequency of RR motif is virtually the same). For fission yeast the underrepresentation of RR motif with respect to wild-type network is even stronger; in the biological network there are 4 RR motifs whereas in our ensemble the frequency is almost 20 times smaller. One of the reasons for this absence of RR motifs for the yeast cell cycle pattern imposed, is that the full cell cycle process exhibits checkpoints\footnote{Checkpoint is a molecular mechanism that detects problems in cell cycle processes and prevents progression to later events until the earlier processes can be successfully completed.} \cite{171, 172} that can be interpreted as switches between enabling or blocking progression in the cycle. Since, the target gene expression patterns used to obtain yeast ensembles does not exhibit any types of checkpoints, it is not surprising that also no RR motifs are present. It is likely, that if additional phenotypic constraints are introduced (\textit{i.e.} mimicking the checkpoint functionality) the RR motifs will emerge in our ensembles.

Differences exist also for a number of other motifs. For example, in the wild-type network of baker’s yeast, the numbers are: A (C1-FFL) = 2, C (C4-FFL) = 1, D (neg 3-loop) = 2. In the case of fission yeast, the numbers are G = 2, H = 2 (for the two types of diamonds) and I = 1 (for bi-fan). These numbers can be compared to those in the gene regulatory network ensembles using Tab. 6.5.

Apart from the network motifs presented in Tab. 6.5, it is worth to discuss the activatory cascade motif\footnote{This network motif is not included in Tab. 6.5 as it has leaves (loose edges).}, whose presence was already indicated by cascade patterns in Fig. 6.3 and 6.4. For the baker’s yeast wild-type network (Fig. 6.6) we see for instance
6.2 Experimental target phenotypes

a long sequence of activators: Cln3 → MBF → Clb5,6 → Mcm1 → Swi5 → Sic1, which is reproduced in some of networks in our ensemble (for instance in the regulatory network with the highest fraction of hits). Concerning shorter cascades consisting of four genes, we find on average 3.79(2) of these motifs in the baker’s yeast ensemble (for randomized networks the frequency is 0.61(1)). The function of such a cascade is clear: if provided with external input (i.e. a “block” of ON genes) the motif will propagate gene expressions as a cascade of falling dominoes \[173\] until the block of 1’s is pushed out and the whole region consists of 0’s.

<table>
<thead>
<tr>
<th>Motifs</th>
<th>Baker’s yeast</th>
<th>Fission yeast</th>
<th>Mammalian cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (C1-FFL)</td>
<td>1.04(2)</td>
<td>0.0075(11)</td>
<td>0.0077(10)</td>
</tr>
<tr>
<td>randomized</td>
<td>0.30(1)</td>
<td>0.162(7)</td>
<td>0.038(3)</td>
</tr>
<tr>
<td>B (C3-FFL)</td>
<td>1.93(2)</td>
<td>0.23(1)</td>
<td>0.0018(5)</td>
</tr>
<tr>
<td></td>
<td>0.22(1)</td>
<td>0.094(5)</td>
<td>0.016(2)</td>
</tr>
<tr>
<td>C (C4-FFL)</td>
<td>1.93(2)</td>
<td>0.21(2)</td>
<td>0.0011(5)</td>
</tr>
<tr>
<td></td>
<td>0.22(1)</td>
<td>0.099(5)</td>
<td>0.019(2)</td>
</tr>
<tr>
<td>D (neg 3-loop)</td>
<td>0.80(1)</td>
<td>0.091(6)</td>
<td>0.054(3)</td>
</tr>
<tr>
<td></td>
<td>0.34(1)</td>
<td>0.258(8)</td>
<td>0.099(4)</td>
</tr>
<tr>
<td>E (neg 4-loop)</td>
<td>0.81(2)</td>
<td>0.063(5)</td>
<td>0.0001(2)</td>
</tr>
<tr>
<td></td>
<td>0.061(3)</td>
<td>0.023(3)</td>
<td>0.0021(5)</td>
</tr>
<tr>
<td>F (neg 4-loop)</td>
<td>0.98(2)</td>
<td>0.023(3)</td>
<td>0.0014(4)</td>
</tr>
<tr>
<td></td>
<td>0.065(3)</td>
<td>0.020(2)</td>
<td>0.0020(6)</td>
</tr>
<tr>
<td>G (diamond)</td>
<td>0.41(2)</td>
<td>0.098(5)</td>
<td>0.0000(0)</td>
</tr>
<tr>
<td></td>
<td>0.045(3)</td>
<td>0.030(4)</td>
<td>0.0036(9)</td>
</tr>
<tr>
<td>H (diamond)</td>
<td>0.48(2)</td>
<td>0.050(4)</td>
<td>0.060(3)</td>
</tr>
<tr>
<td></td>
<td>0.071(3)</td>
<td>0.050(3)</td>
<td>0.013(2)</td>
</tr>
<tr>
<td>I (bi-fan)</td>
<td>0.10(1)</td>
<td>0.0037(9)</td>
<td>0.021(2)</td>
</tr>
<tr>
<td></td>
<td>0.023(3)</td>
<td>0.0035(8)</td>
<td>0.0000(0)</td>
</tr>
</tbody>
</table>

Table 6.5: Frequencies of motifs found in regulatory networks generated with the baker’s yeast, fission yeast and mammalian cycle target patterns. For each motif, the first line provides the frequency in the MCMC ensemble and the second line provides the frequency in the randomized networks. Numbers in bold indicate motifs for which the frequency in the gene regulatory network ensemble is at least two times higher than in randomized ensemble. Motifs corresponding to the symbols are shown in Fig. 6.9.
6. RESULTS FOR EXPERIMENTAL GENE EXPRESSION PATTERNS
Chapter 7

Conclusions and outlooks

It is quite remarkable that a model, based on rather elementary thermodynamic and probabilistic principles, succeeds in reproducing many topological properties of gene regulatory networks. Our *in silico* approach is based on considering all possible networks constrained to exhibit a given gene expression pattern. In effect, this approach allows one to infer how the network regulatory architecture is constrained by the network functional capabilities both in idealized and experimentally motivated systems. But one should be aware that functional constraints, on their own, would not result in regulatory networks resembling biological designs, that is networks being sparse, with a few incoming interactions, having pleiotropic genes, exhibiting specific network motifs, *etc.* These properties are recovered due to balance between mutation and selection, where the latter picks regulatory designs that most faithfully reproduce imposed expression patterns.

The concept of mutation-selection tradeoff reduces to a simple argument: mutations try to remove any strong interaction that emerged by chance while selection favours regulatory connections that increase genotype fitness. Hence, these two opposing forces, in a sense, compete with each other, and the obtained regulatory networks are an outcome of this struggle. If the selection is too weak, no viable genotype can emerge, and the mutation wins. On the contrary, if the selection is too strong it outperforms the mutation: the system will find a viable genotype, but the corresponding regulatory network will not be evolvable. Ideally, it is to sample a system in a regime where both mutation and selection are balanced out: the resulting network topologies have
properties observed for regulatory designs in living organisms, and they are evolvable, *i.e.* many network topologies realize the same function.

First, on a very simple case with a single fixed point imposed on the system dynamics we demonstrated that indeed our sampling procedure results in a very sparse networks that are evolvable. Typically, there is only one incoming interaction per gene, and the in-degree distribution is narrow (possibly exponential). The number of these interactions is set by the mutation-selection balance. We also proved that the system can gradually evolve from a completely random genotype with a very low fitness to a viable genotype that is highly fit. Concerning the out-degree distribution we gave a qualitative reason (rooted in our modelling) why it should be broad. More precisely, to validate our reasoning we performed a simulation of large population within a separate balls-in-boxes type of model in our work [19] that confirmed the qualitative argument. Furthermore, generated viable genotypes have heterogeneously distributed robustness: most of the mutations are innocuous to regulatory network fitness, but a few that change essential interactions are typically lethal.

Second, we showed that depending on the type of functional constraint (multistability resembling cell development vs. time-periodic expression found in cell cycles) imposed on the *in silico* network evolution, different classes of network motifs emerged. In the case of multistability the double negative feedback loop with autoregulation was overrepresented, and for the periodic constraint four node motifs like bi-fan, diamond and frustrated loop were highly significant. Quite surprisingly, the double negative feedback loop was completely absent in the cyclic case, and conversely, no network motifs present for the cycle were found in for multiple fixed points. This result is even more striking if we realize that no motif structures are incorporated in our framework on any level. Instead, motifs emerge due to mutations and selection towards imposed functional constraints.

As could be anticipated, essential networks obtained under multi-stability and cyclic conditions are sparse and make use of inhibitory interactions parsimoniously. Inhibitory interactions within our transcriptional dynamics act as a *veto*, hence their role in regulatory networks is typically related to turning OFF a given gene (as it is in the case of cycle where a gene gets activated and then it needs to be inhibited) or keeping it OFF all the time (as it is in the case of the bistable switch with autoregulation). If we analyse results for the multiple fixed point constraint, we see that the bistable switch
arises once when imposing two fixed points, twice when imposing three fixed points, etc. From a design perspective: the choice of going to one fixed point rather than to another can be implemented most simply by using bistable switches that operate in this logical fashion. What is very gratifying, is that a similar mechanism is present during cell differentiation process: the cell fate is decided by a series of definite choices between two alternative expression patterns.

Third, from idealized gene expression patterns we moved on to expression profiles observed in biological systems. Our modelling was able to give a good indication of the number of interactions and of the actual edges that are used in the two yeast species. We also considered network motifs and edge usage in the generated ensembles. We found a few types of the feed-forward loop (the most abundant three node motif found in biological networks) that was missing in the idealized cell cycle case. Further, we revealed a design principle in the form of an activatory cascade that manifest in the cell cycles studied. Specifically, we derived regulatory networks of two yeast species and the mammalian cell cycle, which resulted in quite a good agreement for the yeast (from 50% to 70% of edges reproduced) and a rather poor performance for the mammalian cell cycle. The negative finding in the latter case was actually a very important lesson showing limitations of our modelling since we used transcriptional dynamics to approximate other types of regulation. It is reassuring that the model works better for unicellular organisms than multicellular ones which have far more levels of regulatory control; the contrary would be suspect!

Finally we showed that our model is applicable for a biologically relevant range of parameters. Nonetheless, we are aware that neglecting chemistry in a biological process may be responsible for some of the features arising in our ensembles, but not in biological networks. Hence, the following future research directions are possible: (i) by forbidding interactions known to be unlikely on biochemical grounds, the resulting essential networks for two yeast species could have even a larger overlap with the wild-type regulatory structures, (ii) another option is to explore dynamics stability [174] of regulatory networks in order to see which of the obtained network topologies are most robust to fluctuations in the gene expression or against a node removal (gene knockout), (iii) one might also think of constraining the system to have not only certain functional capabilities, but also to fulfill other types of criteria (e.g. predefined sensitivity, level of robustness to fluctuations in gene expressions), and identify these constraints that
7. CONCLUSIONS AND OUTLOOKS

bring *in silico* gene regulatory networks closer to their biological counterparts (similar to what was done in [175] for metabolic networks), (iv) lastly, as we indicated in [20], the search for functional biases between experimental network motifs in close-by organisms can be performed.
Chapter 8

Appendices

8.1 Appendix A: Transcription factor search kinetics

In order for transcription factor (TF) to effectively find the target sequence in the DNA, all the three thermodynamical modes of TF molecule are important (c.f. Sec. 2.3; see review paper [152]). The task that need to be solved by TF requires locating a short sequence (consisting of several base pairs) on a DNA that have about $10^5 - 10^6$ bp for simple bacteria and can have up to $10^{11}$ for eukaryotes. Since the double-stranded DNA is typically compacted in a small space (for eukaryotes it is cell’s nucleus), it can be proved that the best way for TF molecule to find a specific binding involves mixture of one and three dimensional diffusion (see [103] for single-molecule level experiments). Particularly, the unspecific attraction causes the TF to be bound to DNA with a finite probability, such that the TF spends about equal amount of time on and off the DNA backbone. In the case of one-dimensional mode the TF molecule, depending on its conformation, can either slide along DNA backbone in a flat energy landscape with energy $E_u$, or diffuse in the landscape $E(\mathbf{s}_j)$ if it is in the conformation of specific binding site. As a result, the TF molecule can sample the low energies of the landscape $E(\mathbf{s}_j)$ while avoiding being trapped between energy barriers.

The process of search kinetics proves to be remarkably fast [104], so the times typical for the TF molecule to locate its target site are on time-scale of seconds (up to one minute). When compared with at least a few minutes that are required for gene transcription and translation into a protein the search times are roughly an order of magnitude smaller. As a consequence, we can quantify the regulatory effect of the TF
related binding site, by calculating the probability of binding site being occupied by the TF molecule at thermodynamical equilibrium.

8.2 Appendix B: Abundance of functional essential networks

Another question of interest concerns the number of distinct functional essential networks. The number of distinct genotypes is of little interest, being trivially enormous since all not essential interactions can be changed at will without affecting the phenotype. First, we find representative essential networks: a cluster analysis of essential networks in the generated ensemble is performed, and the cluster centers are identified with representatives. Let the numbers of such networks be $M$ and define a distance between a pair of them, for example

$$\text{Distance}(A, B) = \sum_{ij} (A_{ij} - B_{ij})^2$$  

(8.1)

where $A_{ij}$ (or $B_{ij}$) is $\pm 1$ for essential interactions and 0 otherwise.

Our question can now be reformulated more precisely: does the number of clusters, considered as a proxy for the number of representative essential networks, saturate at some moderate value as $M$ grows? (It must saturate somewhere, of course, but that may be for very large $M$ values.) To answer this question, we use the modern affinity propagation algorithm [176], where the number of clusters is not pre-assigned but it is determined by the algorithm; the code can be downloaded from www.psi.toronto.edu/index.php?q=affinity propagation (checked on 5th June, 2013). As an illustration, for 4 fixed points we find that the number of clusters grows at large $M$ roughly like $M^{2/3}$ (with a prefactor approximately equal 0.3) and shows no sign of saturation up to at least $M = 4000$. Other values of $k_{FP}$ lead to similar results.

8.3 Appendix C: Results for mammalian cell cycle
8.3 Appendix C: Results for mammalian cell cycle

<table>
<thead>
<tr>
<th></th>
<th>Cdh1</th>
<th>E2F</th>
<th>CycE</th>
<th>CycA</th>
<th>CycB</th>
<th>Cdc20</th>
<th>UbcH10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Table 8.1: The expression profile for mammalian cell cycle [27]. There are \( N = 7 \). The rest of conventions are analogous to Tab. 6.1.

Activatory interactions

<table>
<thead>
<tr>
<th></th>
<th>Cdh1</th>
<th>E2F</th>
<th>CycE</th>
<th>CycA</th>
<th>CycB</th>
<th>Cdc20</th>
<th>UbcH10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cdh1</td>
<td>0.92</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.08</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>E2F</td>
<td>0.95</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>CycE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>CycA</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.99</td>
<td>0.02</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>CycB</td>
<td>0</td>
<td>0.02</td>
<td>0.37</td>
<td>0.91</td>
<td>0.07</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Cdc20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.31</td>
<td>0.83</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>UbcH10</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.04</td>
<td>0.72</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Inhibitory interactions

<table>
<thead>
<tr>
<th></th>
<th>Cdh1</th>
<th>E2F</th>
<th>CycE</th>
<th>CycA</th>
<th>CycB</th>
<th>Cdc20</th>
<th>UbcH10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cdh1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E2F</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CycE</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.32</td>
</tr>
<tr>
<td>CycA</td>
<td>0.79</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.62</td>
</tr>
<tr>
<td>CycB</td>
<td>0.6</td>
<td>0.41</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cdc20</td>
<td>0</td>
<td>0</td>
<td>0.17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>UbcH10</td>
<td>0</td>
<td>0.34</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 8.1: Frequencies of activators (left) and inhibitors (right) for networks produced in the ensemble relevant for the mammalian cell cycle. Array element \((i,j)\) corresponds to the probability of finding a link \(i \leftarrow j\) in the analysed ensemble of gene regulatory networks. Dark grey entries have frequencies higher than light gray entries; a thick solid frame indicates that the link is present in the mammalian cell cycle [27] (see also Fig. 8.2).
Figure 8.2: Left: The wild-type regulatory network for the mammalian cell cycle [27] has 22 links, of which 9 are activating (solid) and 13 inhibitory (dashed). Middle: Regulatory network in the MCMC ensemble with the highest fraction of hits (71% of the essential interactions are present in the wild-type). The network has 12 hits (black) and 5 non-hits (light grey) for a total of 17 essential interactions. For mammalian cell cycle the total number of predicted edges is typically much lower than in the wild-type network. Thus the fraction of hits differs significantly from the fraction of hits in all experimentally found edges, which for this network equals to 55%. Right: Most frequent network in the mammalian ensemble (8.2% of all gene regulatory networks, 8 hits (black) and 5 non-hits (light grey) for a total of 13 essential interactions).
References


REFERENCES


REFERENCES


109
REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


