Topological Properties of Gene Regulatory Networks with Qualitatively Different Gene Expression Dynamics in Spatially Organised Systems

by

Konstantinos Bougioukos

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

UNIVERSITY OF EAST ANGLIA

Faculty of Sciences
School of Computing Sciences

February 2010

UNIVERSITY OF EAST ANGLIA

ABSTRACT

FACULTY OF SCIENCES
SCHOOL OF COMPUTING SCIENCES

Doctor of Philosophy

by Konstantinos Bougioukos

In biological systems, gene expression takes place when genes generate gene products and gene expression levels correspond to concentration levels of these products. Gene expression levels within a single cell are determined by a network of regulatory interactions among genes mediated by gene products. In spatially extended systems consisting of multiple cells gene expression levels within a cell are also affected by gene activity taking place in neighbouring cells. The interplay between spatial interactions among neighbouring cells and the gene regulatory network (GRN) within each cell may qualitatively alter the gene expression dynamics and affects spatially extended essential biological processes such as cell differentiation, pattern formation and morphogenesis.

This thesis dealt with:

1. Computational modelling of the interplay between GRN and cell spatial interactions and simulating the spatially organised gene expression dynamics.

2. Reproduce phenomena of gene expression heterogeneity in a spatially extended system and not in a null model and scoring GRNs according to their capacity to organise these phenomena.

3. Investigate associations between network topological properties of GRNs and the capacity of networks to organise gene expression heterogeneity in spatially extended systems.

Network density is significantly correlated to the GRNs potential to generate heterogeneity in spatially extended systems, small network diameter also constitutes a characteristic of spatial heterogeneity. Several networks that scored for higher spatial heterogeneity, individual element measures such as gene centralities and membership in cycles have correlated with the capacity of spatial heterogeneity.

Initial condition choices exert limited impact on GRNs capacity to organise spatial heterogeneity and it is the network topology together with the parameters specifying gene interactions and properties of gene products that account for the spatial heterogeneity generation.

GRNs with smaller diameters have identified to have greater degree of robustness to initial conditions. The “small world” network phenomenon is associated with the capacity biological gene regulation networks to generate spatial heterogeneity.

Contents

List of Figures

List of Algorithms

Acknowledgements

This thesis would not have been possible without the contribution of many people. In particular, I would like to thank my supervisor Dr Jan T. Kim who offered generous supervision, invaluable assistance, support and guidance from the first day right up to completion, and my secondary supervisor Dr Steven Hayward whose pep talks kept me going when the light at the end of the tunnel seemed no more than a glimmer.

Special thanks also to all my friends from the School of Computing Sciences and in particular, my colleagues from the Computational Biology Lab who cheered me on and believed in me.

I wish to express my gratitude to my family in Greece for their encouragement and support and for keeping me in constant supply of cheap Greek cigarettes throughout my PhD.

And finally, my thoughts are with my wife Muslin who kept me alive while I was writing this thesis, in more ways than one.

To Despina, Evelina, Eleni and Muslin,
the women in my life!
and in memory of my father Symeon

Chapter 1

Introduction

"Verily, very first of all Chaos came into being"
Hesiod, *Theogony* 116–25

Every living entity uses nucleic acid macromolecules (DNA, RNA) to store and transmit information essential to any process that characterise life (energy exchange, respiration, metabolism, development, reproduction). DNA and RNA hold the information that is needed to encode for another class of biological macromolecules, proteins. Proteins are biological polymers, consisting of amino acid chains, that constitute the building blocks of structure and control the functions that characterise living organisms. The particular genetic constitution of a living entity is known as the genotype and the particular biochemical procedures, or morphological characters the phenotype. The total of DNA molecules of an organism makes up its genome and encompasses the majority of the genetic information. There exists certain regions of the genome which are central in terms of the information they contain and are called genes. What characterises genes is that through a process called transcription the nucleotide (DNA) sequence of a gene serves as a template for the formation of an RNA sequence called the messenger RNA (mRNA). The mRNA sequences then, through a process called translation, provide the information for the formation of amino acid sequences that is proteins. The whole process of transcribing DNA to RNA and translating RNA to proteins is also called gene expression. Gene expression has been encapsulated to what is known as the *central dogma* [1] in molecular biology (figure 1.1). The direct link

[1]The central dogma with its additions (DNA and RNA replication, RNA editing, ribozymes and prions) is still a representative scheme of the way that biological information is disseminated.

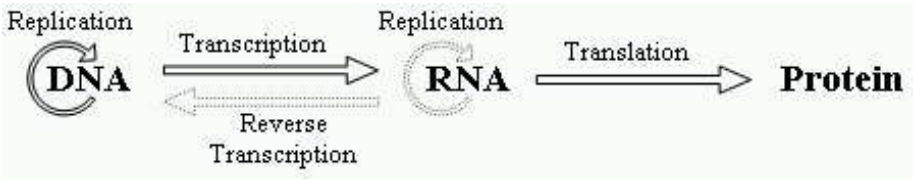


FIGURE 1.1: The central dogma of molecular biology. Arrows depict the flow of information between biological macromolecules. (image taken from http://mlkd.csd.auth.gr/TIS/background.html)

between DNA and proteins implied by the central dogma signifies one more relationship central to biology, that that protein function can be attributed to genes and their expression. Whether of not a protein is present in the cell –in a broad biological sense the function of a protein can be assigned to a particular phenotype– is strictly related to whether or not a gene –which is a part of the genotype– is expressed. Thus the expression of a gene is a prerequisite for the function of the protein encoded by this gene. At least two distinct regions can be identified in a gene. One is the region that gets transcribed to RNA (and consecutively translated to a protein) which is the structural region of the gene. The second is the region, situated usually upstream the structural region, which controls the rate and the time of the expression of a gene and is called the regulatory region. In addition to the regulatory region various other regions, not necessarily situated close or upstream the structural region of a gene, control the amount of gene expression. The information incorporated in the regulatory region together with other regions scattered around the genome constitutes the regulatory information of a gene. The genome contains the full set of all the genes and their respective regulatory information that encode for proteins which build up and carry out all the processes that constitute a living entity.

In multicellular organisms each cell contains a copy of the same genome [2], at the same time cells have a variety of different functions and morphologies. For instance, in animals, a hair cell shares the same genome with a cell in the liver, however the morphology and the functions that each cell sustains are different. These profound differences in cells are the product of the process of cell differentiation. Genes are either expressed or not, or expressed differently in different cells corresponding to the existence of different proteins with different concentrations in different cells leading to different functions. In unicellular organisms there is only one type of cell however unicellular organisms respond to their environment by changing the expression of their genes, e.g. in order to feed or to move towards nutrients or light. The expression of genes is not a static phenomenon distinct cells

[2]This statement neglects somatic mutations which are not the concern of this thesis

express different sets of genes in differential expression levels and even within a single cell that responds to environmental changes numerous signals are realised by differentiating gene expression. Gene expression is regulated by numerous factors.

The regulatory information of a gene determines the expression of the gene. The regulatory region incorporates a significant amount of the regulatory information. A constitutional part of the regulatory region of a gene is the promoter, the promoter contains numerous specific nucleotide sequences that can be recognised by factors that initiate transcription. Transcription of DNA to RNA is carried out by an enzyme called RNA polymerase II (RNApolII). In order for this enzyme to reach the regulatory region of the gene a set of factors (coenzymes, proteins) need to bind in the promoter of the gene and facilitate the binding of the RNApolII. An additional set of factors, i.e. proteins, can identify characteristic sequences in the promoter of the gene and elsewhere in the genome and enhance the activity of the RNApolII increasing the transcription rate of the gene. Conversely there are factors that recognise other specific regions of the regulatory information of a gene and decrease the transcription rate. Most of these factors are proteins and the function of these proteins is to regulate the transcription, these proteins are defined as transcription factors (or *trans*-regulatory elements). The specific regions where these proteins bind in the genome are called *cis*-regulatory elements.

Transcription factors are proteins that are encoded by genes called regulatory genes which have also promoters that other transcription factors can bind on and regulate the expression of regulatory genes and so forth. Therefore there are regulatory genes which control the expression of other genes through the activity of the products the regulatory genes encode for. The regulatory relationships between genes (through the transcription factors which genes are encoding for) can be represented by a network of genes and gene products. These networks are called Gene (or Genetic) Regulatory Networks (GRNs) and comprise genes and gene products, where genes, through their products, are controlling the expression of other genes. In this thesis a GRN is represented by a graph (or a network)[3] where a gene is represented by a node in the network and regulatory interactions between genes by edges[4].

The more activating transcription factor bound on a gene (both different types or more molecules of the same one) the higher the expression of this gene will

[3]The terms graph and network refer to the topological object and the GRN respectively. However, will be treated as topologically equivalent in this thesis and might be used interchangeably

[4]the graph representation is just one of numerous mathematical abstractions that are employed to describe and study GRNs

be, and more repressing transcription factors lead to a decrease and can finally abolish the gene's expression. Thus the arrangement and the nature of the interactions between genes on a GRN are crucial determining factors for gene expression. This arrangement of interactions which is known as the topology (glossary entry: Topology) of the network is a central property that determines gene expression and it is the reason why topological studies of networks have attracted a considerable amount of research the last years ((Barabási and Oltvai, 2004; Bray, 2003) for a general introduction to biological networks studies).

1.1 Gene Expression Regulation in Detail

Each gene encodes a product, either a protein or an RNA molecule (assuming that there is an one-to-one relationship between gene and gene products). The process of generating a gene product begins by the binding of the enzyme RNApolII on the promoter of the regulatory region of a gene. Several factors that are needed for the transcription also bind in the regulatory region, these factors are called general transcription factors and are essential for the assembly of the transcriptional apparatus. In the absence of any other transcription factor, transcription takes place in a relatively low rate, the basal transcription rate. If only these general transcription factors are bound in to the regulatory region the gene is constitutivelly expressed. The general transcription factors are not taken into account when one refers to GRNs however the constitutive gene expression is often incorporated in GRN realisations.

Gene expression is not a static process. A molecule of an activating transcription factor binds to a specific site of the regulatory region of a gene, increases the probability for the RNApolII to bind to the promoter and actively assists the transcription apparatus to move on and transcribe the DNA. After that the transcription factor binding site will remain free and another molecule must bind on it for the transcription to be continuously enhanced. The higher the concentration of the transcription factor the faster the empty *cis*-regulatory position will be occupied again and the faster a transcription event will initiate again. The rate at which the gene will be expressed depends positively on the concentration of the activating transcription factor (respectively for repressing factors the transcription rate depends negatively on the factor concentration).

When a gene is actively transcribed, the rate by which its product is synthesised is increasing and consequently the concentration of the factor that this gene encodes

for is also increasing, the factor concentration is associated to the expression level of this gene. On the contrary when a repressing regulator is bound on the promoter of a gene the rate of transcription is decreased and the factor concentration decreases as a result. Therefore the levels of gene expression are subject to variation in terms of time.

Factors mostly consists of protein molecules, however there are cases where other biological macromolecules like RNA can be the gene products and consequently regulate the expression of genes. Some novel classes of RNA molecules called generally small RNAs are currently under increasing attention from the scientific community for their role as regulatory elements. In addition a plethora of post-transcriptional and post-translational modifications mechanisms contribute to the regulation of the gene expression levels. Gene expression levels can be affected by: regulation in the splicing level (and alternative splicing in higher eukaryotes), post-transcriptional regulation, including microRNAs, post-translational regulation e.g. protein phosphorylations, proteolysis and also protein degradation, to mention a part of the numerous processes that cells are using in order to regulate the concentration of gene products. However this thesis approaches the phenomenon of regulation as a general biological process regardless of the underlying mechanism.

GRNs in this thesis take into account factors that increase the rate by which a gene gets expressed and are called activating factors (or enchancers) and factors that decrease the rate by which a gene gets expressed and are called repressing factors (or inhibitors).

The amount of gene product after a gene gets transcribed is actually an approximation of the concentration of the protein, however the two terms here are used and treated as equivalent. As described, several additional regulatory mechanisms can be involved after a gene gets transcribed and affect a factor's concentration, however transcription regulation is the principal one. Evidence suggests that the mRNA level (the product of transcription that which is predominantly determined by the GRNs) and the degradation rate are essentially the two mechanisms that determine the concentration of a protein (Ben-Tabou de Leon and Davidson, 2009). Furthermore, analysis of collections of microarray experiments has revealed that the gene expression profiles (mRNA levels captured by microarray experiments) characterise much better the state of a cell than protein concentrations levels (Hughes, Marton, Jones, Roberts, Stoughton, Armour, Bennett, Coffey, Dai, He, Kidd, King, Meyer, Slade, Lum, Stepaniants, Shoemaker, Gachotte, Chakraburtty, Simon, Bard, and Friend, 2000).

1.1.1 Importance of transcription regulation

Transcription regulation constitutes an absolutely central process in biological systems. It organises responses to intracellular signals (metabolism), extracellular signals and various environmental stimuli. Transcription regulation also orchestrates the complex phenomena of cell differentiation, morphogenesis and development in multicellular organisms, regulates the cell cycle and maintains the internal state of the cell (homoeostasis) in every single life form. The apparatus that implements transcription regulation is a gene regulatory network.

1.1.2 Gene expression dynamics

Gene expression is a dynamic phenomenon, as the rate a gene is expressed is determined by the concentration of the transcription factors of this gene. In addition to that, the degradation rate of the gene product also affects the factor concentration. These two processes (synthesis and degradation) are shaping the concentrations of gene products in terms of time. So the term gene expression dynamics refers to the variation of factor concentrations in terms of time and is an omnipresent phenomenon in biological systems. Cells alter factor concentrations to respond to changes in the environment, to metabolise different substances and to control internal processes (including cell cycle). Characteristic studies of a case of gene expression dynamics that follow an oscillatory pattern and the role of GRNs in organising the oscillatory behaviour of the circadian clocks can be found in (Locke, Millar, and Turner, 2005; Rand, Shulgin, Salazar, and Millar, 2006).

1.1.3 Gene expression heterogeneity

The phenomenon of cells having different states resulting form different gene expression levels (or different factor concentration levels) is generally called gene expression heterogeneity. Gene expression is a dynamic process, the state of all the factor concentration values in a given time determines the state of the cell at that particular time. Factor concentrations, as described, change over time and consequently the cell state changes over time. States with characteristic different concentration sets may correspond to different cell types.

The difference between cell differentiation and gene expression heterogeneity is that heterogeneity is a prerequisite for cell differentiation, and that not all the

different cell states correspond to a distinct and characterised cell type. Cell differentiation is based in the changes in the cell states and on the gene expression heterogeneity, and consequently on the networks regulating the gene expression. A motivating introduction to the mathematical background of transcriptional regulation, degradation rates and the role of GRNs in cell differentiation can be found in (Ben-Tabou de Leon and Davidson, 2009).

Furthermore, two aspects of gene expression heterogeneity have particular importance for biological systems. First, homoeostasis (Homeostasis) pertains to the property of cells to retain their biochemical stability under limited external perturbations. The term "limited" refers to the limits that the biological organisation of the matter imposes. For instance is impossible for the phenomena (due to constructional and thermodynamical limitations) to occur in temperatures much lower than the freezing point (0^oC) or in temperatures that will break biological membranes and cause the proteins to denature. Second, multistationarity (Multistationarity) refers to the existence of more than one stable state in a dynamical system. Multistationarity in biological systems has been found in bistable switches, memory switches –apparatuses that maintain the response to a transient initial signal stable– and –perhaps the most notable– the cell cycle organisation, where the G1 phase has been suggested to be a bistable switch (Tyson, Chen, and Novak, 2001).

1.2 Gene Regulatory Networks

A gene regulatory network (Gene Regulatory Network), as described, comprises gene and gene products: the gene products regulate the expression of genes and the interactions between genes (mediated by their respective gene products) can be represented by a graph. The vertices of the graph represent genes and the graph edges the regulatory interactions. GRNs represent biological networks known with more specific names, like genetic regulatory networks, gene transcription networks and gene expression networks. All these networks are subclasses of GRNs. Other widely known biological networks like protein-protein interaction networks, neuronal networks, ecological networks and food webs and phylogenetic networks constitute different types of biological networks and are not the subject studied in this thesis.

Two aspects of GRNs are central for understanding the properties of these systems, the one is already described, is the is the arrangment of the interactions between

genes or the topology (Topology) of the graph in graph theoretical terms; and the second is the dynamical parameters of the network (Dynamical Parameters). The dynamical parameters of a GRN consist of a set of real valued parameters that specify the nature and the strength of the regulatory interactions as well as other properties of gene products like the degradation rate and the ability of a gene product to diffuse between different cells.

Therefore, both properties of the topology of GRNs ((Longabaugh and Bolouri, 2006)) and properties of the dynamical parameters of the interactions ((Kim, 2006; Prill, Iglesias, and Levchenko, 2005)) need to be included for any qualitative model of GRNs to simulate gene expression dynamics.

GRNs are organising the complex dynamics of gene expression heterogeneity and cell differentiation. Moreover GRNs, it is suggested that, are behind complex biological phenomena like gene additivity, dominance and epistasis (Omholt, Plahte, Øyehaug, and Xiang, 2000). Phenomena that however important are outside of the scope of study of this thesis.

1.3 Spatial Organisation

Many biological systems are spatially extended. A tissue is a collection of cells that is extended in a spatial structure. The notion of spatial extension in this work is close to the general mathematical definition of space which is a set with an added structure. So in a biological context a spatially extended system is a set of cells with an added spatial structure. Spatially organised cells form tissues and spatially organised tissues form more complex structures like organs and organisms. Gene expression levels can vary along a tissue as neighbouring cells exchange gene products through a number of different physical process (e.g. diffusion, active transport, osmosis). Gene expression heterogeneity induces changes in cells states leading to cell differentiation and that together with spatial organisation constitute a mechanism able to generate more complex biological phenomena like development and morphogenesis. In such spatially extended systems, GRNs (comprising the topology and the dynamical parameters) alone are not sufficient to determine the dynamics of gene expression, as gene expression levels are determined by the process of gene regulation and exchange of factors together.

Numerous biological processes, such as pattern formation, can only take place in spatially extended systems and as such spatial organisation can impact the

dynamical properties of these systems. In a classical example of the hypercycle, a system of linked chemical reaction cycles, Boerlijst and Hogeweg have studied the qualitatively different dynamical properties of a spatial extended version of the system, in terms of the number and the nature of its attractors (Boerlijst and Hogeweg, 1995).

Studying dynamical properties of systems where gene expression dynamics in spatially organised systems are determined by both the exchange of factor concentrations –realised as diffusion– and by the regulatory network is a central objective of this thesis.

1.3.1 GRNs and development

Different cell states may correspond to different cell types and links between GRNs, different cell types and cell differentiation has been establish early by Stuart Kauffman in (Kauffman, 1987) suggesting that different cell states can be associated with stable states of gene expression. Arguably, the fundamental level of understanding developmental processes is the level of cell differentiation and GRNs are able to organise differential gene expression. Within the biological development community this level of organisation is called the developmental program and gene networks and network modules which perform a set basic functions, are considered to be the fundamental building blocks of this developmental program. The time activation of these blocks organises the complex processes behind development including regulatory state maintenance, exclusion of alternative fates, and subcircuit shutoff (Ben-Tabou de Leon and Davidson, 2007).

GRNs have the capacity to control whether a gene will be expressed, at what time window, in which level and at what part of a tissue. GRNs consisting of time or site (or tissue) specific transcription factors are organising the processes required for the development from an undifferentiated embryo to an adult organism with hundreds of different cell types both in animals (Davidson, Rast, Oliveri, Ransick, Calestani, Yuh, Minokawa, Amore, Hinman, Arenas-Mena, Otim, Brown, Livi, Lee, Revilla, Rust, Pan, Schilstra, Clarke, Arnone, Rowen, Cameron, McClay, Hood, and Bolouri, 2002; Stathopoulos and Levine, 2005) as well as plants (Mendoza and Alvarez-Buylla, 1998).

1.4 Modelling and Dynamics

The process of transcription of a gene can be represented by an enzymatic chemical reaction. A transcription factor can be represented by an enzyme and the DNA *cis*-regulatory elements as the substrate of this enzyme. Then using the well established abstractions from the theory of chemical and enzyme kinetics the dynamics of gene transcription can be simulated. Taking also in to account the degradation rate, gene expression dynamics can be simulated in terms of time and incorporating a factor exchange mechanism (diffusion) the dynamics of gene expression can be simulated on spatially extended systems.

Models, as abstract representations of biological systems, are subject to mathematical, ontological and computational constraints. To address these constrains every model should come up with a set of clearly defined assumptions. In this thesis the modelling assumptions are designated as follows:

- Factor concentrations levels calculations are deterministic.
- Only transcription factors are considered as gene products and not other biological macromolecules (e.g. RNAs).
- Gene expression levels are equivalent to factor concentrations.
- Gene expression levels are real valued.
- Time and space are discrete.

1.5 Motivation

This work is related to and motivated by the field of network biology (Barabási and Oltvai, 2004), is also motivated by the role of GRNs in complex biological phenomena (Omholt et al., 2000) and envisages to model and study phenomena that organise developmental processes (Davidson and Levine, 2008; Davidson, McClay, and Hood, 2003). The behaviour of biological systems is not only depending on the parts but on how and on the way its parts are linked together (Motter, Matías, Kurths, and Ott, 2006) and thus the aspect of topology of GRNs and how it is connected with complex dynamical processes in biological systems is central in the course of this thesis.

Cell differentiation is an elementary processes for the organisation and development of multicellular organisms and modelling of cell differentiation has captured the interest of many researchers since quite a while ago. A necessary condition for cell differentiation is spatial heterogeneity of gene expression and mechanisms to generate that include (Turing, 1952) and as studied and extended further in (Gierer and Meinhardt, 1972; Meinhardt, 1982) are is generaly known as Turing-Meinhardt systems. Research of GRNs in development is focused on the reconstruction of the regulatory interactions, on the understanding of the role of topological features of GRNs in biological systems and on how these elements of topology have been shaped by evolution.

The turning point in the motivation of this thesis is the fact that gene expression on spatial systems is determined both by the GRN as well as by the exchange of gene products in neighbouring cells, and that spatial organisation has impacts on the dynamical properties of the systems organised by the GRN. In chemical reaction systems (Boerlijst and Hogeweg, 1995) space lead the hypercycle system to additional attractors, in plant cells (Espinosa-Soto, Padilla-Longoria, and Alvarez-Buylla, 2004) spatial organisation and the direction of diffusion changes the attractors of the system and controls hair root formation and stem hair in plants. Finaly, on developmental biology (Jaeger and Martinez-Arias, 2009) –for a revision of the classical example of positional information– spatial organisation together with gene expression fluctuations generate patterns.

There is an interplay between the network topology and the spatial organisation that coordinates biological processes related to cell differentiation and pattern formation and this work looks forward to systematicaly elucidate it.

1.6 Central Aims of the Thesis

The central aim of the work presented in this thesis is to characterise network topological properties, both whole network properties as well as local network elements properties, of gene regulatory networks that are capable for generating gene expression heterogeneity higher in two-dimensional spatially organised systems than in systems which lack spatial organisation.

This broadly defined central aim of the thesis has informed a set of secondary specific objectives that are summarised as follows:

1. Model and study spatial gene expression heterogeneity phenomena that arise from the interplay of network structure together with the spatial structure. The phenomenon of interest is the emergence of gene expression heterogeneity in spatially organised systems that falls in the category of Turing-Meinhardt systems. Reproduce simulated instances of phenomena of that type.

2. Devise a measure to characterise and quantify gene expression heterogeneity in a spatial extended system and not in systems that lack spatial structure. The measure should take in to account heterogeneity in a spatially organised system compare to a background model of a system that lacks spatial organisation.

3. Characterisation of GRN topologies based on network statistical properties. Employ the measure for spatial gene expression heterogeneity to associate network statistical properties with the capacity of GRNs to exhibit gene expression heterogeneity.

4. Investigate the effects of modifications in the capacity of GRNs to generate spatial heterogeneity. Modifications constitute changes in network topologal characteristics as well as external perturbations. Assess the robustness of GRNs to such modifications.

1.7 Outline

This thesis constitutes one block of work organised in alignment of the central aims and objectives introduced in the above section 1.6. The respective chapters are organised as follows:

Chapter 1 Sets the introduction and motivation of this work in an –as possible– non-technical style.

Chapter 2 More formal and technical motivation and connection of this thesis with the rest of the universe of publications. Review and discussion of the relevant literature and setting of the area where this thesis can be related to. Systematic review of GRN modelling approaches and network sciences theory, tools and advances.

Chapter 3 Formal introduction to the computational modelling framework developed to conduct all the experiments and generate the data.

Chapter 4 Description and motivation of the analytical tools and methodological framework developed to interpret the generated data.

Chapter 5 Presentation and discussion of the experimental results of network properties studies.

Chapter 6 Presentation and discussion of the experimental results of the initial reactor state studies.

Chapter 7 Presentation and discussion of the experimental results of robustness and network pruning studies.

Chapter 8 Conclusions, outlook and potential future research directions.

Chapter 2

Literature Review

"Standing on the Shoulders of Giants"
Bernard of Chartes, c. 12^{th} century

2.1 GRN Modelling

This section will cover a critical review and outline of GRNs' modelling approaches published in the literature. Work that has established the field of modelling of gene regulation and have a significant impact on the network modelling in biology for historical reasons, as well as analytical studies and modelling approaches of specific biological systems that inspired key elements of the work in this thesis, will be presented and critically discussed.

2.1.1 Models lacking spatial structure

The models discussed in the first part of this section have been considered classical both for their level of abstraction, which captures significant properties of regulatory systems, as well as for their pioneering systems perspective that have introduced.

The early models of Stuart Kauffman and René Thomas will be described for historical reasons, as the work on Boolean networks and on logical analysis of gene regulation has paved the way for a systems perspective in biology. Before reviewing individual models it is worthwhile mentioning three books which epitomise the

work of these two researchers and demonstrate the pioneer nature of early models which provide novel approaches in the description of biological systems. The books of Stuart Kauffman "The Origins of Order" (Kauffman, 1993) and "At Home in the Universe" (Kauffman, 1996) and Réne Thomas "Biological Feedback" (Thomas and D'Ari, 1990) have motivated and heavily inspired the work of this thesis.

2.1.1.1 Discrete state space models

In some discrete models, gene expression is considered binary and is assigned either binary (0, 1) or logical values (On // Off, or True // False in a Boolean representation). A gene is represented as active if it is in a 1 (or On) state and as ceased otherwise. The expression of the gene is determined by a Boolean function with input the binary values of its regulators. The initial attempts to model GRNs was as randomly connected networks of genes, as the lack of any large scale data prevented any representation of a specific biological system. In the Random Boolean Network model developed by S. Kauffman (Kauffman, 1969b), a GRN is represented by a randomly constructed network of N genes where each gene has a specified number of K regulators. The number of inputs can either vary among all genes or be a fixed value for each gene in the case of NK networks. There is a potential of 2^{2^K} number of different Boolean functions for a gene with K inputs, Kauffman in the original NK model has assigned one of the potential Boolean functions randomly to each gene. An NK network is generated randomly by two means: the regulatory Boolean function of each gene and the topology of the network that connects genes. The network is placed at an arbitrary state T and the state at time $T + 1$ is calculated after each Boolean function consults its input. The model has one control parameter and this is the number of inputs K per gene. Variation of the K parameter enables the study of the dynamics of various NK networks. Networks with minimum inputs per gene ($K = 1$) have extraordinary long state cycles (the length of a cycle is the time the system needs to reach the same state) and fully connected graphs ($N = K$) a cycle length of 2^N also extremely long for relatively small number of genes N. However networks with K inputs between 2 and 3 have most of their cycle lengths heavily skewed towards small numbers. Moreover, the number of attractors of these networks was approximately equal with the number of different cell types in higher organisms Kauffman (1987). The studies of RBNs by S. Kauffman (also in (Kauffman, 1969a) for an equivalent "continuous" deterministic model) provided for the first time (although influenced by some earlier results of (Walker and Ashby, 1966) on random Boolean networks) a description of biological phenomena based on a high

level statistic of the system the number of regulatory inputs K of a gene. Yet additional striking observations were enabled by the Boolean network abstraction, when the majority of Boolean functions were assigned a certain type –functions called canalising functions (Kauffman, 1974). An NK network with its genes regulation controlled by a randomly chosen canalising function, exhibits a cycle length of $\sqrt{N}$ and the same number of distinct recurrent patterns as well as robustness to random perturbations (homoeostasis) (Kauffman, 1974).

The Boolean formalism has been used extensively in the work of a second researcher, R. Thomas. He has introduced a Boolean approach to model gene regulation, genes have a logical value and the state of a cell is represented by a logical vector. Thomas has employed principles from logical analysis and formal methods to unambiguously represent regulatory systems (Thomas, 1973) and proposed the use of simplification techniques, known to logical analysis, for biological systems. The logical analysis of Thomas lead the way to systematically characterise GRN's behaviour in terms of circuits and make the first attempt to analyse feedback in biological systems. In (Thomas, 1978) a comprehensive logical analysis of numerous feedback mechanisms (e.g. positive feedback loop) can be found and the first attempts for converging to laws of biological circuits are presented. These laws, briefly, that positive feedback is responsible for cell differentiation and negative feedback for homoeostasis, are formalised in (Thomas, 1998) and constitute one of the major contributions of the application of formal methods in modelling biological regulation.

Beyond the Boolean discrete network formalism for GRN modelling, there exists attempts to study and explore the dynamics of more complex events of gene regulation. In cases where the effect of TFs in gene regulation is not additive but additional phenomena are taking place, phenomena like synergy or antagonism in the transcription factor binding. The "logic" behind these phenomena is surveyed in (Schilstra and Nehaniv, 2008) and concludes that the rules for combinatorial logic apply only for the independent binding of TFs, in other cases the behaviour is similar to logical operators and when there is competition for the binding site a whole non-Boolean continuum of behaviours is observed.

Discrete modelling of biological systems. Boolean networks have been extensively used to model gene regulatory networks and simulate the dynamics of biological systems. Two classical examples are presented here the *Arabidopsis*

thaliana flower morphogenesis network and the cell cycle network of the yeast *Saccharomyces cerevisiae*.

Modelling the *A. thaliana* flower morphogenesis: Using a binary approach to gene expression and the Boolean network abstraction, Mendoza and Alvarez-Buylla have modelled the dynamics of the gene regulatory network that controls the flower morphogenesis in *A. thaliana* (Mendoza and Alvarez-Buylla, 1998). The topology of the network comprising 11 genes and 24 regulatory interactions as well as the regulatory strengths of the interactions have been retrieved from the literature and an exhaustive analysis of the network dynamics has reproduced the 4 distinct states of the ABC model for flower morphogenesis, implied as 4 stable attractors in the model. The model has also revealed a 5^{th} attractor that corresponds to the vegetative state and a 6^{th} that is not present in wild type flowers but exists in laboratory strains. The group has moved forward the study of the GRN that underlies the ABC model and a recent review collects all the refinements of their models (Chaos, Aldana, Espinosa-Soto, León, Arroyo, and Alvarez-Buylla, 2006) and an extensive version –including 15 genes and 29 regulatory interactions– of the *A. thaliana* flower morphogenesis GRN.

Modelling the yeast cell cycle: Cell cycle regulation is one of the most well studied biological systems. Based on the accumulated knowledge build up over years of research the network topology of the yeasts' cell cycle key regulators can be mined form the literature. Knowledge of the topology of the network is sufficient to simulate the dynamics of the system in terms of successive stable states of biological activity. Indeed by using a discrete Boolean approach (Li, Long, Lu, Ouyang, and Tang, 2004) and (Davidich and Bornholdt, 2008) have been able to simulate the dynamics of the yeasts *S. cerevisiae* and *Schizosaccharomyces Pombe* cell cycle respectively. The cell cycle network in both organisms has a robust design, with the majority of the initial conditions to be members of the largest basin of attraction of the system which equilibrates to a fixed point attractor corresponding to the G1 control point of the cell cycle. The biological pathway of the cell cycle corresponds to one of the attracting trajectories of the Boolean network dynamical trajectories.

2.1.1.2 Continuous state space models

Factor concentrations, transcription activation and repression coefficients are represented and measured as positive real values, gene expression levels are also measured as continuous real variables, therefore it is natural that continuous state space models of GRNs have been developed to represent and study such systems. Continuous modelling is based on the premise of sets of Ordinary Differential Equations (ODEs) that are coupled and are used to calculate the changes in gene products concentrations with regard to time. Modelling is based on principles of chemical reaction kinetics, with the Michaelis-Menten kinetics (extended by the Hills equations) to be regularly used to derive the functions which simulate gene expression. The systems of ODEs are either solved numerically, by numerical integration in discrete time intervals, or solved analytically. In numerical simulation approaches, a timeseries of simulated gene expression levels is generated by the model and then is subjected to analysis by established methods of gene expression data analysis, (Eisen, Spellman, Brown, and Botstein, 1998) –is a mainstream example of gene expression data analysis. Analytical approaches are focused on finding steady state solutions to the equations, reveal oscillatory dynamics and characterise critical points, these are points where relatively minute perturbations can lead the system to quantitatively different dynamics and potentially correspond to stable cell states in biological systems. The analytical approaches reviewed here have inspired the design of regulatory systems with anticipated dynamics and have also inspired the study of GRN properties, both topological properties and dynamical parameter settings, in this thesis.

Early work to model and analyse the dynamics of regulatory control circuits include the studies by J. Tyson and A. Othmer (Tyson and Othmer, 1978), where a comprehensive analysis of the dynamics of biochemical networks as well as genetic control circuits were modelled as continuous systems. The work studied activating and repressive systems separately and derived formal mathematical conclusions for steady states and local and global stability of GRNs. Continuous modelling of GRNs provides insights that logical models are unable to capture. This advantage is facilitated by the extensive body of dynamical systems analysis tools which are used to study the dynamics of gene regulation. Tyson and Othmer have described invariants of the dynamics of regulatory systems that only continuous modelling can derive including that one unique steady state is asymptotically stable in activating and repressing systems and when 3 steady states exist in activating systems the second one is always unstable. In an equivalent approach Berding ((Berding

and Harbich, 1984)) has modelled the dynamics of the operon by a set of ODEs, as a system itself and also as part of a cascading pathway. The dynamic analysis include the calculation of the Lyapounov exponents of the system for a range of different dynamical parameters and constitutes an early analytical study that the feedback loop is able to express a variety of different dynamics depending on different parameter settings for the regulatory strengths.

A recurrent topic in continuous modelling of GRNs is the application of dynamical systems analysis to relate invariant sets (such as equilibrium points and periodic orbits) to biological questions relating gene regulatory mechanisms. Analytical studies of the dynamics of gene regulation with respect to the regulatory elements organisation have shown that the number and the stability of equilibria relates to the number of binding sites of a transcription factor (TF) in the regulatory region of a gene (Wolf and Eeckman, 1998). This work attempts an early connection between structure of regulatory networks (the number of *cis*-regulatory sites) and the dynamics of gene expression. The authors have drawn a theoretical conclusion for the minimal mechanism that exhibit an "on-off" switching behaviour, that is a two gene and two binding sites per gene system where one gene acts as a switch for the other, and suggested that this might be a constituent part of networks controlling cell differentiation and development.

The concept of gene switches was analytically studied by Cherry and Adler (Cherry and Adler, 2000)) as a "flip-flop" switch system. A two genes system that has two stable states one where the first gene in "on" and the second "off" and another in which the states are reversed. This work has introduced a functional to characterise the shape of functions that are able to give rise to "flip-flop" phenomena, certain criteria regarding the dynamical parameters should be met for a system to act as a switch. Functions based in Michaelis-Menten type of repression alone can not generate a switch-like behaviour in a two genes system, but functions incorporate cooperativity, effects from multiple binding sites (i.e. Hills coefficient higher than one) or depletion of the repressor should be employed such that a system will exhibit a "flip-flop" behaviour.

Stochastic modelling: Continuous state space modelling by using systems of coupled ODEs provides a realistic representation of most gene regulatory systems. However there are cases where the phenomena that control a gene's regulation appeared to have a random and infrequent nature. Small number of regulatory

molecules (transcription factors, RNApolII) and the random intervals of transcription initiation events result in a considerable degree of biological noise that any ODE approach can not capture due to its deterministic nature. Therefore, modelling approaches based on stochastic differential equations (SDEs) have developed with (McAdams and Arkin, 1997) to be one of the earlier and most characteristic ones. In this study the authors model a single gene where the time interval for transcription initiation events was random. The stochasticity of the system incur significant differences in temporal mode of gene expression. The random expression of factors can lead to probabilistic behaviour of regulatory switches and thus generate different cell types. The random nature of gene regulation can generate diversity on gene expression by non genetic means, a stochastic type of regulation.

Continuous modelling of biological systems. Continuous models have employed to study numerous biological systems, as a comparison example with the discrete approach discussed before (section 2.1.1.1) the *A. thaliana* root and leaf hair development modelling and the yeast cell cycle modelling based on continuous models will be discussed here.

Modelling the *A. thaliana* root and leaf hair: The *A. thaliana* root and leaf hair development has been modelled in an activator / inhibitor continuous model by the same group that model the flower morphogenesis in the same plant (Benítez, Espinosa-Soto, Padilla-Longoria, Díaz, and Alvarez-Buylla, 2007) (discussed in section 2.1.1.1). The pattern generated by the continuous model was in agreement with patterns generated by the logical equivalent of the model, supporting the conjecture that stable states found by logical models are always present in the equivalent continuous model (a further discussion on discrete-continuous modelling comparisons follows in section 2.1.1.3).

Modelling the yest cell cycle: One of the major applications of continuous modelling of GRNs is the yeast cell cycle analysis using tools from dynamical systems. In the work of J. Tyson and B. Novak the dynamics of one of the most well studied physiological systems of the cell –the cell cycle– were analysed in terms of networks and dynamical systems properties ((Tyson et al., 2001) for a review on the work of the group). Critical points in the yeast cell cycle are characterised as steady states of the dynamical system and bifurcation analysis reveals that the G1 control point is a bistable switch. These results are in an extent in agreement with the discrete modelling of the yeast cell cycle that were discussed in section 2.1.1.1.

Results from the application of continuous state space modelling, which is routinely accompanied by analytical studies of the invariant sets of the underlying dynamical systems, demonstrate that dynamical systems analysis when is coupled with elementary concepts of network theory (at least for relatively small networks) can pave the way for a unified theory for modular cell physiology according to (Hartwell, Hopfield, Leibler, and Murray, 1999)

Software packages Here software suites which are based on continuous modelling and computational simulations of the dynamics of gene expression are reviewed. The packages are composed of a computational representation of a gene regulatory systems and a numerical simulator of the dynamics of the represented networks.

The group of Pedro Mendes has developed two software packages, Gepasi to model and simulate gene expression (Mendes, 1997) and Copasi to simulate complex pathways and parameter optimisation (Hoops, Sahle, Gauges, Lee, Pahle, Simus, Singhal, Xu, Mendes, and Kummer, 2006). The underlying models of the Mendes group software are presented in Mendes, Sha, and Ye (2003) and consist of random network generation algorithms to produce GRN topologies and a set of ODEs for reaction kinetics, incorporating the Hill's coefficients (Gepasi also includes SDE modelling capabilities). The system design requirements were focused to facilitate topological studies of GRNs as well as studies of their dynamical parameters.

An artificial gene expression data generation software named SynTReN has been developed by T. van de Bulcke et. al. (van den Bulcke, van Leemput, Naudts, van Remortel, Ma, Verschoren, de Moor, and Marchal, 2006). SynTReN uses a sampling from biological networks approach to generate different GRN topologies, then assigns a regulatory function to each interaction and calculates the systems steady state, it needs to be pointed out that SynTReN calculates directly the steady state gene expression of a GRN as it accepts only acyclic graphs as network topologies and does not calculate any dynamics. However SynTReN has been successfully used to assess the accuracy of several GRN reconstruction algorithms.

The BioComputing group in the University of Hertfordshire has developed a software package able to represent and simulate the dynamics of continuous and discrete GRN models. The package is called NetBuilder, is reviewed in Titus Brown, Rust, Clarke, Pan, Schilstra, De Buysscher, Griffin, Wold, Cameron, Davidson, and Bolouri (2002), it comprises of a Petri-Net approach to model gene regulatory

systems and can simulate the dynamics of GRNs using both deterministic as well as stochastic equations sets.

2.1.1.3 Comparisons between discrete and continuous models

Here approaches which incorporate discrete together with continuous systems will be discussed, in the contet of the studies of (Kappler, Edwards, and Glass, 2003) which have pointed out discrete systems are able to predict the number of attractors of continuous systems, for GRNs of relatively small size.

The central point in comparing the Boolean network approaches to continuous modelling is the ability of the model to adequately capture the dynamical properties of the biological system. A straightforward remark is that in the lack of detailed knowledge for the parameters that control the strengths of gene regulation for relatively large systems the Boolean formalism becomes a favourable way to study the behaviour of biological systems. Indeed this is valid if one considers that the first attempts were discrete models (Kauffman, 1969b; Thomas, 1973). However, as the knowledge of biological systems become more detailed continuous models have gradually started to develop, especially for relatively small and exhaustively studied systems (e.g. λ-phage, lactose operon). Continuous modelling is apparently more biologically realistic as the measured quantities in biological systems are taking continuous values. Moreover, continuous modelling offers a competitive advantage, that is it can capture the full spectrum of dynamics that otherwise is lost to 0, 1 or On, Off in discrete modelling. However, analytical studies of systems modelled by both a discrete and a continuous approach (Glass and Kauffman, 1973), tried to analyse continuous models by their logical equivalents have shown a degree of agreement between the two modelling approaches. Effectively, every stable state in a logical (Boolean) models corresponds to an attractor in the continuous equivalent and every transient to transitions in the logical system, e.g. oscillations will correspond to cycles in the logical mapping. Analysis of the dynamics of discrete systems in terms of equivalence with the continuous systems suggests that all the steady states of a discrete system qualitatively correspond to steady states of a continuous system but not the opposite.

These results can be summarised in the following two individual publications dealing with equivalent systems both continuous and logical.

A theoretical study by R. Thomas in both continuous (Thomas and Kaufman, 2001a) and discrete (Thomas and Kaufman, 2001b) systems with time delays suggested that there exists qualitatively similarity between continuous and discrete approaches regarding the laws of regulatory circuits (as they appeared first in (Thomas, 1973) and are discussed in section 2.1.1.1). In (Thomas and Kaufman, 2001a) the concepts of a full (a circuit that takes into account all the variables of the system) and an ambiguous (a circuit that its sign depends on the location in the state space) circuit were introduced. A formal mathematical survey of the dynamics explored the requirements for multistationarity, periodicity and deterministic chaos. Subsequently, in (Thomas and Kaufman, 2001b) groups of logical parameter settings for GRNs were shown to have qualitatively equivalent dynamics with the continuous approach above. Note also that most of the work of Thomas has been theoretically corroborated further and proved by the studies of Christoph Soulé, including formal requirements for multistationarity (Soulé, 2006) and (Soulé, 2003) which is a proof that negative circuits is a sufficient condition for multistationarity.

Furthermore, a motivating review to the discussion of comparisons between Boolean and continuous regulatory network models can be found (Hasty, McMillen, Isaacs, and Collins, 2001). A more exhaustive review of various different methods of the discrete and the continuous approaches as well as for an introduction to methods that combine modelling elements from both the approaches (e.g. piecewise differential equations) is published by (de Jong, Gouzé, Hernandez, Page, Sari, and Geiselmann, 2004). Finally, in (Smolen, Baxter, and Byrne, 2000) a formal mathematical account of the differences between the Boolean and continuous approaches to modelling is rigorously explored.

2.1.2 Models including spatial component

All the models reviewed so far were referring to modelling GRNs and gene expression dynamics in individual cells or were pertaining to averages of gene expression along tissues. However, numerous biological processes in multicellular organisms, such as cell differentiation and morphogenesis, are taking place in collections of cells that are spatially organised and where genes are expressed differently in different cells. Therefore models that take into consideration the notion of space have been developed to model such biological systems. Furthermore, as the processes taking place during the development of an organism shape the mapping from genotype to phenotype, modelling and understanding developmental processes can shed

light on determining the relation between genotype and phenotype (Solé, Salazar-Ciudad, and Newman, 2000).

The process of pattern formation was the first that attracted the interest in modelling. Which mechanisms are able to reproduce the phenomenon where a collection of cells organised in a 2-dimensional sheet and having minute differences in their factor concentration can generate patterns. Allan Turing has been motivated by this problem and he was the first that introduced a mathematical model of partial differential equations able to generate patterns in a 2-dimensional space (Turing, 1952). The model was based on the principle of differences in the diffusion of two molecules. An inhibitor molecule could diffuse an order of magnitude faster than an activator molecule and this disproportion between activation and repression was the generating factor of patterns in a 2-dimensional space. The lateral activation global inhibition principle was further extended and examined by Hans Meinhardt and Alfred Gierer, who established a mathematical framework of pattern formation mechanisms (Gierer and Meinhardt, 1972) and connected their theoretical work with aspects of developmental biology and GRNs (Meinhardt, 2006).

An innovative approach for spatial modelling of biological systems was introduced by Franco Bignone who was the first who introduced the concept of discrete orthogonal 2-dimensional lattices in order to model cells spatial organisation. His work, (Bignone, 1993), has incorporated gene regulation and diffusion to simulate gene expression dynamics and it has been very inspirational for the development of the spatial models in this thesis.

Studying the mechanisms that give rise to developmental phenomena has been motivating for researchers from the artificial neural network community. A cell interaction model, which combines chemical, electrical, cellular and genetic interactions to model development has been developed by Kurt Fleischer (Fleischer and Barr, 1993). The open development of the computational framework of this model provides a testbed for the study of numerous mechanisms of cell differentiation, pattern formation and multicellular development (e.g. genetic coupled with physical interaction between cells). A variety of results from experimentations with the model are reported in (Fleischer, 1996). Although the core motivation behind this work (as well as the work of (Geard and Wiles, 2005) to model cell lineages in *Caenorhabditis elegans*) were to develop better models of artificial neural networks which imitate biological developmental networks and are capable of solving

problems in perception and control, some fundamental insights and principles of biological development have been examined from their analyses.

2.1.2.1 GRNs in developmental biology

Gene regulatory networks constitute the key mechanism to orchestrate the complex processes that control biological development, as introduced in section 1.3.1. Consecutively, modelling developmental processes has always attracted considerable efforts from the modelling scientific communities. In an insightful approach (von Dassow, Meir, Munro, and Odell, 2000) have modelled the segment polarity GRN in the *D. melanogaster* embryo development using non-linear ODEs. The modelling approach included the reconstruction of the network topology from the literature, the assignment of a non-linear ODE to each interaction and the inclusion of a spatial component as a string of cells with periodic boundary conditions. The insightful findings of this study were that the segment polarity GRN has found to be, after an extensive search of the parameters spaces, a robust network in terms of the dynamical parameters choices and in terms of the modelling initial conditions. Results that have motivated a series of experiments in this thesis.

The majority of the work in developmental GRNs is conducted by collaborations between biological laboratories which have a strong interest in deciphering the networks of gene regulation of one specific biological system with computational groups. Arguably the most comprehensive work in developmental GRNs in animals has been carried out by the group of Eric Davidson for the sea urchin *Strongylocentrotus purpuratus* embryonic development GRN. In an approach by incorporating computational modelling (using continuous ODE modelling) of the GRNs dynamical properties and data integration from multiple sources (proteomics, transcriptomics), the group has reconstructed and modelled the dynamics of the GRN that controls the specification of the endoderm and the mesoderm in the sea urchin embryonic development. The need for high quality computational models has lead to the development of a software suite for modelling the dynamics of developmental networks in animals (Longabaugh, Davidson, and Bolouri, 2005). The close collaboration of computational modelling with bioscientists can improve our understanding of biological processes that span various levels of complexity such as the evolution of development (evo-devo), see (Davidson and Erwin, 2006) for a characteristic key work on the subject.

Equally comprehensive work with animals has been conducted by the group of Elena Alvarez-Buylla in modelling plants developmental GRNs using *A. thaliana*

as a model organism. The group contributed to the extension of the well studied ABC model for flower morphogenesis (Coen and Meyerowitz, 1991) and proposed an extended gene network for the ABC flower morphogenesis molecular mechanism (Espinosa-Soto et al., 2004), which was based in previous computational and logical analysis of the dynamics of the flowering network (Mendoza, Thieffry, and Alvarez-Buylla, 1999). The concept of the meta-GRN, a network of gene regulatory networks which are connected together via diffusion of proteins between neighbouring cells, has been introduced as a means to model phenomena which combine space. Simulating the dynamics of this meta-GRN has revealed mechanisms that control the hair morphogenesis in stems (Benítez, Espinosa-Soto, Padilla-Longoria, and Alvarez-Buylla, 2008) and also the equivalent network for hair morphogenesis in plant roots (Benítez et al., 2007), where a contrasting pattern between stem and root hairs can be generated by equivalent networks and that a spatial parameter (the cell space) can affect the patterns on the roots.

Moving form the 2-dimensional modelling to the 3rd dimension a comprehensive model of continuous ODEs embedded in a 3-dimensional lattice has been used to model differentiation in gene expression levels in the shoot apical meristem (SAM) of a system of morphogens that control the development of the SAM (Jönsson, Heisler, Reddy, Agrawal, Gor, Shapiro, Mjolsness, and Meyerowitz, 2005).

Evo - Devo The models of this section, apart from including a representation of space, introduce a further component in the developmental aspect of modelling that of a model of biological evolution.

Using the lattice modelling abstraction proposed by (Bignone, 1993) a model of a lattice with periodic boundaries has been developed by (Keränen, 2004). The lattice represented an early embryo and the motivation was to study the complexity of embryo cell differentiation. A set of ODEs were used for simulating gene expression data on a discrete toroidal lattice. The study links the differentiation patterns with gene connectivity and gene interaction strengths. The complexity (in terms of differential gene expression) of the patterns and the complexity of the network (both in terms of topology and dynamical parameters) are explored and the implications of evolution in the mode and the increase of network complexity are studied. The topologies of GRN were the focus of the study by Salazar-Ciudad (Salazar-Ciudad, Garcia-Fernandez, and Sole, 2000), topologies that are capable for pattern formation in a reaction-diffusion model. The model comprises ODEs to simulate gene expression data and a string of cells as the spatial pattern, an

evolutionary algorithm has been employed to search the topologies' space for candidates that generate patterns on the string of cells and the study proposed a set of converged topologies of small (2-3 genes) to medium size (7-9 genes) as candidates for generating spatial patterns. The work developed in the last two papers has motivated the development of the spatial models and the optimisation approach in this thesis.

In line with the last model, using again a string of cells as the spatial representation (Munteanu and Solé, 2008) have been able to exhaustively search the whole space of topologies of a relatively small (2 genes and 2 hormones) system of a GRN that controls the stripe formation in the *Drosophila melanogaster* embryo. They were able to identify stages of the stripe patterns of the Drosophila embryo in the string and derive some conclusions about evolutionary neutrality and robustness of the *D. melanogaster* stripe formation GRN, in accordance with von Dassow et al. (2000).

Arguably one of the most well known abstractions in the evolution of development is the concept of the "French flag" that L. Wolpert has introduced in (Wolpert, 1969) and it represents three different cell states, like the three different colours in the flag, which are specified by a morphogen gradient. Since then, numerous computational models have had as an objective to reproduce this pattern, using different computational approaches. Two representative papers that use an evolutionary algorithm approach to optimise network topologies so that the system can reproduce the desired pattern of a "French flag", one uses a cellular automata to realise space (Chavoya and Duthen, 2008) and the second a 2-dimensional cellular potts model (Knabe, Nehaniv, and Schilstra, 2008b).

2.1.3 The transsys framework

The computational background upon which all this work is developed is *transsys*. Transsys is a computational framework developed to comprehensively represent GRNs and simulate the gene expression dynamics organised by the network. The transsys software consists of the transsys language, a formal language to unambiguously describe GRNs, a facility for computational simulation of gene expression dynamics and a collection of various tools for analysing, visualising and developing specific regulatory network models. Transsys had been initially used to model a gene regulatory network for flower morphogenesis (the ABC model (Coen and Meyerowitz, 1991)) by (Kim, 2001).

Two constructs are central in transsys, the *transsys program* containing a set of transsys language instructions to represent a GRN and the *transsys instance* of a transsys program containing the gene expression state vector. A concise presentation of the key elements of the framework follows.

2.1.3.1 The transsys language

The transsys language is a formal language for the representation of GRNs. A set of valid statements from the transsys language formal specifications constitute a *transsys program*. A transsys program represents a GRN. Conceptually, two biological entities of regulatory networks are central -and thus present- to any GRN representation the gene and the product of the gene (or factor) and these two constitute the two central transsys language elements. A transsys program contains the declarations of gene definitions and factor (gene product) definitions that comprise a GRN.

transsys Factor A factor (or a gene product, which can be protein, RNA, etc.) is specified within a transsys program by the word `factor` immediately followed by the name of the factor (technically a transsys identifier). The body of a factor declaration consists of one block containing the decay and diffusibility expressions. The decay expression represents the rate of degradation of a factor as the aggregated result of many biochemical processes. Decay rate is denoted by the keyword `decay` followed by an expression. The expression can either be simple (e.g. a real number representing the decay rate) or complex (i.e. involving interactions with other gene products). The factors diffusibility is denoted by the keyword `diffusibility` and followed by a real value parameter. The diffusibility represents a general ability of a factor to diffuse and it can be used to implement various different diffusion models.

transsys Gene Genes are fundamental units of genetic information and following a straightforward biological representation are partitioned into the regulatory part and the structural part (section 1.1). The structural part encodes for a gene product (which can be linked with a specific biological activity, that is a functional molecule like an RNA or a protein). The regulatory part –a component of the regulatory information of the gene– determines the expression rate of the gene by *cis*-acting elements placed in the promoter of the gene upstream the structural part (however *cis*-elements of the regulatory information of a gene are scattered

around the genome) which interact with a class of proteins called transcription factors (*trans*-elements).

In transsys a gene declaration begins with the keyword `gene` followed by the gene's name. The regulatory / structural partitioning is represented by the `promoter` block and the `product` block respectively. The product block contains the specification of the factor which this gene encodes for, that is the name of the factor and its type (currently only `default` type corresponding to a protein is implemented). The promoter block comprise a list of promoter elements which describes the transcriptional conditions of the gene, the elements are of three types constitutive, activating and repressing. The constitutive promoter element represents the basal transcriptional activity of a gene and specifies the constitutive expression, which can be a real number representing the basal amount of gene product concentration or a complex expression (including interactions with other factors). The activation / repression promoter elements are representing the regulatory interactions of transcription factors that bind to the promoter. Declaration of every activation or repression element includes the keyword `activation` or `repression` respectively, preceded by the name of the factor that is regulating the gene and followed by a list of two expressions as arguments. The arguments determine the kinetic parameters or the regulation. The first specifies the binding specificity of the regulating factor with the element α_{spec} and the second the maximal rate of regulation that this element can cause a_{max}. α_{spec} and a_{max} are analogous to the Michaelis-Menten chemical kinetics parameters K_M and v_{max} respectively. Both the parameters are specified by transsys expressions which allows apart from simple designation of a real value the modelling of more complex phenomena such as protein-protein interactions, where a protein can control the activity of another protein.

transsys Expressions In the expression parts of the statements of transsys genes and factors complex expressions are allowed apart from real numbers. Transsys expressions are designed to be similar to that of standard programming languages (like C/C++ or Java), thus the transsys arithmetic and logical operators are identical with those of most of the standard programming languages. With one exception, the usage of identifiers, transsys identifiers in transsys expressions refer to factor concentrations C_{factor}. Complex expression statements can be used to model interactions that can not be represented by a simple regulatory interaction concept (e.g. one transcription factor binds in one cis-regulatory element). Complex transsys statements are mentioned here to give a complete account of the

transsys language specifications however they have not been used in the models presented in this work.

transsys Components A transsys program can be divided into two major components, the network topology and the dynamical parameters. A transsys program network topology comprises the set of genes and the set of all the regulatory interactions between factors and genes, that is the transsys promoter elements for activating and repressing (sec. 2.1.3.1). The network topology can straightforwardly be represented by a graph where the set of transsys program genes $\mathcal{G}$ is the vertex set and each regulatory interaction between a factor and a gene forms an equivalent graph edge from the factor encoding gene to the regulating gene. Figure 2.1 illustrates the topology of an example GRN represented by a transsys program.

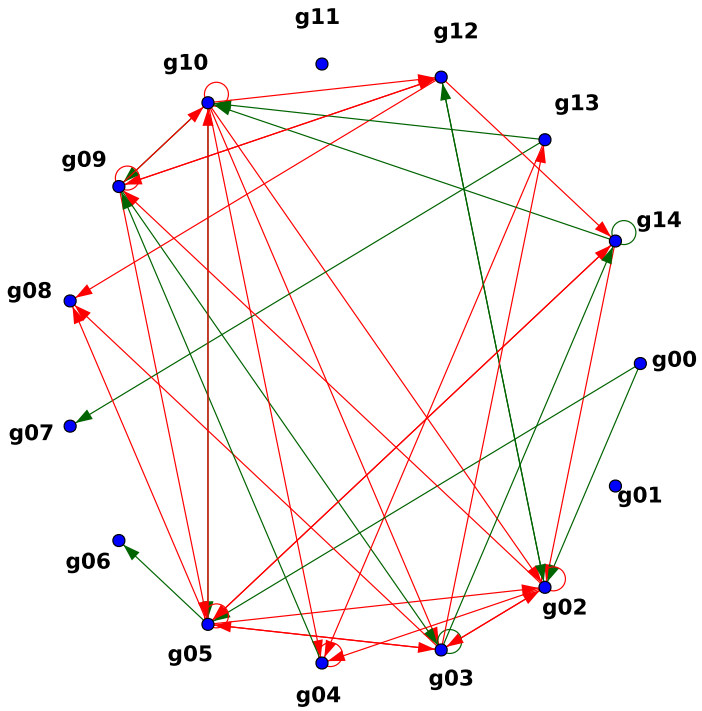


FIGURE 2.1: Illustration of a GRN topology. The arrows show the direction of the regulation, (base of the arrow at the regulatory factor, tip of the arrow indicates the regulating factor). Activating interactions are depicted in green and repressing in red.

The transsys program dynamical parameters are all the real number transsys expressions that quantitatively describe the properties of the genes and factors of a transsys program. The set of all the decay rate, diffusibility, constitutive, $a_{\max}$, and α_{spec} parameters for all the factors and genes of a transsys program consists

the dynamical parameter set. A demonstration of the lexical structure of transsys and an example of a dynamical parameter set can be found in the appendix A.

2.1.3.2 transsys instance

The second fundamental construct of the transsys framework is the *transsys instance*. A transsys instance can be generated once a transsys program is declared. For a transsys program P specifying a set of factors $\mathcal{F}$, a transsys instance p holds the following information: The list of the concentrations of all factors $f \in \mathcal{F}$, this list represents the state of the instance p, is also referred as the sate vector in dynamical systems, and the transsys program.

A transsys instance (in an Object Oriented programming language analogy) has the same relationship with the transsys program as an instance of a class has with a class and the factor concentrations may be considered as the member variables of the transsys instance. A transsys instance provides an *update method* which takes the instance at time t and using information from the transsys program computes a transsys instance at time $t + 1$, thus simulating gene expression dynamics.

2.1.3.3 Simulating gene expression dynamics

The simulation of artificial gene expression, in a biological analogy, receives information from the regulatory part of the gene which is the promoter block in transsys. Each of the three possible different types of promoter elements described in section 2.1.3.1 are contributing to the expression of a gene, let q_i denote the contribution of an individual promoter element i. Thus for a promoter element the amount of gene expression for each type will be:

Constitutive Constitutive is a generic type of promoter element represent the basal transcriptional activity of the promoter. The evaluation of the constitutive expression determines the rate q_i at which the gene product will be synthesised. Thus the contribution of a constitutive promoter element to the synthesis of a factor is:

$$q_i = \text{result of evaluating expression}$$

Activation Activation promoter elements are preceded by a factor name f followed a list of two expressions as arguments. The arguments represent the

α_{spec} and the a_{max} parameters respectively as described in section 2.1.3.1. The synthesis is calculated according to the Michaelis-Menten equation for chemical reaction kinetics. The contribution of an activating promoter element into the rate of synthesis q_i of a product U relative to the concentration of the regulating factor C_f is given by the formula:

$$q_i = \frac{a_{\text{max}} \cdot C_f}{\alpha_{\text{spec}} + C_f}$$

Repression For a repression promoter element the same equation applies however with a minus symbol representing the negative impact that repression has in the rate of synthesis of a product U, thus:

$$q_i = -\frac{a_{\text{max}} \cdot C_f}{\alpha_{\text{spec}} + C_f}$$

For a promoter of a gene g consisting of a set $\mathcal{I}$ of different promoter elements the total contribution of all the promoter elements of g that affect the rate of synthesis of the product U that is synthesised by the expression of this gene in a given timestep t is:

$$\Delta_g C_U(t) = \begin{cases} q_{\text{total}} := \sum_{i \in \mathcal{I}} q_i & \text{if} \quad q_{\text{total}} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

The overall change in the concentration of the product U that is encoded by a set of genes $\mathcal{E}_{\mathcal{U}}$ in this particular timestep t constitutes the first step to the calculation of the gene expression level, thus:

$$\Delta C_U(t) = \left(\sum_{g \in E_U} \Delta_g C_U(t) \right) - r_U(t) C_U(t) \tag{2.2}$$

Where the term r_U denotes the decay rate of the product U.

Having calculated the change in the concentration of product U at timestep t from 2.2, the concentration in the next timestep will be given by the equation:

$$C_U(t+1) = C_U(t) + \Delta C_U(t) \tag{2.3}$$

Equation 2.3 constitutes the update function which calculates factor concentrations for the next time-step (i.e. gene expression dynamics).

For all factors belong to the factor set $\mathcal{F}$ of a transsys program P the above equation 2.3 can be written using set notation for all products $U \in \mathcal{F}$:

$$C_{\mathcal{F}}(t+1) = C_{\mathcal{F}}(t) + \Delta C_{\mathcal{F}}(t) \tag{2.4}$$

The term $C_{\mathcal{F}}(t)$ represents the set of factor concentrations of a transsys instance of the state of the transsys instance at time t as described in section 2.1.3.2. The equations described in this section (sec. 2.1.3.3) constitute the mathematical representation of the update function as described in section 2.1.3.2, which takes the state of a transsys instance in time t ($C_{\mathcal{F}}(t)$) computes the expression of all the genes and returns the state of the transsys instance at time $t+1$, that is $C_{\mathcal{F}}(t+1)$.

To conclude, the transsys framework has been used as the main modelling software for a set of publications including: (Bouyioukos and Kim, 2009; Kim, 2001,0,0; Repsilber and Kim, 2003). For further details, a user manual and a copy of the current version of transsys one can visit (Kim, 2009)

2.1.4 Further Reading

For a more comprehensive coverage of the numerous approaches in GRN modelling the following reviews of different categories of modelling are characteristic.

The most recent review paper in computational modelling (Karlebach and Shamir, 2008) is a major and recent review, focused on computational methods for modelling GRNs a comprehensive supplement of tables is reviewing current computational tools for modelling. In another review (Smolen et al., 2000) are presenting studies motivated by the modelling of specific biological systems to highlight the need for further analytical and computational studies of genetic regulatory systems in parallel with the experimental ones.

A review that connects the fields of evolutionary biology with systems biology through computational modelling (Loewe, 2009), contains a summary of the strong and weak points of computational modelling in general (without references to individual models). The insight of this review is that is focused on describing a framework for systems biology and evolutionary biology crosstalk. The work is more focused on the evolutionary question and hierarchical organisation, however

is engaging is a very good discussion on the benefits of modelling. What are the pros and cons of abstract models how one balances between abstraction and realistic representation, which are the connections with theoretical evolutionary questions are all aspects that are answered by this review.

Hidde de Jong in (de Jong, 2002) has a comprehensive review of most of the available mathematical approaches to model gene regulatory systems. The review provides a table to categorise mathematical model following similar principles to the one used for the categorisation of the models in this thesis.

A comprehensive comparison between the Boolean and the continuous approaches to model networks as well as analyses of the role of time-delays and expression noise in qualitatively changing the dynamics of gene networks are reviewed in (Smolen et al., 2000).

2.2 Networks

Based on the fundamental work on graph theory and discrete mathematics an explosion of the studies of networks and dynamical systems represented by networks has been seen the past decade. These advances, spanning among sciences, arts and humanities and social sciences disciplines have characterised as the "new science of networks" by some of the most cited researchers in the field (Barabási, 2003; Watts, 2004a,0). The "new" to the new science of networks is justified in (Newman, Barabasi, and Watts, 2006) as:

1. The fact that deals both with networks constructed from observations (real world networks) as well as with the underlying theory
2. Networks are not static but an (explicit or implicit) dynamical procedure alters their topologies.
3. It aims not only to study networks as topological object but to understand principles of dynamical systems that can be represented by networks.

A network is a high level mathematical abstraction that can be used to represent a vast number of phenomena and elements of the physical world. Therefore, here after setting up the background and discussing some of the central publications in network sciences, the focus will concentrate on network studies in gene expression

regulation and more precisely in studies of network topology and its relationship with function in biological systems.

2.2.1 Network generation

The traditional model to describe the topology of a network is the random graph theory of Paul Erdös and Alfred Rényi (ER Network) (Erdös and Rényi, 1959), where a network is generated by assigning edges to a node according to a predetermined fixed probability p (for a full description (Bollobás, 1985)). According to the ER random network model the distribution of the degree (the number of edges that are connected to a vertex) is expected to follow a Poisson distribution when the number of nodes n tends to infinity and the average degree $\langle k \rangle$ to be $\langle k \rangle = n \cdot p$. The ER model has provided a random mechanism to generate networks with certain topological features. The huge data acquisition of several modern large scale projects (like genome projects in biosciences, or the fast expansion of the Internet) has made possible the study of topological properties of networks generated by natural processes. What the first studies of real world networks revealed was that the degree distribution was characterised by a fat tail and indeed follow a power low distribution instead of a Poisson. A power law characterises phenomena that lack a characteristic size (or scale) and thus the term scale-free is also used to characterise the topology of several complex networks. Scale-free topologies were observed in the routers connecting WWW servers in (Faloutsos, Faloutsos, and Faloutsos, 1999) and in actor collaboration, power grid and the *C. elegans* neural network data in (Barabási and Albert, 1999). The scale-free property of network degree distribution has been suggested to be the result of a procedure called preferential attachment, where a vertex acquires new edges with a probability proportional to its current degree. The preferential attachment is the generation mechanism of a random graph model that is able to reproduce a power-law degree distribution, the model is also known by the initials of the authors of the publication (Barabási and Albert, 1999) (the BA model). These studies revealed properties that systems with many interacting parts have in common regardless of the background generating mechanisms.

In addition to the power-law degree distribution characteristic, another topological principle of complex networks has been discovered at the same period. That many biological, technological and social networks have small paths like random graphs yet are highly clustered like regular lattices. By using a rewire mechanism (Watts and Strogatz, 1998) were able to generate networks lying between regular and

random and they introduce the clustering coefficient as a measure to quantify clustering in networks (Watts and Strogatz, 1998, figure 2). The "small world" phenomenon –named after the famous social experiment by Stanley Milgram– has been found in the topologies of numerous examples of networks reconstructed from the real world.

2.2.2 Topological properties

Following the advances on the new science of networks, comprehensive studies of topological properties of complex networks have developed. Three are the general characteristics of complex networks: high clustering coefficient, small-world phenomena and degree distribution that deviates form Poison. In (Albert and Barabási, 2002) additional properties inspired by statistical mechanics and spectral theory have been used to analyse a group of 15 real world networks (Albert and Barabási, 2002, table 1). The preferential attachment network evolution mechanism is also analysed and the statistical properties of networks generated according to this mechanism are compared with those of real networks. More structural properties of networks, random generation models and dynamical processes that taking place on the networks are presented in the comprehensive review of (Newman, 2003) and discussed as tools to understand the function of systems build upon complex networks. Topological structure thus, provides evidences for the dynamical properties of systems that can be modelled by networks and measures of these properties valuable tools for the analysis of systems behaviour as it is suggested in (Barabási, 2005). It is not by chance that the structure vs. function relationship has been initially studied in a random Boolean network model context (the NK model reviewed in section 2.1.1.1).

Studies that relate the topological structure of networks to the dynamics of systems that can be represented by networks are giving insights into the dynamics of biological systems and also are essential for the transition from molecular to systems descriptions in the biosciences. The population structure (May, 2006), the dynamics of epidemic spread (Pastor-Satorras and Vespignani, 2001), the organisation of metabolic networks (Jeong, Tombor, Albert, Oltvai, and Barabási, 2000) and of transcription regulation networks (Farkas, Jeong, Vicsek, Barabási, and Oltvai, 2003) constitute just a sample of studies that connect topological properties and organisation of large complex networks of biological entities with dynamical properties. Furthermore, in an explicit connection between topology and dynamics Aldana has studied the dynamical robustness against variations of

the internal parameters of Boolean networks with regards to their topology (Aldana and Cluzel, 2003) and conclude that scale-free networks of certain parameters are more robust than the NK alternatives.

However several studies have raise some issues regarding the generalisations that have followed the analysis of large scale high-throughput data with regards to sampling. (Stumpf, Wiuf, and May, 2005) explore the sampling properties in the degree distribution of networks. The paper shows that the sample degree probability distribution is expected to be the same with the original network, however in networks with scale-free topology it is shown that this is not the case. Similarly a study by (Han, Dupuy, Bertin, Cusick, and Vidal, 2005) has simulated the partial sampling at yeast two-hybrid (Y2H) high throughput data. Sampling biases can result in the appearance of scale-free topologies. Four different network types have been used to sample from and all resulted to networks with the same characteristics. Scale-free topology cannot be confidently assigned to complete interaction networks.

In addition to sample bias the generality of power-law has been criticised in (Fox Keller, 2005) as not been the universal architecture of complex biological networks, that numerous context specific mechanisms are able to generate power law degree distributions and that preferential attachment is one among them. In a work discussing the variability of complex phenomena (Willinger, Alderson, Doyle, and Li, 2004), debates the "emergence of scaling" property of power law networks and supports that non-normal distribution is a typical phenomenon for systems with complex behaviours. Additionally, the preferential attachment is not considered to be a universal mechanism of power-law degree distribution as (Salathé, May, and Bonhoeffer, 2005) proposed a diametrically different generation mechanisms the "selective removal" able to generate power law degree distributions, and considered its biological implications such as attack tolerance to mutations as debatable.

Therefore, despite some initial enthusiasm that studies of topological structures of complex biological networks can contribute to a systems level of understanding of biological systems and reveal some universal organisational principles, this conclusion is still far to reach. As discussed, both experimental biases based on the current methods used for data acquisition as well as methodological reasons based on the generality of the mechanisms that are used render the structure vs. function studies in biological systems a central, however still open, question.

2.2.3 Network topological measures

A variety of measures have been devised to capture the characteristics of complex networks, (da Costa, Rodrigues, Travieso, and Villas Boas, 2007) have presented a comprehensive and exhaustive survey of measures that are used to analyse complex networks. In a similar fashion (Boccaletti, Latora, Moreno, Chavez, and Hwang, 2006) are reviewing a large array of studies of the structure of dynamical systems, as represented by networks of interacting parts, and their dynamics. The work reviewed in (Boccaletti et al., 2006) consists of a considerable account of the current state of the art research that aims to use topological measures of complex networks to understand the function and the dynamics of the underlying systems. This section reviews network measures that pertain to global properties of the network as well as network measures pertaining to local properties and individual elements of the network (genes, regulatory interactions) that have been used to analyse the dynamics of biological systems.

2.2.3.1 Global network measures

Characteristic properties of the topology of a network include: the degree distribution, the diameter and the clustering coefficient. For the rest of this section the standard graph notation is used and a graph $\mathcal{G}$ is defined as the tuple $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertices and $\mathcal{E}$ the edges set.

Degree and degree distribution: The degree $d_G(v) = d(v)$ of a vertex v is the number of edges incident to that vertex, an edge e and a vertex v are incident if the vertex v is on edge e. The number:

$$d(\mathcal{G}) = \frac{1}{|\mathcal{V}|}\Sigma_{v\in\mathcal{V}}d(v)$$

is the average degree of $\mathcal{G}$ (Diestel, 2005, Sec:1.2), which essentially is the ratio of edges over vertices or the $|\mathcal{E}|/2|\mathcal{V}|$ for directed graphs. The nature of the distribution of the degrees of all the vertices in a graph is distinctive of the generating mechanism of the graph. The Erdös-Rényi (ER) network mechanism generates networks with Poisson degree distribution (Erdös and Rényi, 1959) and the preferential attachment is one of the mechanisms to generate a power-law (PL) degree distribution (Albert and Barabási, 2002).

Diameter and clustering coefficient: The diameter of a graph diamG is the largest distance among all the vertices pairs $\text{diamG} = \max d_{\mathcal{G}}(x, y)$. The distance $d_{\mathcal{G}}(x, y)$ between vertices x, y is defined as the shortest path that connects the two vertices. The clustering coefficient is a graph measure with two versions, one referring globally to the whole graph and the other locally to individual vertices. The local clustering coefficient C_v of a vertex v is a measure of the of the *cliqueness* of the neighbourhood of the vertex and it is given for undirected graphs by the ratio of the number of edges between all the neighbours of a vertex $\mathcal{E}_{jk}$ over the number of edges that could potentially exist within the neighbourhood of a vertex with degree k_v

$$C_v = \frac{2|\mathcal{E}_{jk}|}{k_v(k_v - 1)}$$

. The network clustering coefficient is the average of all local clustering coefficients of each vertex $v \in \mathcal{V}$ (Watts and Strogatz, 1998). Networks with equal number of vertices and edges that have been generated with the ER model have on average larger diameter than networks generated with the PL process (Albert and Barabási, 2002), the small-world phenomenon is observed more frequently on power-law degree distribution networks. In addition small-world networks tend to be more clustered than their random graph equivalents and tend to have higher average clustering coefficients. The network *transitivity* has also been suggested by (Newman, 2003) as an alternative which is defined as the density of triplets on a network, is also considerably higher in complex networks.

Cycles In the context of this study a cycle is considered as a directed path where the last vertex on the path is connected to the first vertex of the path, the cycle thus is a directed cycle and all edges are pointing to the same direction. Cycles in regulatory networks have been studied since the early models of gene regulatory networks (Thomas, 1978) (also reviewed in section 2.1.1.1). The role of cycles in the dynamics of regulatory systems has been studied theoretically, leading to the formalisation of a set of simple laws for feedback circuits in biology. Cycles are characterised as positive or negative depending on the parity of the negative interactions in a cycle, positive cycles have even number of negative interactions and negative odd. The laws for the dynamics of regulatory circuits are summarised in (Thomas, 1998) as follows: positive cycles are a prerequisite for multistationarity and thus for differentiation and memory; negative cycles are required for the existence of a single attractor (either a stable steady state or a limit cycle) and thus homoeostasis in biological systems. Theoretical extension of R. Thomas work has

provided a general proof that positive feedback is a necessary condition for the existence of multistationarity and differentiation (Soulé, 2003) and an extension of this proof for differentiable systems in (Soulé, 2006). In terms of positive and negative feedback loops (NFLs and PFL respectively) cyclic structures have been evolved in an evolutionary computation approach to favour hysteresis and multistationarity (Kim, Kim, Jung, Kim, Park, Heslop-Harrison, and Cho, 2008). The computational simulation results have revealed that GRNs decrease the number of NFLs to enforce hysteresis and to accomplish multistationarity GRNs have been evolved to decrease the number of NFLs and increase the number of PFLs, results that come as a computational reproduction of the theoretical work of R. Thomas.

Cycles have also been studied in known biological networks (as well as to other known complex networks) and dynamic behaviours have been connected with the characteristics of cycles. In studies of gene regulation in the yeast *S. cerevisiae* (Luscombe, Babu, Yu, Snyder, Teichmann, and Gerstein, 2004) have revealed that cycles are involved in endogenous activities of the cells (such as the cell cycle) and (Jeong and Berman, 2008) have more strongly associate cycles with the regulation of the cell cycle and stress response. In modelling studies the transmission of signals in biological systems have been associated with cycles, negative cycles are enabling robust signal processing (Ziv, Nemenman, and Wiggins, 2007). Ma'ayan et. al. have introduced the concept of ordered cyclic motif and found that cycles where consecutive edges have opposing directionality are overrepresented in real complex networks (Ma'ayan, Cecchi, Wagner, Rao, Iyengar, and Stolovitzky, 2008), this topological property appears to increase dynamic stability of large networks. The study of cycles and their topological properties as regulatory features in complex biological networks constitutes an active topic in network biology.

2.2.3.2 Local network properties

Motifs The concept of motifs as characteristic patterns of complex networks has introduced to the topological studies of complex networks by (Milo, Shen-Orr, Itzkovitz, Kashtan, Chklovskii, and Alon, 2002). Network motifs are relative small (3 or 4 nodes) connected subgraphs that have been found to be overrepresented in numerous real world complex networks. Motifs are considered to be formed by common design principles in biological, ecological or engineered networks and specific motif classes are represented in higher frequencies in these real world networks than in randomised networks of the same size. Network motifs have been identified by the same approach in the transcription regulatory network of

the bacterium *Escherichia coli*, most motifs comprise feed-forward loops which is considered as an information processing mechanism that filters transient signals and responds only to persistent ones (Shen-Orr, Milo, Mangan, and Alon, 2002).

Relatively small families of motifs additional to the 3 or 4 nodes motifs discovered above are considered the building blocks for the regulatory networks of yeast and *E. coli*. Feed forward loops (FFLs), single input motifs (SIMs) and dense overlapped regulons (DORs) appeared to shape the GRNs for signal transduction in these organisms and form the design principles of networks that require fast responses to external signals (Alon, 2007). Regulatory networks that control developmental processes, as these processes are spanned to longer time periods, use all the motifs described above plus positive feedback loops, longer transcription cascades and larger, and more complex FFLs organised in modules (Alon, 2007).

The family of size-3 motifs, the motifs that comprise 3 vertices, that are not structurally isomorphic to each other is depicted in figure 2.2.

However simulation studies have suggested that motif function is not determined by the motif structure. Motifs exhibit a functional variability depending on the dynamical (kinetic) parameters of the system (Ingram, Stumpf, and Stark, 2006; Prill et al., 2005), or on the network context that the motif is embedded within (Knabe, Nehaniv, and Schilstra, 2008a).

2.2.3.3 Individual elements measures

Individual element (either vertex or edge) topological properties take into account the positioning of an element in the topology of the whole network. Measures that capture significant information of the importance of an element within the network structure are reviewed here.

Vertex / Gene Centralities Centrality measures for individual vertices are used in the network literature to describe the topological properties of an individual element in a graph (da Costa et al., 2007). Formally a centrality is a function that assigns a real number value to an individual element of a network, the formal definition of a centrality function can be found in the respective (Centrality) entry of the glossary. Vertices with high centralities play a prominent role in the dynamics of a network and this is the case for biological networks. The first application of centrality measures in biological networks involved protein-protein interaction

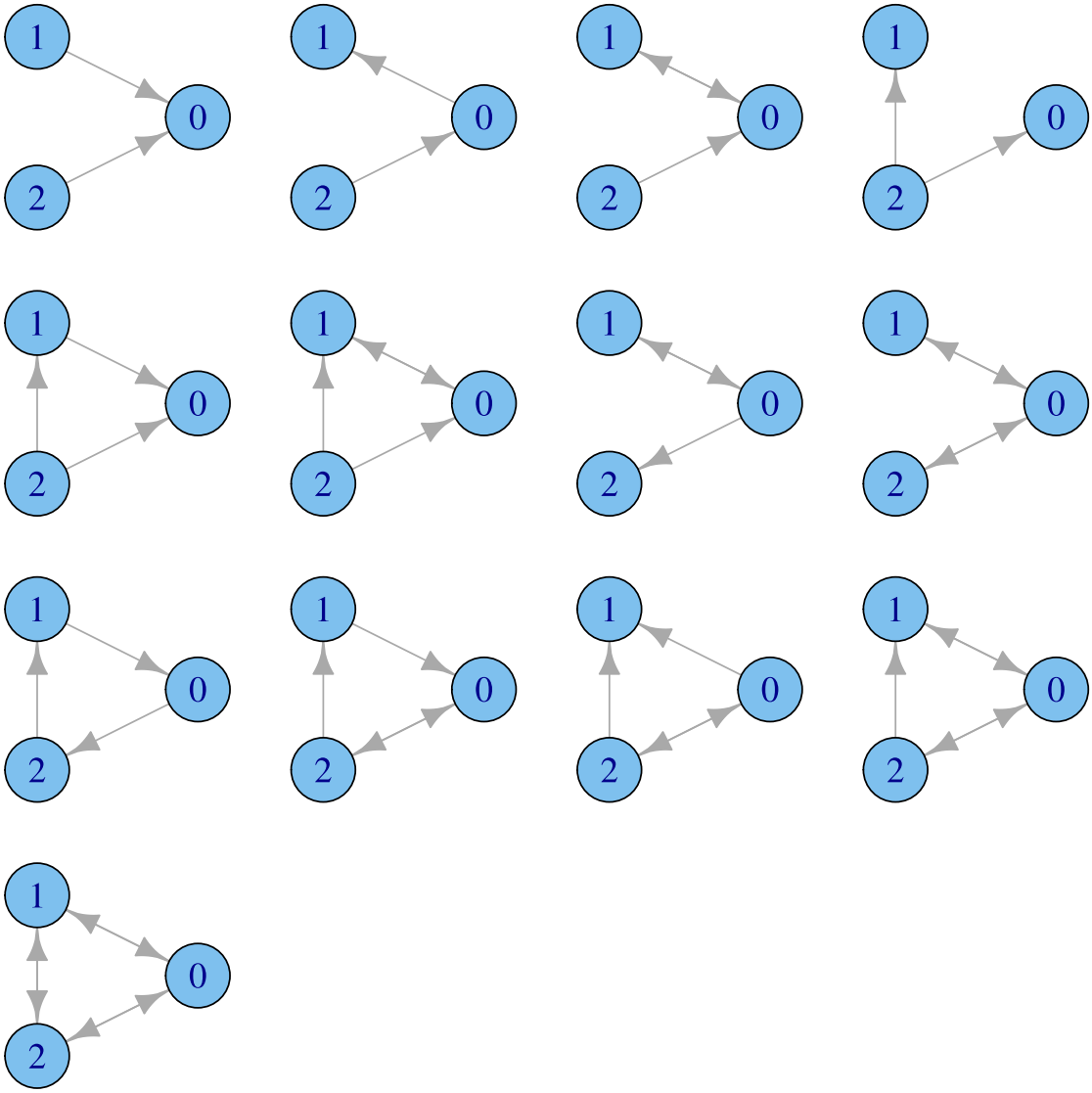


FIGURE 2.2: Enumeration of all the possible, distinct and non-isomorphic network motifs of size 3 (13 in total).

networks (PPI) where the essentiality of proteins (in terms of the survival rate of the single knock-out mutant) found to be correlated with the centrality measures of the protein in the PPI (Yu, Kim, Sprecher, Trifonov, and Gerstein, 2007) and a biological explanation of this property, that is that high centrality proteins are involved in essential biological modules has been demonstrated recently (Zotenko, Mestre, O'Leary, and Przytycka, 2008). Here the application and tools to study gene centralities in GRNs will be reviewed with a focus on the degree, closeness, betweenness and eigenvector centrality vertex properties. A vertex degree is the elementary centrality measure and it has been defined in section 2.2.3.1. *Closeness centrality* of a vertex is defined as the average shortest path between the vertex and all the vertices that can be reached from it. *Betweenness centrality* of a vertex v is the number of all the shortest paths between all the rest of pairs of vertices on

the network that pass through vertex v over the number of all shortest paths (excluding vertex v). In this thesis the random-walk betweenness algorithm is used to calculate betweenness (Newman, 2005). *Eigenvector centrality* for each vertex in a network equals with the respective component of the eigenvector of the biggest eigenvalue of the adjacency matrix. A set of centrality measures, namely the degree, the eccentricity, the closeness, the betweenness centrality and the eigenvector centrality has been used to characterise genes in the transcription regulatory network of *E. coli* (Koschützki and Schreiber, 2004). This initial work on gene centralities of GRNs has identified significant correlations between different centrality values and paved the way for additional research that relates centralities with biological functions. A software tool is available for calculating a set of 17 different centralities in biological networks centralities (Junker, Koschützki, and Schreiber, 2006). The latest work (del Rio, Koschützki, and Coello, 2009) is identifying essential genes in GRNs by systematically measure 16 different centralities (del Rio et al., 2009, table 2), none of the measures alone is able to identify essential genes however combinations of 2 or more centrality measures can separate essential genes in *S. cerevisiae*.

To my knowledge, there is a lack of measures that connect individual genes with cycle measures. And as reviewed in section 2.2.3.1 cycles have a role as regulatory futures in GRNs that requires further investigation. Contributing to that direction this thesis proposes the concept of participation of a gene (vertex) in a cycle and uses the number of cycles a gene is a member of, as an additional individual gene measure (formal definition follows in section 4.4).

Edges / Regulatory interactions Centralities Vertex centralities are measures of the information flow that takes place in an individual node. The edge equivalent centrality measure is the *edge betweenness centrality*. The measure has been developed by Girvan and Newman (Girvan and Newman, 2002) to detect communities in social directed and biological networks. Edges with high betweenness tend to connect communities and thus by removing them network communities can be identified. In biological networks edge betweenness is used for the identification of modules and an extension to the Girvan-Newman community detection algorithm that is able to deal with directed and weighted networks has been proposed (Yoon, Blumer, and Lee, 2006).

The lack of a measure to connect individual regulatory interactions with cycles has also been observed for edges centralities. Therefore this thesis introduces

the concept of regulatory interaction (edge) participation to a cycle and proposes the number of cycles a regulatory interaction participates in, as an additional topological measure for individual interactions (see section 4.4 for an introduction to this measure).

As a final remark, the current graph abstraction that is widely used to represent interconnected biological entities starts to become inadequate to incorporate the increasing details of biological systems descriptions that become available in an increasing rate. Thus, groups of researchers concentrating on the developing the next abstraction, one of them have proposed the use of hypergraphs as a representational object for a more accurate and complete description of complex biological relationships. For an extension of the graph based abstractions to hypergraph representation a paper by S. Klamt (Klamt, Haus, and Theis, 2009) introduces the concepts. The same group has developed a tool for analysing biological networks based on hypergaphs (Klamt, Saez-Rodriguez, and Gilles, 2007).

2.2.4 Biological networks

The notion of robustness and evolvability are central concepts in studies of genetic regulation (Wagner, 2005). Robustness as a generic term refers to resilience to change, in biological systems is a multilevel property and appears with different definitions in different levels of biological organisation. In the gene regulatory networks level, robustness is realised by various aspects including: robustness to network topological perturbations, that is robustness to gene knock-outs, regulatory interaction deletions or network rewiring; robustness to alterations in dynamical parameters, that is variations to dynamical parameters of regulation by mutations of the transcription factor coding gene or single nucleotide substitutions of the transcription factor binding site; and robustness to noise or external perturbations, that is robustness to random perturbations in factor concentrations (noise) and/or robustness to factor concentrations fluctuations from to environmental changes (although GRNs should also be capable to elicit responses out of a cell/organisms from several environmental signals).

Characteristic early studies that connect the gene regulatory network structure with dynamics and function of biological systems can be found on (Mjolsness, Sharp, and Reinitz, 1991), where a modelling framework for development was introduced and a model combined discrete (for cell differentiation) with continuous

(for variables updates) time and grammatical rules to model growth and differentiation. The model was used for the study of the segment polarity network in *D. melanogaster*, by the two morphogens *bicoid* and *hatchback* in (Reinitz, Mjolsness, and Sharp, 1995), and highlight the underlying biochemical relationships of the regulation as important to study the dynamics of the model. In another early work, central for this thesis, Mendoza and Alvarez-Buylla study (Mendoza and Alvarez-Buylla, 1998), motivated by the ABC model for flower morphogenesis, have suggested a more comprehensive network of gene regulatory interactions compatible with the ABC model and studied the dynamics in a discrete state model.

Chapter 3

Modelling Framework

"What I cannot create, I do not understand"
Richard Feynman

This chapter formally introduces the full complement of the computational framework developed to study GRN gene expression dynamics in spatially extended systems and discusses and motivates the control parameters set.

3.1 Spatial Model

Space in the modelling framework is represented as a 2-dimensional discrete orthogonal lattice with periodic boundaries. Similar types of spatial structures have been used in previous studies to represent the spatial organisation of cells and to study the gene-cell interaction dynamics in coupled maps (Bignone, 1993), the effects of signalling networks in the developmental complexity (Keränen, 2004) and analytical studies of pattern formation (Plahte, 2001). The equivalent topological object of this structure is a torus (a doughnut shaped arrangement of discrete elements). Every site in the lattice is occupied by a cell, that is a transsys instance in the model (introduced in section 2.1.3.2) which comprises the transsys instance state, the (x, y) coordinates in the discrete space and the transsys program. All the cells in the lattice are occupied by a transsys instance of the same transsys program. The set of all the transsys instances $\mathcal{P}$ occupying a lattice constitute a lattice reactor and is denoted by $\mathcal{P}_{\text{lattice}}$.

Each transsys instance on a lattice reactor exchanges gene products with all its nearest neighbours in a 5-cell von Neuman cellular automata neighbourhood. The

gene product exchange is based on a diffusion mechanism where the diffusibility d_f of a factor f is used to calculate the amount of factor concentration that is diffused to the 4 neighbours as illustrated in figure 3.1. In each timestep the amount D_f of

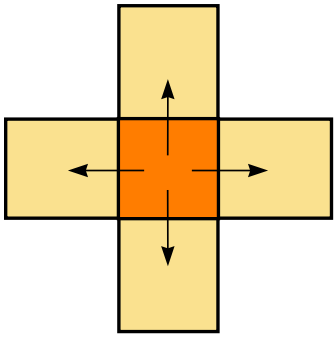


FIGURE 3.1: Illustration of a 5 cells neighbourhood on a lattice. The arrows indicate the net diffusion from the site with higher factor concentration in the middle, to the four neighbouring sites with lower concentration.

a factor f with concentration $C(f, p)$, in a transsys instance $p_{(x,y)}$ located in the (x, y) position of the lattice, which is diffused to each of its 4 neighbours ($p_{(x+1,y)}$, $p_{(x-1,y)}$, $p_{(x,y+1)}$, $p_{(x,y-1)}$) is given by the formula:

$$D_f = \frac{C(f,p) \cdot d_f}{(4d_f + 1)} \tag{3.1}$$

The diffused quantity D_f of factor f from the transsys instance $p_{(x,y)}$ is added to the factor concentration values of each of its neighbours. The volume of the sites in the lattice remains constant and the concentration of all factors inside each lattice site is considered uniform. This discrete diffusion mechanism is designed such that it will not generate any heterogeneity, meaning that the state of a lattice reactor where factor f has the same concentration in every transsys instance (homogeneous) will remain homogeneous and identical after the calculation of the diffused amounts D_f.

The update method of a lattice reactor comprises the synchronous calculation of diffusion for each transsys instance as described before followed by the invocation of the update function as specified in section 2.1.3.3, for each transsys instance. Gene expression within each transsys instance is not determined solely by the regulatory interactions of genes, as instructed in the transsys program, but also from the factor concentrations in its 4 neighbouring cells. Thus spatial organisation is materialised through the exchange (by diffusion) of gene products among neighbouring cells.

3.1.1 Null model

As a null model, a model that lacks spatial organisation, a well stirred reactor has been designed. The well stirred reactor is constructed the same way as a lattice reactor, 2D orthogonal space, occupied with a set of transsys instances (denoted $\mathcal{P}_{\text{wellstirred}}$), periodic boundary conditions and the same diffusion mechanism. The fundamental difference is in the update method, which, in addition to calculating diffusion and simulating gene expression, randomly shuffles the position of each transsys instance in the reactor at the end of each invocation at each timestep. Therefore a 5 cell neighbourhood does not consist of the same instances at two consecutive timesteps. Thus any notion of spatial organisation is distorted in the well stirred reactor.

To delineate further the last statement, consider a case where a factor has a very high concentration in only one instance of a lattice (a peak) and the same for a well stirred reactor, there is only diffusion in the system and no gene expression or decay is taking place. After diffusion calculations for a sufficient amount of timesteps the factor concentration of the transsys instances around the peak will gradually be lower as one moves away from the peak, whereas in the well stirred reactor the factor concentration in the transsys instances (apart from the peak) will approximate the average amount of factor concentration that has diffused.

3.1.2 Spatial gene expression dynamics

Both the lattice and the null model reactors are central objects in every experiment of this study, with the lattice representing systems with spatial organisation and the well stirred reactor the control ‘9or background) experiment. For all the experiments both the lattice and the well stirred reactor have to have the initial factor concentrations of all the factors in all the instances initialised. Each factor in both reactors takes the same initial factor concentration value, drawn out of a random uniform distribution, to sample uniformly and unbiased the set of possible initial states of a reactor. The initial factor concentration state of the lattice is always identical with the one of the well stirred reactor. The range of the random uniform distribution is a user control parameter and it represents the initial inhomogeneities that are inherent into most biological systems. For instance the factor concentrations along a tissue exhibit stochastic variations that may be used

from pattern formation mechanisms to generate patterns. A factor f has a heterogeneous gene expression profile if its concentration levels vary along the different instances of cells in a reactor.

The objective of the computational model so far is to be able to reproduce phenomena where gene expression is more heterogeneous in the lattice (spatially organised model) than in the well stirred reactor (null model) as they have been introduced in the objectives of the thesis in section 1.6. To measure the level of heterogeneity of gene expression in both the lattice and the well stirred reactor and be able compare them, a measure to quantify heterogeneity of gene expression levels has been devised.

3.2 Quantifying Gene Expression Heterogeneity

To quantify heterogeneity in factor concentration in a set of transsys instances a Shannon information based measure is induced. The measure is inspired by the concept of information in biology as described by J. Maynard-Smith (Maynard Smith, 1999, 2000) and is based in the information theory by Claude Shannon (Shannon, 1948). Shannon introduced the concept of entropy in a telecommunications based context as a measure of the information content of a message, however this concept has been extended and Shannon entropy is used as a statistic measure in different contexts and also in biosciences. Shannon entropy measures have been used in biosciences, among other applications, to describe heterogeneously expressed genes in different treatments and identify potential drug targets (Fuhrman, Cunningham, Wen, Zweiger, J., and Somogyi, 2000). A factor with homogeneous distribution of gene expression levels throughout a set of transsys instances is in maximum entropy state and contains no information. Whereas distributions of factor concentrations that exhibit heterogeneous expression profiles among different transsys instances on a reactor have a positive information content. Shannon entropy is expressed in terms of the the relative frequency of each individual factor concentration level. Thus, in a transsys instance p from a set of transsys instances $\mathcal{P}$, a factor f has relative concentration:

$$R(f,p) = \frac{C(f,p)}{C_{\text{total}}(f,\mathcal{P})} \tag{3.2}$$

where $C_{\text{total}}(f, \mathcal{P})$ is the sum of concentrations of factor f in the set $\mathcal{P}$. The Shannon entropy of this factor f in $\mathcal{P}$ is then calculated by:

$$H(f, \mathcal{P}) = -\sum_{p \in \mathcal{P}} R(f, p) \log_2 R(f, p) \tag{3.3}$$

The maximum Shannon entropy is reached when a factor concentration is equal among every transsys instance p of the set $\mathcal{P}$ and is given by:

$$H_{\max}(\mathcal{P}) = \log_2 |\mathcal{P}|$$

Having calculated the Shannon entropy and the maximum entropy of factor f then the information content $I(f, \mathcal{P})$ of a factor f in the set $\mathcal{P}$ is:

$$I(f, \mathcal{P}) = H_{\max}(\mathcal{P}) - H(f, \mathcal{P}) \tag{3.4}$$

Equation 3.4 provides a measure of heterogeneity of the gene expression of factor f in a set of transsys instances $\mathcal{P}$. If the gene expression of f is homogeneous among $\mathcal{P}$, the $I(f, \mathcal{P})$ will be equal to zero. The unit of the information based measure $I(f, \mathcal{P})$, as the logarithm with base 2 is used, is bits, therefore a homogeneous factor expression profile carries 0 bits of information. In another trivial case, a factor which exhibit a zero concentration level in half the instances of a transsys instance set $\mathcal{P}$ and a concentration level of 1 in the other half will carry 1 bit of information.

For a set of transsys instances $\mathcal{P}$ of a transsys program with factor set $\mathcal{F}$, the information based measure for all the factors in a transsys program $I(\mathcal{P})$ is:

$$I(\mathcal{P}) = \sum_{f \in \mathcal{F}} I(f, \mathcal{P}) \tag{3.5}$$

The last equation (eq. 3.5) is a measure of heterogeneity of gene expression of a particular transsys program from which all the elements of the set $\mathcal{P}$ have been instantiated. $\mathcal{P}$ is an arbitrary set of transsys instances. If $\mathcal{P}$ is constituted by a reactor (either a lattice or a well stirred reactor), the level of gene expression heterogeneity of the particular transsys program in the reactor $\mathcal{P}_{\text{reactor}}$ will be returned. The measure is used as the basis to compose an objective function to be used in an optimisation approach.

3.2.0.1 Heterogeneity measure discussion

The heterogeneity measure described in section 3.2 is able to quantify the heterogeneity on a collection of transsys instances where factors acquire different concentrations in a fraction of the available instances and differentiate this score from a collection where factors have homogeneous concentration levels.

However, the information based heterogeneity score is unable to distinguish between transsys instance collections where a factor is differentially expressed at the same number of transsys instances regardless the arrangement of these instances in the grid. The information based heterogeneity measure is invariant to spatial arrangement of heterogeneity and therefore will be unable to quantify spatial patterns on lattices, even though is able to distinguish heterogeneous transsys instance collections.

Measures that captures the spatial arrangement of differential factor concentration values are spatial correlation types of measures. Spatial correlation (specified formally in the glossary entry Spatial Correlation) measures the tendency for sites that are near to each other to have more similar or dissimilar values of their statistics. Spatial correlation measures therefore will be able to quantify the difference between the pattern in the middle of fig. 3.2 -no spatial pattern- and the top of fig. 3.2. Here a spatial correlation measure is used, which calculates the Pearson correlation coefficient between the Manhattan distance of each pair of cells in the lattice and the Euclidean distance of their respective factor concentrations (i.e. the gene expression profile).

Figue 3.2 illustrates examples of transsys instance collections arranged in a lattice which result to the same information content but exhibit different spatial arrangement and therefore their respective spatial correlation measures are different. However as this thesis is concerned with the emergence of gene expression heterogeneity in general and not particularly with the studies of types of spatial patterns that can be risen the need to use a spatial correlation based measure is limited. In conclusion the heterogeneity measure as it is defined in the equation 3.5 is a measure that can be applied to distinguish heterogeneous collections of transsys instances and not spatial patterns of factor concentrations.

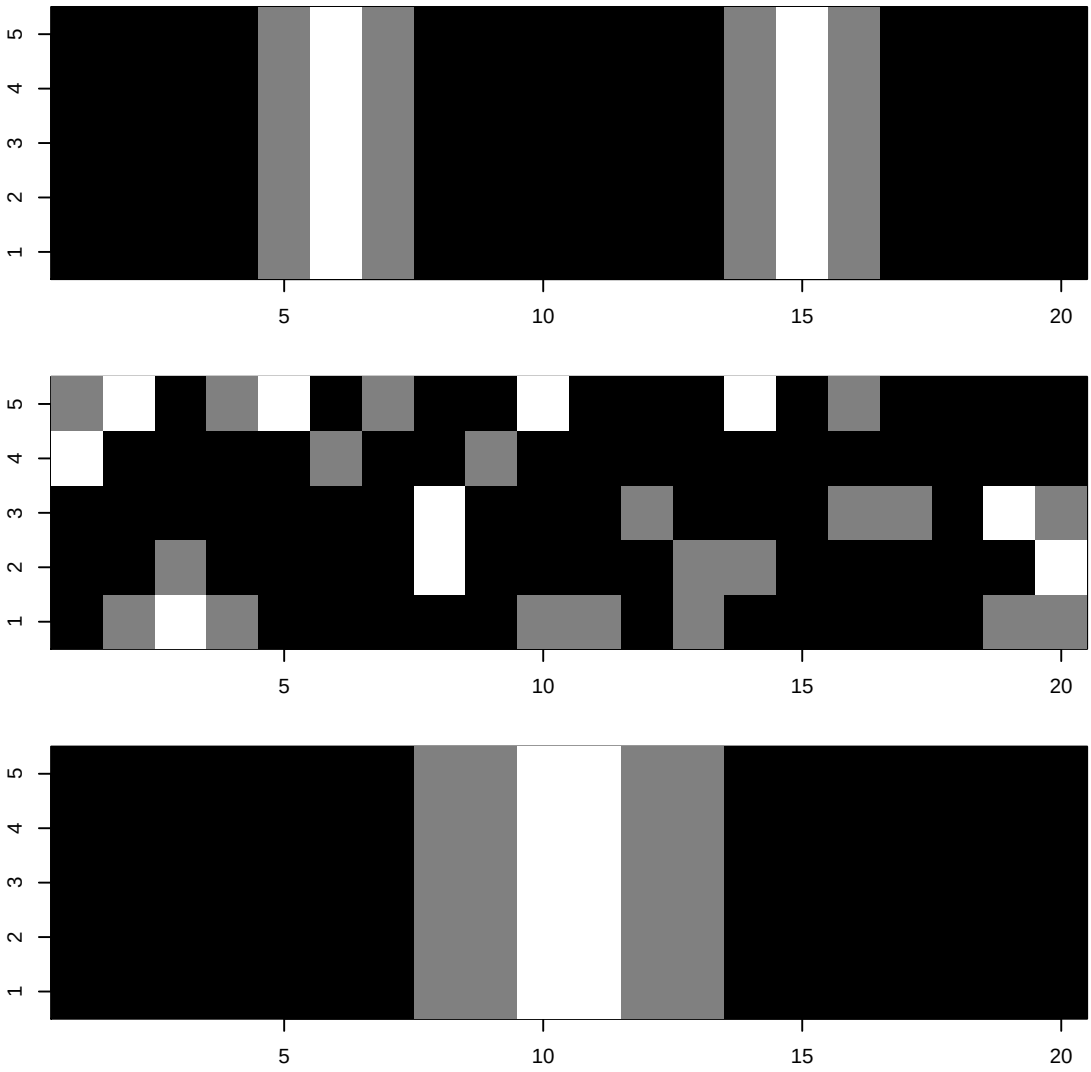


FIGURE 3.2: Greyscale images of three distinct spatial patterns on a 5x20 lattice. Top a pattern of two highly expressed stripes, middle a random pattern bottom a pattern of a stripe exactly twice the size of the stripe on top. All the above patterns have exactly the same information based score $I = 1.821$ bits, however their respective spatial correlation scores are: for the stripy pattern on top 0.036 for the random arrangement in the middle -0.037 and for the blob pattern in the bottom 0.311.

3.2.1 Objective function

A central objective in this thesis is to devise a measurement of difference of heterogeneity of gene expression on a spatially organised system compared to the null model. To quantify this difference an objective function has been devised to calculate the difference of the information content of a lattice from that of a well stirred reactor (WSR). For a lattice and a well stirred reactor populated with instances of the same transsys program p, initialised with identical initial reactor

states and both updated for equal number of timesteps t, the objective function is defined as:

$$O(p,t) = I(\mathcal{P}_{\text{WSR}})_t - I(\mathcal{P}_{\text{Lattice}})_t \tag{3.6}$$

or using function notation and expressing the objective score as a function of the transsys program and the gene expression simulation parameters the objective function f is defined as:

$$f : \text{SimParams} \times \text{TranssysPrograms} \to \mathbb{R} \tag{3.7}$$

or

$$f(s,t) \mapsto \text{objectiveScore} \tag{3.8}$$

Also, as the $\log_2$ is used Shannon information is calculated in bits, the objective score units are bits of information as an indication of the difference in heterogeneity between different reactors. To observe higher information content in the lattice than in the well stirred reactor, and thus having higher gene expression heterogeneity, the score in equation 3.6 should be negative, the largest the gene expression heterogeneity in the lattice than in the well stirred reactor the more negative the objective score is.

3.2.1.1 Objective function evaluation

A transsys program P enters the objective function evaluation procedure. Two sets of identical initial reactor states are generated one for the lattice and one for the well stirred reactor. Gene expression levels are simulated in both the lattice and the well stirred reactor for as many timesteps t as required such that any initial transients will vanish. Then the information content measure is calculated for both the lattice and the well stirred reactor according to the equation 3.5. Each objective function evaluation returns the difference of the objective score in the lattice form that on the well stirred reactor (eq. 3.6, following the process illustrated in the objective function evaluation activity diagram in figure 3.3).

This objective score is a quantification of the property of a transsys program to exhibit heterogeneous gene expression patterns on spatially organised systems. As the transsys program is the same in both the lattice and the well stirred reactor and all the other parameters of the experiments are kept constant in each objective

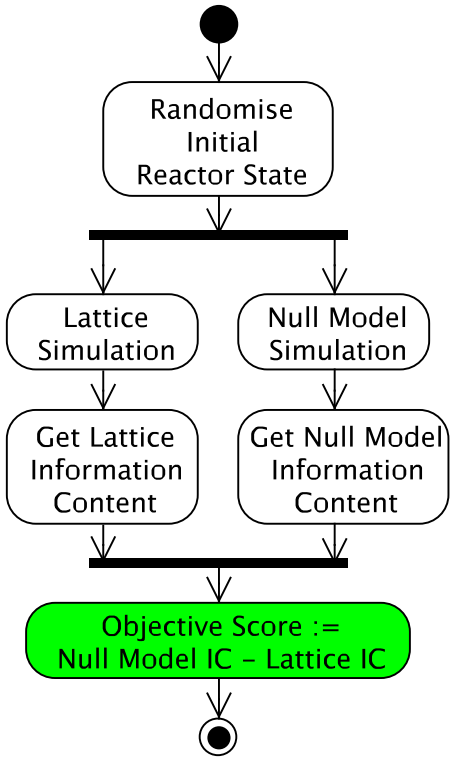


FIGURE 3.3: Activity diagram of the objective function evaluation procedure. The information based heterogeneity score (equation 3.6) is calculated for both a lattice and the a stirred reactor starting from the same initial random conditions.

function evaluation, the difference in the information content between the lattice and the null model will be the result of the spatial organisation only. This result of spatial organisation, as it is quantified by the objective score, constitute the score for an optimisation approach.

As more negative scores imply higher heterogeneity in the lattice than in the WSR, the optimiser operates with an objective to minimise this score. Thus the optimiser is technically a minimiser trying to minimise the objective score as much as possible. I hypothesise that the mechanics of the optimiser are able to separate network topologies which have an increased capability to generate heterogeneity of gene expression in lattices than in well stirred reactors. The following section will formally introduce this optimisation approach.

3.3 Optimisation

The computational framework involves a transsys program optimisation approach. The aim of this approach is to reproduce the biological property described in

section 3.1.2, that is to find transsys programs such that the gene expression heterogeneity is higher in the lattice than in the well stirred reactor. In terms of the information based measure, the $I(\mathcal{P}_{\text{lattice}})$ should be higher that the $I(\mathcal{P}_{\text{wellStirred}})$.

3.3.1 Optimiser

The motivation behind the design of the optimiser is that networks with topological properties such that can exhibit higher gene expression heterogeneity on a lattice than in a well stirred reactor will be able to be distinguished by the optimisation procedure. For that purpose the optimiser gets a transsys program as an input and keeps the topology unchanged throughout the whole optimisation procedure. The optimiser operates on the dynamical parameters space of a transsys program (sec. 2.1.3.1), and searches for certain parametrisations such that the objective score is minimised. By optimising the transsys dynamical parameters only, networks with the capacity to generate higher gene expression heterogeneity in a lattice than in a well stirred reactor will be parametrised more efficiently by the optimiser. By keeping the topology stable and try to optimise big collections of transsys programs with topologies generated by random graphs generation mechanisms, the optimiser will pick up the particular networks whose topology enable them to generate heterogeneity in lattices and not in well stirred reactors.

3.3.1.1 Optimisation approach

The optimiser belongs to the Random Local Search family of optimisation approaches. The objective function score for a transsys program is evaluated as follows: $I(\mathcal{P}_{\text{lattice},t})$

Random Local Search Optimisation The optimiser performs a user specified number of optimisation rounds. Each optimisation round consists of the following steps:

1. Any transsys program, with numerical expression values as its dynamical parameters enters the optimiser. The dynamical parameters are kept as the current best solution.

2. A copy of the current best parameters are randomly perturbed, by a specified displacement range to generate the current alternative set of parameters.

3. The current best and current alternative sets of parametrisations are evaluated according to the objective function as described in section 3.2.1

4. If the objective score of the current alternative is lower or equal to the objective score of the current best then the current alternative parameters set is becoming the current best.

5. If the specified number of optimisation rounds is reached the current best transsys program is returned else the current best is set to enter a new optimisation round.

The random local reach optimiser described above is illustrated with the activity diagram in figure 3.4.

More analytically the step 2 of the random local search optimisation approach specified above consists of the addition on the current best dynamical parameters sets (current best parametrisation) of a randomly generated displacement. The displacement is a random number drawn from a symmetric uniform distribution. The range for the displacement has been chosen based on both the potential of the optimiser to explore the parameter space as well as possible and on the ability to distinguish between transsys programs with higher capacity to generate heterogeneity on the lattice than on the null model, from transsys program that lack this capacity.

3.3.2 Transformation functions

The optimiser generates the alternative parametrisation by imposing a random perturbation (the range of which is specified by the displacement parameter `optStep`). This displacement comes from the $[-\texttt{optStep}, \texttt{optStep}[$ interval. The random application of this displacement to a transsys program dynamical parameter is equivalent with an one dimensional random walk starting from the initial dynamical parameter value. An one dimensional random walk of n steps of $[-\texttt{optStep}, \texttt{optStep}[$ range has a displacement expectation $E = \sqrt{n} \cdot \texttt{optStep}$. However the optimiser is a random local search optimiser and always searches on the vicinity of the which is the current best solution, therefore the optimiser performs a directed walk (optimally towards the best solution) and not a random walk in the parameter space. In a good approximation, the directed walk applied in the optimiser, would expected to have a larger expected displacement than the pure random walk. It is

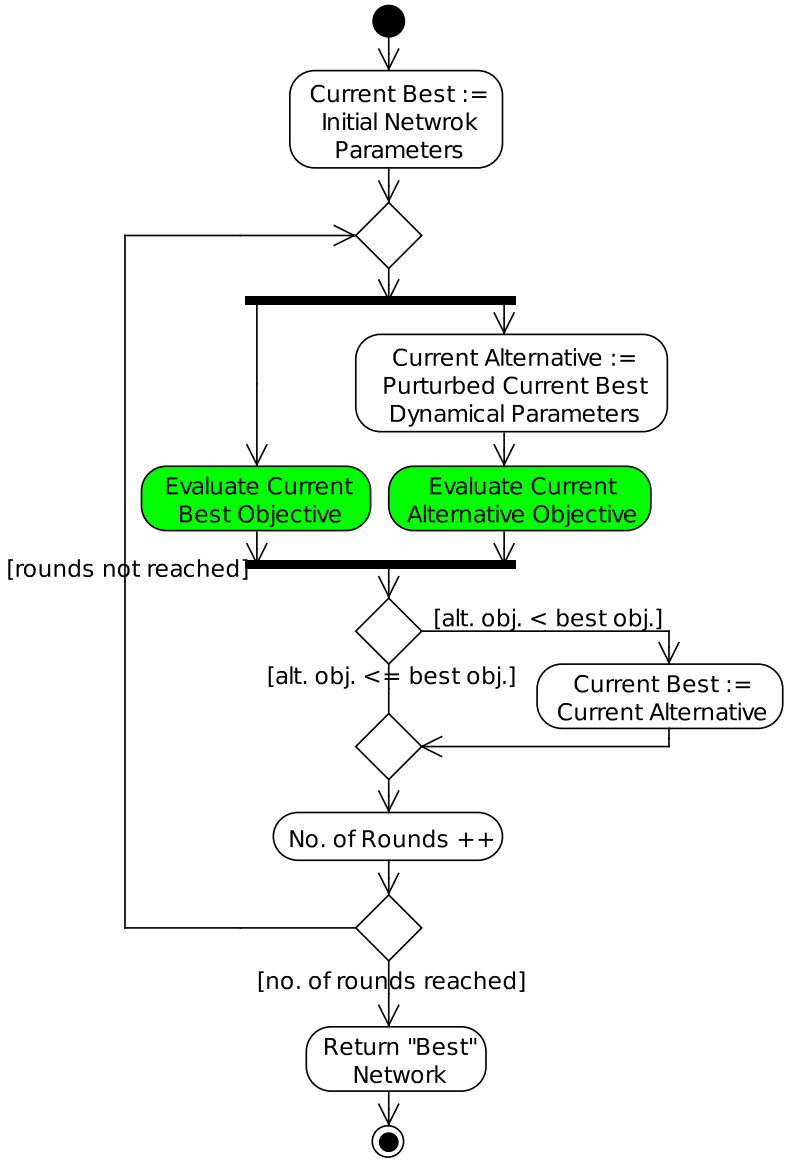


FIGURE 3.4: Activity diagram of the random local search optimisation procedure. Each optimisation round entails two evaluations of the objective function (illustrated in figure 3.3), one for the current best transsys program parameterisation and one after applying a random perturbation to the current best parameterisation.

evident that the expected value of a dynamical parameter can grow relatively fast to biologically implausible negative values after only a few of optimisation rounds.

To avoid this unrealistic behaviour of the optimisation process, a family of functions the transformation functions have been applied. The key role of these functions is to get the dynamical parameters out of a transsys program, transform them to an unconstrained domain, (then the optimisation displacement will be applied on the unconstrained domain) and then transform the unconstrained values back

to the constrained domain of the transsys dynamical parameters. The transformation functions operate within limits for the transsys dynamical parameters and are either upper or lower bounded (or most of the times both, with a trivial lower bound of zero) in the dynamical transsys parameters domain. It is self-evident that a function should be one-to-one and onto and thus to be invertible in order to be a transformation function. Numerous transformation functions are available in the transsys optimisation software package, here in all of the experiments an arc tangent transformation function has been used.

3.4 Random Networks Generation

The random network generation was based on the two random network mechanisms discussed in section 2.2.1 the Erdös-Rényi Erdös and Rényi (1959) random network process (ER) and a precess generating power-law degree distribution based on the preferential attachment method described in Barabási and Albert (1999) (PL). ER networks represent an unbiased generation of networks by sampling the network space and have been used as the baseline model for random networks. PL networks represent a class of networks that although generated with a random process manifest characteristics (see 2.2.1 for description of the network characteristics and measures) that resemble more networks that are present in the real world, refer to (Barabási and Albert, 1999; Faloutsos et al., 1999; Jeong et al., 2000) for particular examples. Networks with power-law degree distribution have been reported to describe more accurately the topological architecture of various biological networks such that protein-protein interaction networks and -more important for this thesis- GRNs. Thus here we treat the ER random graphs as the baseline case of random network generation and the PL as the case that resembles more networks in biological systems.

3.5 Control Parameters

This section contains a description for the random network generation, the simulation and the optimisation control parameters. There is a reference parameters set which was used to generate the majority of the results of this thesis and will be explained in detail however, certain values of these three parameter sets may vary among different experiments. An explanatory background for each of the parameter in the parameters sets will be given here, the reference set will be introduced in

the experimental procedures chapter and whenever there is an experiment where there is a deviation from the reference parameter set it will be individually introduced and discussed in the relative experimental context. The transformation functions parameters are kept stable for all of the experiments and the motivation, description and discussion behind the particular parameters choices can all be found in this section.

3.5.1 Network generation parameters

Every experiment is starting by generating a population of random networks which constitute the input data of the computational procedure described in chapter 3. Transsys programs that represent GRNs are generated according to the following control parameters:

Number of genes: Specifies the number of genes (vertices in a network) of a GRN.

Number of regulatory interactions: Specifies the number of regulatory interactions among genes (edge in a network) of a GRN.

Network seed: Is the random seed of the network generator and specifies the number of different topologies that will be constructed.

Parametrisation seed: Is the random seed of the random number generator for the dynamical parameters of a transsys program. Specifies the number of different initial dynamical parametrisations for a given network topology.

Generation mechanism: Specifies the random network model that will be used for the network generation. For this study the Erdös-Rényi random graphs model (ER) and a random network procedure that generates networks with a power-law degree distribution similar to (Barabási and Albert, 1999) (PL) are used. In both the random network generation algorithms the directionality of the connecting edges is randomly chosen with equal probability, thus on average a network has equal number of incoming and outgoing edges.

3.5.2 Simulation control parameters

The simulation control parameters are specifying each objective function evaluation computation and consist of:

Timesteps: The number of timesteps that a reactor's update function will be evaluated. (integer)

Lattice width: The width, in terms of cell number of a reactor. (integer)

Lattice height: The height, in terms of cell numbers of a reactor. (integer)

Initialisation random seed: The random seed of the random number generator for the initial factor concentration state of a reactor. (integer)

Initialisation range: The interval out of which the random uniform values of the initial factor concentrations will be drawn. (a pair of real values)

Objective Function: The name of the objective function. Throughout all the experiments presented in this thesis only the Shannon information based objective function has been used (as described in section 3.2.1). However the transsys framework provides a collection of different objective functions that can be specified by their names. (string, the name of the objective function)

Null model: The type of the null model that will be used. Throughout all the experiments in this thesis the well stirred reactor null model has been used. However one more null model, an individual collection of cells is also available. (string, the initials of the null model)

3.5.3 Optimisation control parameters

The optimisation control parameters are setting up the optimiser, each one has the following semantics:

Optimisation rounds: The number of rounds of optimisation that the optimiser will perform. (integer)

Optimiser random seed: The random seed for the random number generator of the optimiser's perturbation procedure. (integer)

Displacement: A number specifying the range of an interval from where a random uniform number will be drawn and serve for the perturbation of the current best parametrisation. If a number `s` is specified then the interval is in the $[-\mathtt{s}, \mathtt{s}[$. (real)

3.5.4 Transformation parameters

The following transformers, all based on the arctangent function and the with respective options have been used to transform each dynamical parameter:

Decay transformer: As the decay rate affects the speed a factor concentration reaches equilibrium (according to equation 2.2), it should have a sufficiently large value to allow equilibration. However, very large decay rates are not desired because the system will perform relatively large leaps in the state space and thus avoid to enter attractor basins that might exhibit some of the desired dynamics. Therefore the decay rate has been bounded to 0.50%. Decay is also lower bounded as rates equals to zero will render the system unable to equilibrate, thus a 0.01 lower bound for the decay rate is used.

The decay rate transformer is the arc tangent function, applied in a (0.01, 0.5) interval.

Diffusibility transformer: Diffusibility, or the general ability of a factor to diffuse, needs to be bounded as very high diffusibility values will render the system to a homogeneous state relatively quick and will have a strong homogenising effect on any heterogeneous gene expression might appear. An upper bound of 0.3 has been chosen (the 0.0 lower bound is self-evident as negative diffusibility values are not plausible).

The diffusibility transformer is the arc tangent function, applied in a (0.0, 0.3) interval.

Constitutive transformer: Constitutive expression was chosen to be at relatively low value to represent the basal promoter activity of biological promoters. Thus it is expected that has a relatively small effect on the factor concentrations and most of the activity dynamics of a gene will be a result of gene regulation rather than basal promoter activity.

The constitutive expression transformer is the arc tangent function, from a (0.0, 0.1) interval.

$a_{\max}$ activate transformer: The maximal level of expression rate $a_{\max}$ value is bounded to the unit, (between 0 and 1). By keeping the $a_{\max}$ in he unit one can construct transsys programs with equivalent dynamics by fine-tuning other parameters such as decay rate and α_{spec}. Effectively, this reduces the degrees of freedom of the parameter choices by 1 and provides better estimations of the potential factor concentration values a system can exhibit.

The $a_{\max}$ activation transformer is the arc tangent function, applied in a (0.0, 1.0) interval.

α_{spec} activate transformer: The factor's binding specificity α_{spec} determines (according to equation 2.1.3.3) the speed to which the gene expression rate will reach the maximal value $a_{\max}$. The larger the value of α_{spec} the more time a factor requires to saturate the binding site. The need to bound is to prevent very slow equilibration times and weak interactions. The α_{spec} upper bound is approximately one order of magnitude higher than the $a_{\max}$ upper bound to allow for smoother Michaelis-Menten dynamics.

The α_{spec} activation transformer is the arcus tangent function, in a (0.0, 8.0) interval.

$a_{\max}$ repress transformer: Identical with the $a_{\max}$ activation transformer.

α_{spec} repress transformer: Identical with the α_{spec} activation transformer.

3.6 Network Elements Deletion Procedures

To study the effects of individual graph elements (i.e. vertices or edges) an element deletion framework has been developed. The framework consists of single element deletion approaches for genes (vertices) and regulatory interactions (edges) and two respective sequential element deletion approaches one for genes and one for regulatory interactions.

3.6.1 Single element deletion

The single element deletion, for both genes and regulatory interactions alike, is performed by deleting a single element from the original transsys program at a time. In a wet-lab biological experiment analogy, single gene deletion represents a single gene knock-out experiment and the single regulatory interaction deletion represents either a transcription factor protein modification or a transcription factor binding site mutation experiment. After the element deletion, the objective function for the mutant transsys program is evaluated, as it is described in section 3.2.1 using an identical set of control parameters as the wild type transsys program. The operation is repeated for each of the elements of the transsys program. The difference of the objective score of each single element mutant from the

original transsys program is then calculated together with a set of network related measures for the individual element that has been deleted.

3.6.1.1 Gene knock-outs

A single element deletion operation that is implemented on a graph node is equivalent to a single gene knock-out mutant in a biological experiment. The single gene knock-out operation is repeated for all the genes in a ranssys program and at the end of the procedure the following are returned: The objective score difference between the single gene knock-out mutant and the wild type transsys program, the centrality measures of the gene together with the number of cycles that the gene is a member of, as specified in section 2.2.3.3, as well as the information content of the individual factor that the gene encodes for (calculated according to equation 3.4).

3.6.1.2 Regulatory interaction deletion

For single element deletions implemented on a graph edge the operation is equivalent to the deletion of a regulatory interaction in a biological experiment. The single regulatory interaction deletion operation is repeated for all the regulatory interactions in a transsys program and the following are returned: The objective score difference of the edge reduced transsys program from the wild type together with the edge network centralities and the number of cycle the edge participates in describe in section 2.2.3.3, as well as the dynamical parameters $a_{\max}$ and α_{spec} of the deleted edge.

3.6.2 Sequential element deletion (pruning)

The sequential element deletion approach is based on the cumulative application of the single element deletion operation on a transsys program. The procedure is the equivalent for both gene and edge sequential deletion and consists of the following steps:

1. The elements are sorted according to the effect the single element deletion has procured on the wild type transsys program objective score. The element that its single deletion has procured the smallest difference from the wild type objective score comes first and the rest follow in an ascending order.

2. The first element in the ordered list is deleted from the wild type transsys program and then the objective score and a series of network related measures (described in section 2.2.3.1) are calculated. The transsys programs generated by a sequential element deletion procedure are called pruned transsys programs.

3. For each of the next element in the ordered elements list a single element deletion operation is performed on the pruned transsys program, resulting in a new pruned transsys program. The objective function score is evaluated and network measures are calculated for the new pruned transsys program. At the end of each execution of this step the new pruned transsys program enters the beginning of step 3 as the current transsys program.

4. The operations of step 3 are repeated until the ordered element list is empty.

The sequential element deletion procedure is followed by both the implementations for both gene and for regulatory interactions sequential deletions as follows:

3.6.2.1 Vertices (genes) pruning

The gene pruning procedure returns the objective score difference of the gene pruned transsys program from the wild type one, together with the individual element topological parameters of the gene that has been knocked-out. Including the degree, closeness, betweenness, eigenvector centrality and the number of cycles the gene is a member of, which are returned together with the objective score difference in a tabular format. A transsys file containing all the gene pruned transsys programs is also returned. Note that by the cumulative pruning of genes the transsys program that will be returned last is an empty gene-less and factor-less transsys program.

3.6.2.2 Edges (regulatory interactions) pruning

Similarly, the regulatory interaction pruning procedure returns the objective score difference of the edge pruned transsys program from the wild type one together with the individual element topological properties and the dynamical properties of the regulatory interaction that has been removed. Including the edge betweenness, the number of cycles the edge participates in, the nature, the $a_{\max}$ and the α_{spec} of the interaction, which are returned together with the objective score difference in

a tabular format. In addition a transsys file containing all the regulatory interaction pruned transsys programs is returned. Note that by the cumulative pruning of regulatory interactions the transsys program that is returned consists only of singleton genes.

Chapter 4

Experimental and Analytical Framework

"It requires a very unusual mind
to undertake the analysis of the obvious"
Alfred North Whitehead

A key introduction to the experimentation principles and design will be introduced in this chapter, the reference experiment, an experiment which has generated the core data sets analysed in this thesis, will be introduced and motivated, as well some analytical techniques and procedures that were developed in the context of this thesis.

4.1 Experimental Procedure

4.1.1 Reference control parameter settings

The reference set of experiments presented in this thesis have been conducted by using a reference set of control parameters. The values of the parameters as well as the motivation behind any particular choice are explained in the following section. The reference set of the parameters is used to generate the reference experiment and the reference data. There will be an explicit statement and consequent motivation should any alternative selection of control parameters occur in any parts of this work.

Network Generation Parameters: The parameters interpretation is introduced in section 3.5.1, the choices for the reference set are as follows:

- Number of genes: 15 The number of genes is chosen to be 15, the number is a trade off between reasonable execution time and relatively large network size.
- Number of interactions: 45 The number of regulatory interactions is chosen to be 3 times the number of genes. There is evidence suggesting that biological developmental gene regulatory networks have somewhere between 2 and 4 times as many edges as nodes (e.g. (Alvarez-Buylla, Benítez, Dávila, Chaos, Espinosa-Soto, and Padilla-Longoria, 2007; Oliveri, Tu, and Davidson, 2008)).
- Network generation mechanisms: 2 Random network topologies are generated according to two different generation mechanisms (section 3.4). The Erdös-Rényi random graphs model (Erdös and Rényi, 1959) (to be referred as the ER model thereafter) and a random network procedure that generates networks with a power-law degree distribution (Almaas, 2007; Barabási and Albert, 1999) (to be referred as the PL model thereafter).
- Number of Topologies: 15 The reference network population consists of 15 random networks. Again this number is a compromise between computational time and reasonable sampling.
- Number of initial dynamical parameter sets: 30 The number of different initial random dynamical parameter sets (parametrisations) is chosen to be 30, which again is a compromise between sampling the parameter space and keeping the number of transsys program relatively low for the shake of execution time.

Simulator Control Parameters: They are described in section 3.5.2, the values of the reference set are as follows:

- Lattice width: 60 and
- Lattice height: 5 Lattice size was chosen based on two premises that a lattice with large radius is needed to generate patterns and the number of instances (i.e. cells) on the lattice is kept relatively short for the sake of computational time.
- Number of timesteps: 400 The number of timesteps for the gene expression simulation is 400. It is the minimum required number of timesteps

so that any transient signal form the initial reactor factor concentrations state will disappear. Experiments using significant larger number of timesteps (2000 timesteps) saw the same dynamic behaviour with the 400 timesteps ones.

- Null model: Well stirred reactor (WSR) A well stirred reactor serves as the null model.
- Initial reactor state: $[1, 3[$ The initial factor concentrations on the reactors are drawn from a random uniform distribution on the interval $[1, 3[$. The interval is chosen to start from a non-zero value so that the information content that is initially externally injected to the reactors is reduced (an interval including zero will increase significantly the information content of the initial reactor state). The upper concentration limit is chosen to be a relatively small number so that it represent initial states that biological systems can be exposed to (either owing to environmental or developmental perturbations).

Optimisation Control Parameters: They are explained in section 3.5.3 and the values of the reference control parameter set are as follows:

- Optimisation rounds: 200 The number of rounds that the optimiser should complete is set to 200, a trade-off between having a substantial amount of optimisation rounds and computation time.
- Displacement: $[-0.5, 0.5[$ The random perturbation that the optimiser impose in every round on the transformed values of the dynamical parameters of the current best parametrisation. The choice of a relatively large interval has been taken after conducting a sweep experiment of all the intervals between $[-0.1, 0.1[$ to $[-1.0, 1.0[$ by increment the step by 0.1.
- Optimiser random seed: 1 This is the random seed that controls the optimiser's random number generator and is set to one for all the experiments conducted in the course of this thesis.

4.1.2 Reference experiment

The experimental procedure using strictly the settings for the control parameters specified in the previous section and the transformation values specified in section 3.5.4 will be referred as the reference experiment in the rest of the thesis. It was conducting as follows: A population of 2 generation mechanisms $\times$

20 topologies × 30 parametrisations = 1200 transsys programs ware generated. Each transsys program has entered the optimisation procedure which returns the optimised objective score and a table of all the factor concentrations after the last optimisation round. Single element deletion experiments are then conducted (as described in section 3.6.1) using the reference simulator control parameter set. This experiment conducted using the reference control parameter set will referred as the reference experiment for the rest of this thesis.

As a benchmark for the optimisation performance, that is to make sure that the optimiser is actually working and is able to improve the objective score of transsys programs that their topology is capable for generating gene expression heterogeneity in a lattice, a random sampling approach has been employed. Starting from the same reference control parameters sets (excluding of course the optimisation control parameters) a set or random initial transsys program dynamical parameter setting has been generated of size equal to the optimisation rounds (i.e. 200). The random sampling approach has failed to generate objective scores lower that the ones of the optimiser, and for certain transsys programs the random local search optimiser has managed to generate significantly lower objective score, fulfilling its aim, which was to distinguish transsys program with a topology capable for generating spatial gene expression heterogeneity.

4.1.3 Capture spatial heterogeneity

Transsys programs that exhibit heterogeneity in factor concentrations at the lattice reactor and not in the well stirred reactor will be categorised based on their objective score after optimisation. An optimisation score threshold has been introduced for this analysis, the motivation for the level of the threshold is that at least one factor from the transsys program can have the maximum information content. According to the equation 3.5 and for lattices of the size of the refernce experiment (i.e. 300 cells) the maximum information content that a single factor can obtain is $log_2 300 \approx 8.22$ bits.

This threshold is operationally used to detect transsys programs with low objective scores that exhibit the "stripy lattice" phenomenon (refer to the glossary entry Stripy Lattice for a more formal definition). To visualise this phenomenon the factor concentration levels of a transsys program that its objective score was below the negative value of the threshold (-8.22 bits) are depicted in figure 4.1. This is an arbitrarily chosen limit as the information based score is unable to quantify spatial

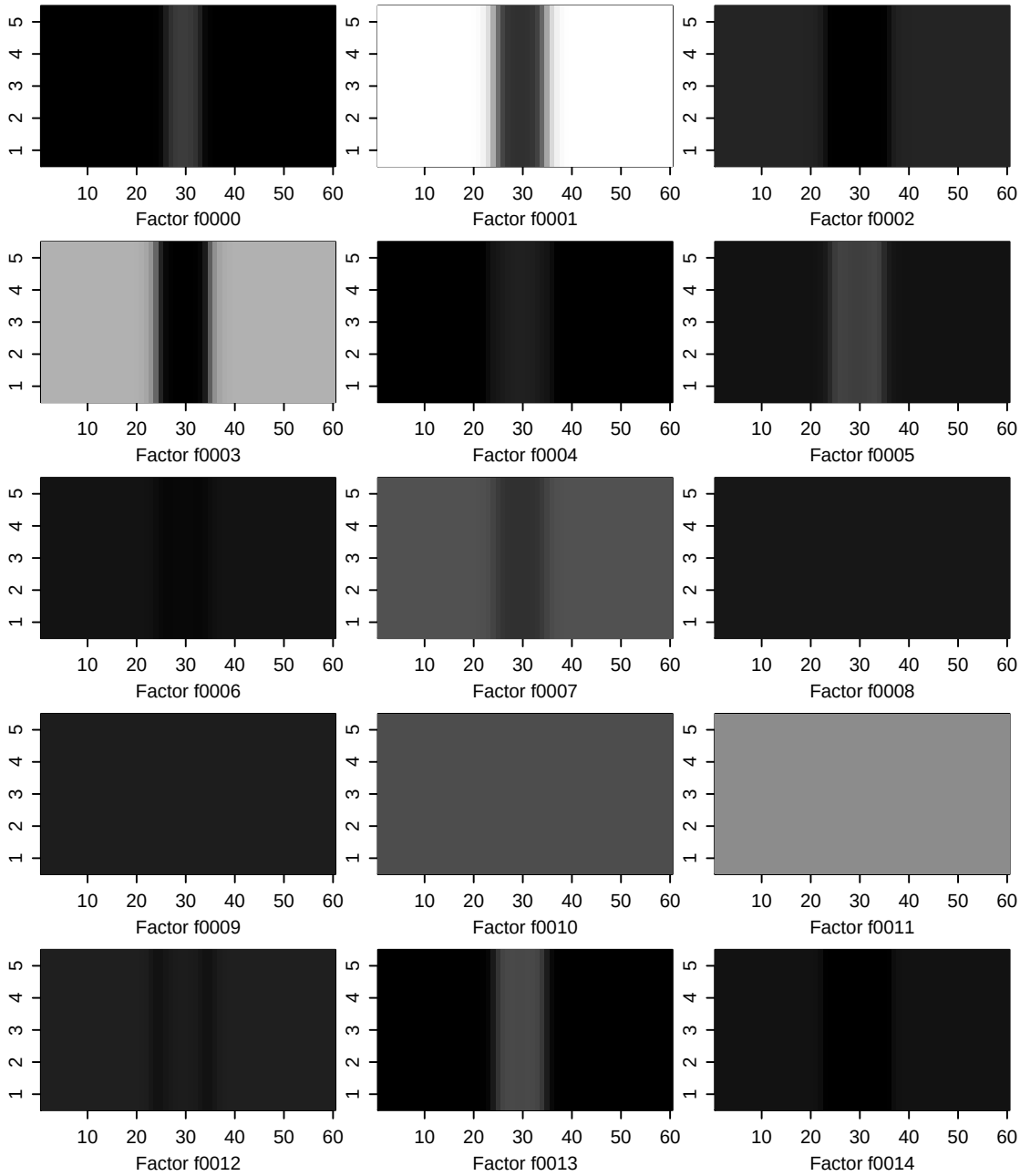


FIGURE 4.1: Greyscale images of factor concentrations from a lattice reactor for each factor of a transsys program that exhibits the "stripy lattice" property. A zone of cells has obtained high concentrations in several factors (e.g. `f0001`), forming the "stripy lattice" property. The depicted transsys program exited the optimisation procedure with objective score ≈ -8.30 bits.

arrangement of patterns and thus different spatial arrangements can have the same information based score (as illustrated in figure 3.2). However, throughout all the experiments with the 5×60 lattice whenever any gene expression heterogeneity was present it was always observed in the form of stripes on the lattice due the sort height and long width of the structure.

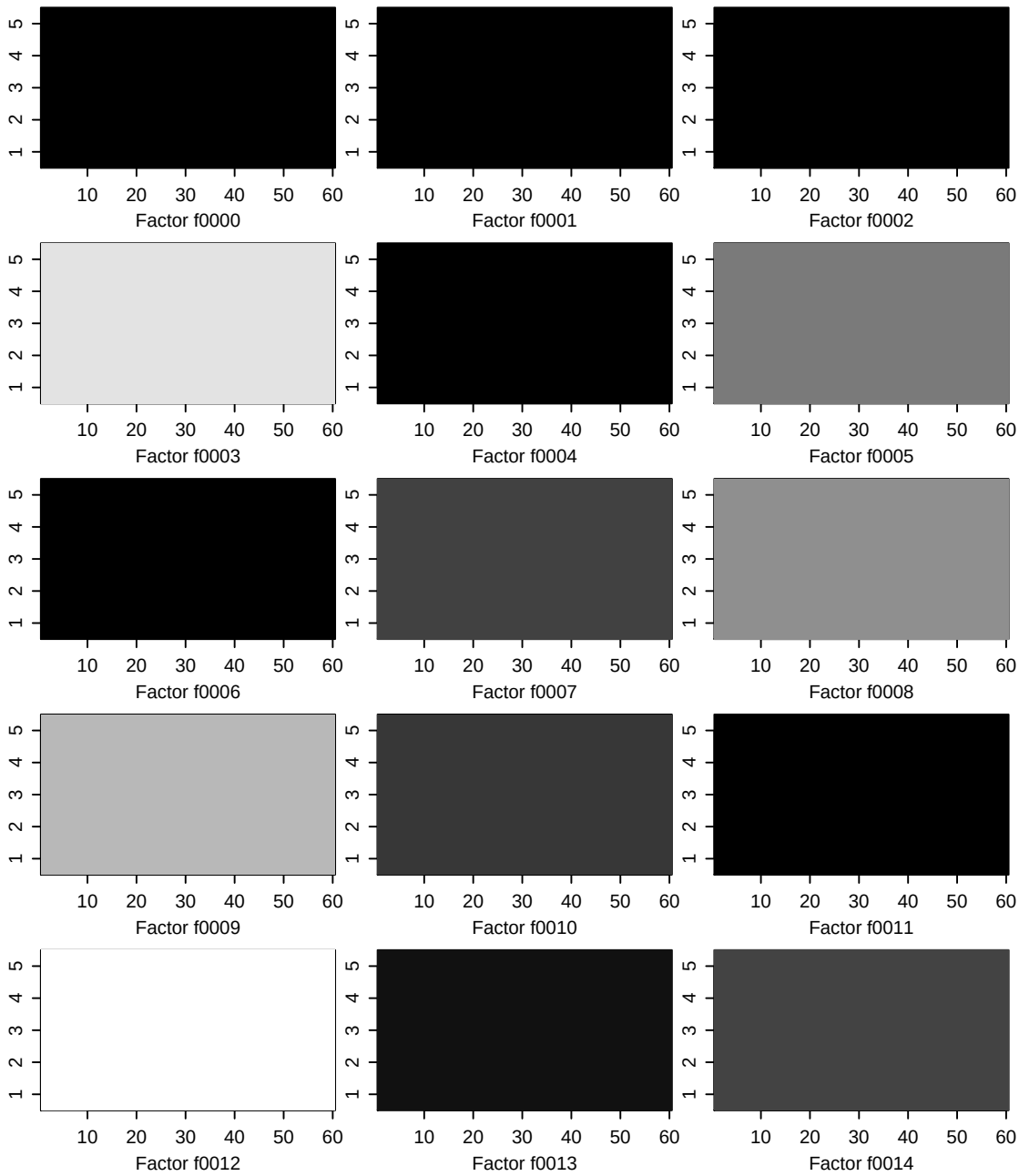


FIGURE 4.2: Greyscale images of factor concentrations from a lattice reactor for each factor of a transsys program that does not exhibit spatial heterogeneity in the factor concentrations (or it exhibits a minute one). The depicted transsys program exited the optimisation procedure with objective score ≈ -8.05 bits.

On the contrary figure 4.2 illustrates the factor concentration levels of a transsys program that exhibit a negligible amount of heterogeneity and no "stripy lattice" phenomenon can be observed. The difference in the objective scores of the two afforementioned transsys programs is 0.2 bits however the defined threshold is able to characterise and distinguish the "stripy lattice" phenomenon. Throughout the rest of this work every reference to a "stripy lattice" pattern or "stripy lattice"

phenomenon pertains to the description of this phenomenon as it is explained in this section.

4.2 Network Analyses

This thesis aims to identify topological properties of GRNs which exhibit increased gene expression heterogeneity on a spatially organised systems compared to a null model. Most of the instances of the above phenomenon came in the form of repetitive stripes of differential gene expression in a lattice reactor, thus this phenomenon will be informally termed as a "stripy lattice" (glossary entry Stripy Lattice for a full definition) in the course of this thesis. A visual illustration from a transsys program where several of its factors on a lattice are exhibiting the "stripy" pattern of factor concentration is presented in figure B.1. Networks will be characterised both as collections (or ensembles) of graphs with certain characteristics (e.g. degree distribution), or characterised in terms of topological properties of individual network. The set of network topological properties that is introduced and discussed at the literature review chapter (section 2.2) will be employed for the purpose of topological characterisation of GRNs.

4.2.1 Global Network Measures

To calculate all the global network topological properties described in section 2.2.3.1 such as the clustering coefficient and the diameter the *igraph* library for network analysis was employed. *igraph* is a set of tools to generate and represent networks and a library for calculation and analysis of topological measurements (Csárdi and Népusz, 2006). *igraph* provides interfaces to high and higher level programming languages and for this analysis the Python interface was used.

4.2.1.1 Cycles

To calculate all the directed cycles that exist in a network a series of algorithms have been devised, a directed network is the input of this algorithmic procedure

and a set containing all the directed cycles as tuples of vertex indices is the output. The algorithms are specified as follows:

Algorithm 1: Calculate all cycles on a directed graph

input : A directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

output: A set of all the cycles as tuples of vertices

allCycles $\leftarrow$ { }
reduced$\mathcal{G}$ $\leftarrow \mathcal{G}$
foreach $v \in \mathcal{V}$ **do**
    vCycles $\leftarrow$ cyclesFromVertex(reduced$\mathcal{G}, v$)
    allCycles $\leftarrow$ allCycles $\cup$ vCycles
    remove v from reduced$\mathcal{G}$
end
return allCycles

Algorithm 1 computes all the cycles in a directed graph and it is based on computations of the `cyclesFromVertex` algorithm described below:

Function cyclesFromVertex($\mathcal{G}, v$): Calculate all the cycles that pass through a vertex v

input : A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a vertex $v \in \mathcal{V}$

output: A set of all cycles as tuples of vertices

cycles $\leftarrow$ { }
paths $\leftarrow$ pathsFromVertices($\mathcal{G}, (v)$)
foreach path $\in$ paths **do**
    lastVertex $\leftarrow$ the last vertex in path
    if (lastVertex, v) $\in \mathcal{E}$ **then**
        cycle $\leftarrow$ (path + v)
        cycles $\leftarrow$ cycles $\cup$ cycle
    end
end
return cycles

The algorithm to calculate all the cycles that pass from a given vertex (as implemented in the function cyclesFromVertex) depends on the computations of all the

paths that pass from a series of vertices, described in the algorithm below:

Function pathsFromVertices($\mathcal{G}$, vSeq): Calculate all the paths that start from the first vertex of vSeq

---

input : A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a tuple vSeq containing a sequence of vertices
$:\forall v \in \text{vSeq}, v \in \mathcal{V}$

output: A set of all paths as tuples of vertices

paths $\leftarrow$ { }
lastVertex $\leftarrow$ the last vertex in vSeq
neighbourV $\leftarrow$ (all the neighbouring vertices of lastVertex,)
foreach $v \in$ neighbourV **do**
  if $v \notin$ vSeq **then**
    extvSeq $\leftarrow$ (vSeq + v)
    paths $\leftarrow$ paths $\cup$ extvSeq $\cup$ pathsFromVertices($\mathcal{G}$, extvSeq)
  end
end
return paths

The successful execution of all the algorithms described in this section will return a set of all the directed cycles of 2 or more vertices (loops, i.e. cycles of one vertex or self-regulatory interactions are no considered as cycles in this analysis) that exist in a network.

4.3 Local Network Measures

For the calculation of the all the motifs and the relevant motif profiles, as introduced in section 2.2.3.2, the *igraph* function `motifs_randesu` have been employed for size 3 and size 4 motifs. `igraph` motif finding function is based on a recently developed fast network motif detection algorithm named FANMOD which is formally described in (Wernicke and Rasche, 2006). The frequency of occurrence of each individual of these size-3 motifs in a graph defines a vector that is referred in this thesis as the 3-motif profile and is used as a characteristic signature of the graph.

In addition a measure has been devised to assess the impact of individual network elements deletions on the size-3 motifs. The measure is based on the size-3 motifs profile which is the tuple of all the frequencies of occurrence for each of the size-3

motifs. The euclidean distance between the size-3 motif profile of the transsys program and the size-3 motif profile of the transsys program after a network element deletion was calculated and serve as a measure to investigate any relationships between size-3 motifs and the loss in the objective score due to the network element deletion.

4.4 Individual Network Element Analysis

All the individual element based network properties, as introduced in section 2.2.3.3 for both nodes (genes) and edges (regulatory interactions) on a network were readily available from the relevant `igraph` functions.

An additional individual element measure has been introduced to connect studies of single graph elements (nodes, edges) with cycle measures, it is the participation of a single element in a cycle, it is defined as the the total number of cycles that an element is a member of and it can be calculated both for edges and nodes. The scores for the mutant transsys programs (both the gene knock-outs and the edge deletion) were calculated at the last step of the reference experimental procedure.

4.5 Implementation

The computational framework presented in the methods chapter (ch. 3) was developed entirely using the Python (Pyhton Softaware Foundation, 1996–2010) programming language. The reference experiment all the additional experiments presented in the next chapters were run in the UEA Linux cluster High Performance Computer (HPC) by using shell scripts to connect the processes together and to distribute the jobs in the cluster. The entire statistical analysis, report and results presentation was conducted in R (R Development Core Team, 2008).

Chapter 5

Network Topological Properties

The results presented in this chapter are an updated and extended version of the results published in the paper (Bouyioukos and Kim, 2009).
Bouyioukos, C. & Kim, J. T.
"Gene Regulatory Network Properties Linked to Gene Expression Dynamics in Spatially Extended Systems"
Advances in Artificial Life (Proceedings of the 10^{th} *European Conference in Artificial Life),*
Kampis, G. (ed.) vol. 5777/5778 LNCS/LNAI, Springer–Verlag, 2009" (in press)

5.1 Network Density Experiments

The experimental design is focused on studying the effects of network edge density on the capacity of GRNs to generate gene expression heterogeneity on the lattice and not on the well stirred reactor. Edge density is defined as the ratio of the edges a network actually has over the number of edges a fully connected graph will have (i.e. the maximum number of edges) (Diestel, 2005, Chapter 7). The glossary entry for Density contains a formal mathematical definition of edge density for directed graphs where, like GRNs, self-regulatory interactions (loops) are allowed.

The experiment was conducting by generating a population of random networks starting from a relatively small number of edges, gradually increase the number of edges by a step of 2 and run all the experimental process described in chapter 4

by using all the rest of the control parameter settings (apart from the number of edges) and the transformers equal to the reference values specified in section 4.1.1. The number of genes was kept equal to the reference value (i.e. 15) and the number of edges vary from 16 to 72 by a 2 edges increment step, resulting to network density varying from $16/15^2 = 0.071$ to $72/15^2 = 0.32$. For each edge density level 4 random network topologies were generated by the ER process and 4 topologies by the PL process. For each of these topologies 3 different initial transsys dynamical parameters settings have constructed. To summarise for each edge density level 2 generation mechanisms $\times$ 4 topologies $\times$ 3 parametrisations = 24 transsys programs were generated and as there are 29 different levels of density the total experiment includes 29 density levels $\times$ 24 transsys programs per level = 696 transsys programs in total.

The objective score for each transsys program after optimisation was correlated with the network density. The correlation plots (figure: 5.1) and the Spearman rank correlation coefficient $\rho = -0.467$ with a p-value ($\approx 10^{-38}$), suggest a significant correlation between lower objective scores and network density. In fact as the density increases the objective score significantly decreases. In addition more and more networks exhibit a lower objective score and thus higher heterogeneity on the lattice than in the well stirred reactor. To make the latter finding more illustrative boxplots of the same data are presented in figure 5.2.

Boxplots of objective scores (figure 5.2), illustrate that the number of transsys programs with lower objective scores out of the total for each density level is increasing as the edge density increases. The median is getting decreased and the sizes of the boxes (representing the inter-quartile range) are increasing as the density increases. The finding is justified as increasing the number of regulatory interactions in a GRN increases the complexity of its dynamical properties, therefore the set of dynamical properties of lower density networks is included (i.e. is a subset) of the dynamical properties of more dense networks. It needs to be noted that a positive bend is observed in the objective scores as the density reaches the highest levels in this experiment. However, as little is known about the nature of the objective function landscape the settings of the current optimisation approach –including the type of the optimiser, the optimisation rounds and the optimisation offset– might not be the optimal for optimising transsys programs with density higher than 0.32. In addition to nature of the fitness landscape, the cardinality of the space of all possible networks increases dramatically as the edges increasing and the current sampling size might also impose limitations to the potential of low objective score.

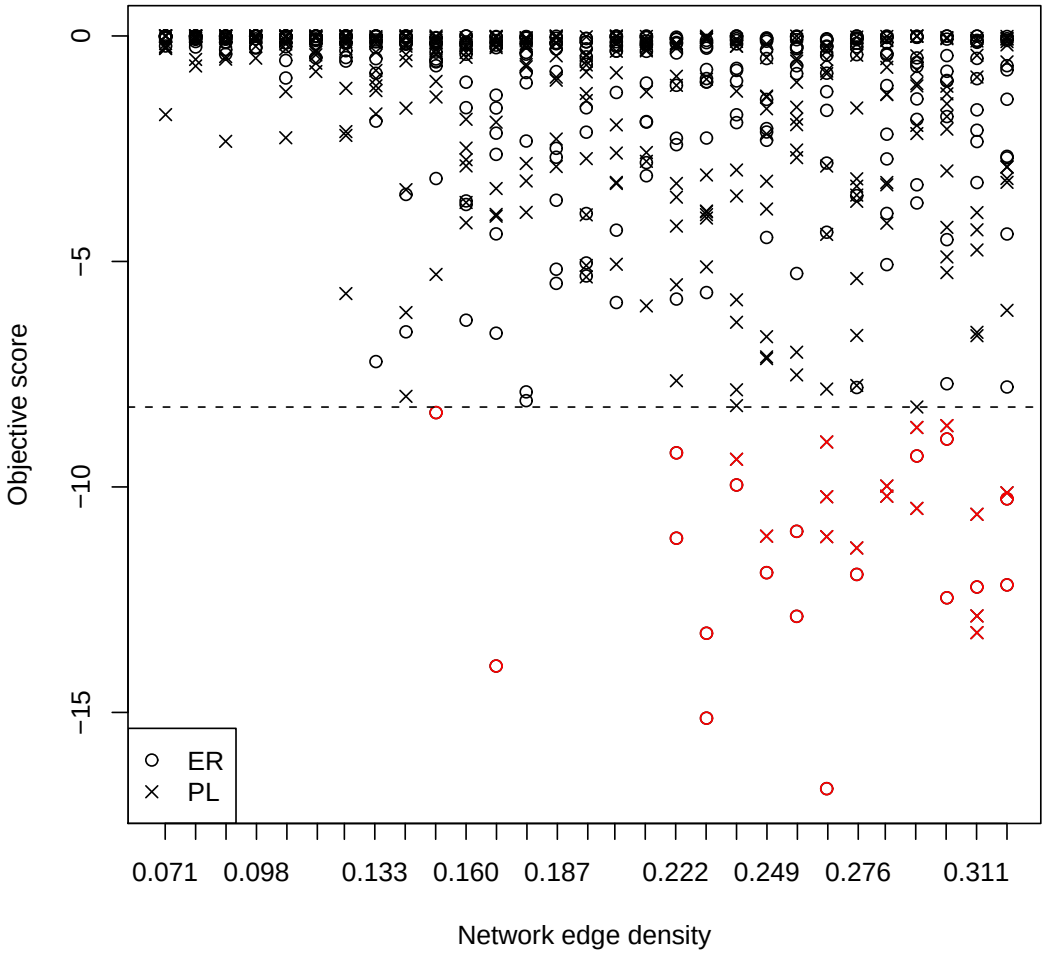


FIGURE 5.1: Scatterplot of transsys programs objective score after optimisation vs. network edge density for the 696 transsys programs of the reference experiment set. Circles designate transsys programs which their network topology has generated by an Erdós & Reyní process (ER) and × transsys programs with power law network degree distribution (PL). The dashed line designates the operational threshold for "stripy lattice" and networks which exhibit this property are coloured red. The Spearman correlation coefficient ρ is -0.467 and p-value $\approx 10^{-38}$. Network density is negatively correlated with low objective scores.

The density findings are in a partial agreement with previous studies of density. Most notable are the studies of edge network density in the work of S. Kauffman, where he identified a threshold of K=2 for NK networks to begin exhibit properties characteristic for biological systems such as homoeostasis and differentiation (Kauffman, 1993). These dynamical properties disappeared as the number of incoming edges K exceeds 3 (Kauffman, 1969b), this means that a gene should have

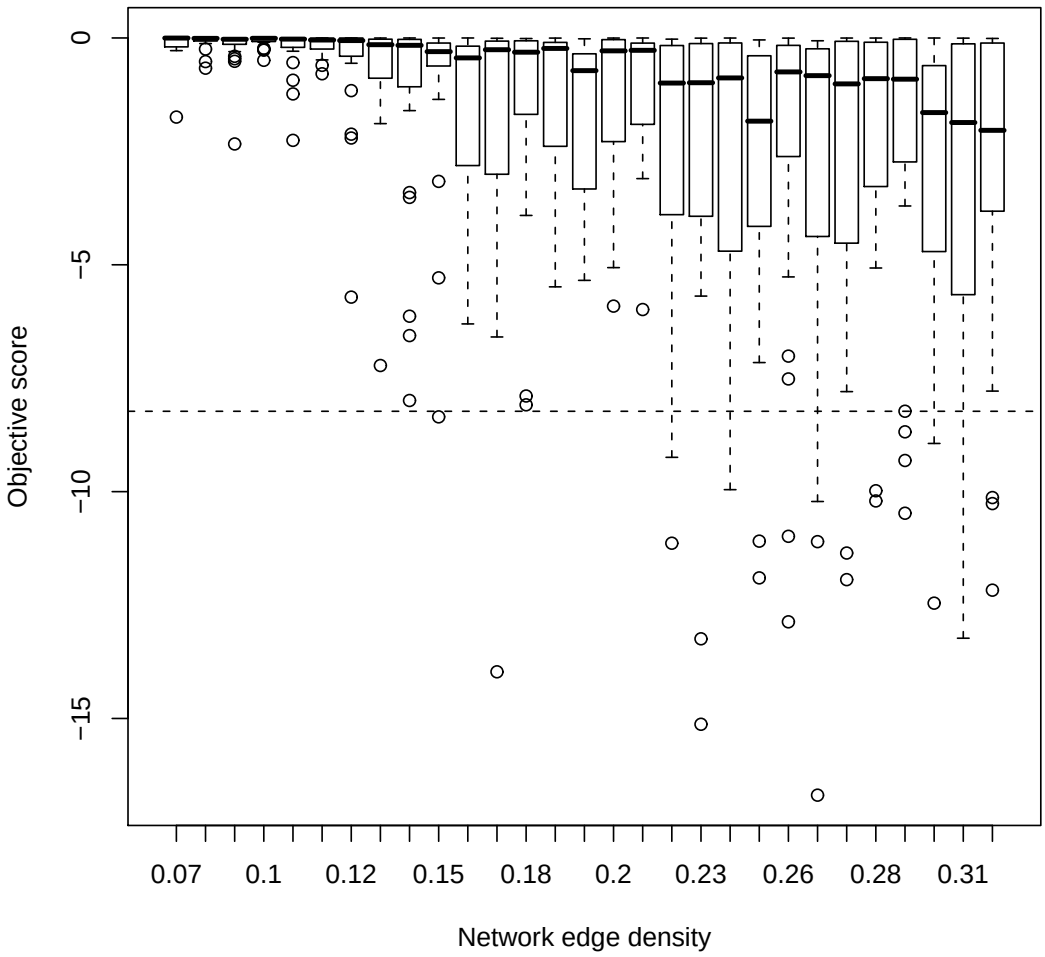


FIGURE 5.2: Boxplots of objective scores after optimisation for all the 696 transsys programs of the reference set at different network edge density levels. The medians for each density level are lower for higher densities and the number of low scoring transsys programs –depicted as outliers in the boxplots– is increasing as the density increases. The horizontal line depicts the operational threshold for stripy lattices introduced in section 4.1.3

an average number of regulatory interactions between 4 and 6. Although different network generation mechanisms were used in this study and the dynamical properties of the NK networks is determined by the complexity of the Boolean functions, in the experiments presented here (with ER and PL topologies) the density range for a network to exhibit low objective score is between 0.14 to 0.28. This density levels correspond to an average degree between 4 and 8.

5.1.1 Connection with the low objective score patterns

Transsys programs which their networks have obtained low objective score have also exhibit the "'stripy lattice" property (section 4.1.3). Indicative results are illustrated in the following figures: In figure 5.3 an example of a transsys program with density $d = 0.1244$ shows no sign of heterogeneity in the greyscale images of all its factor concentrations.

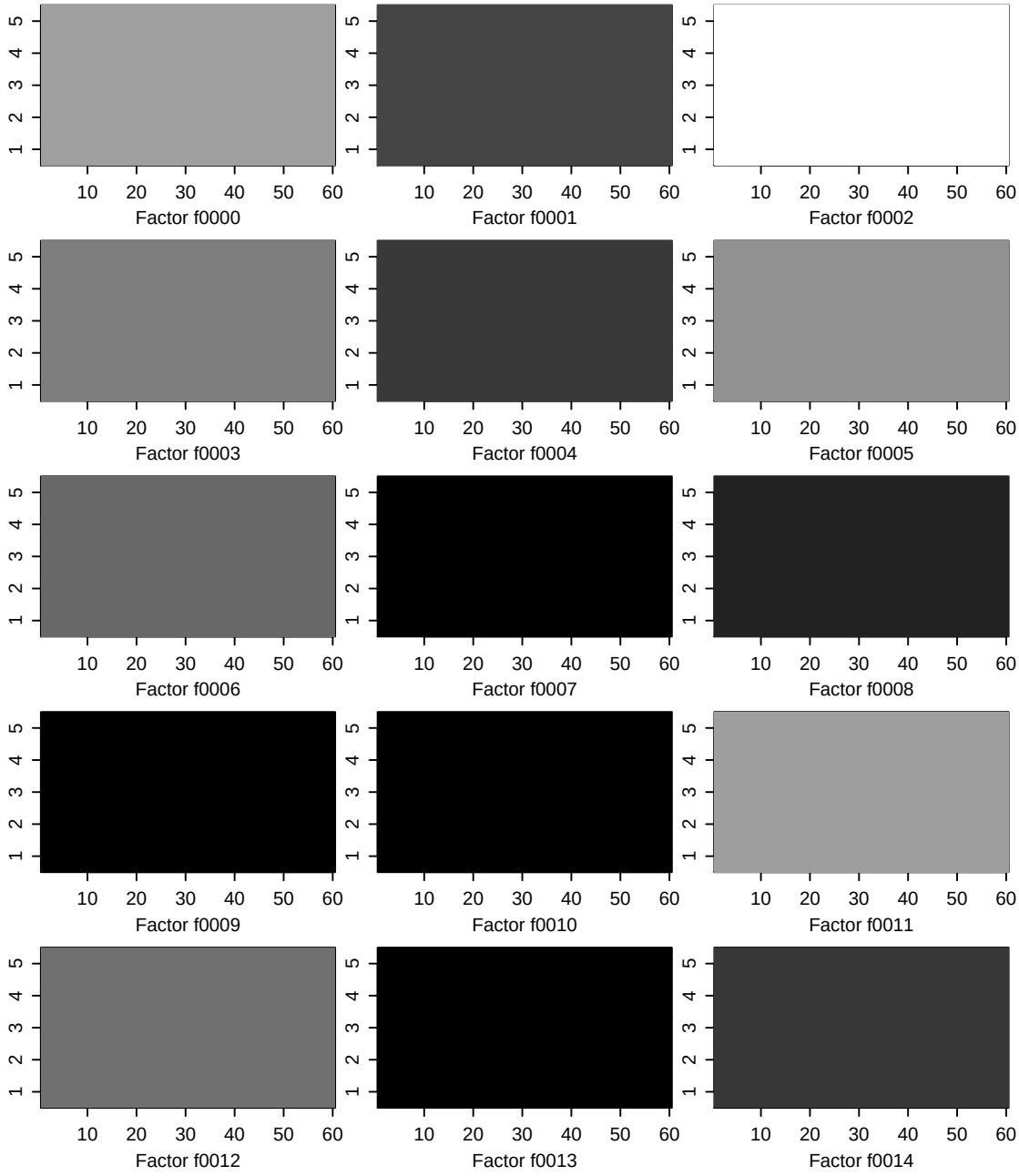


FIGURE 5.3: Greyscale images of factor concentrations from a transsys program with 0.1244 density. The objective score is -0.002 and it is well above the threshold (designated with the dashed line in figure 5.1.

In figure 5.4 a selected transsys program that is below the operational threshold for the "stripy lattice" is illustrated a considerable degree of heterogeneity can be observed in several of its factors. The selected transsys program has got relatively medium density $d = 0.222$ and show heterogeneity in some of its factors. Transsys programs from within this medium region of density have been used for most the experiments in this thesis as well as to form the reference control parameter set (section 4.1.1).

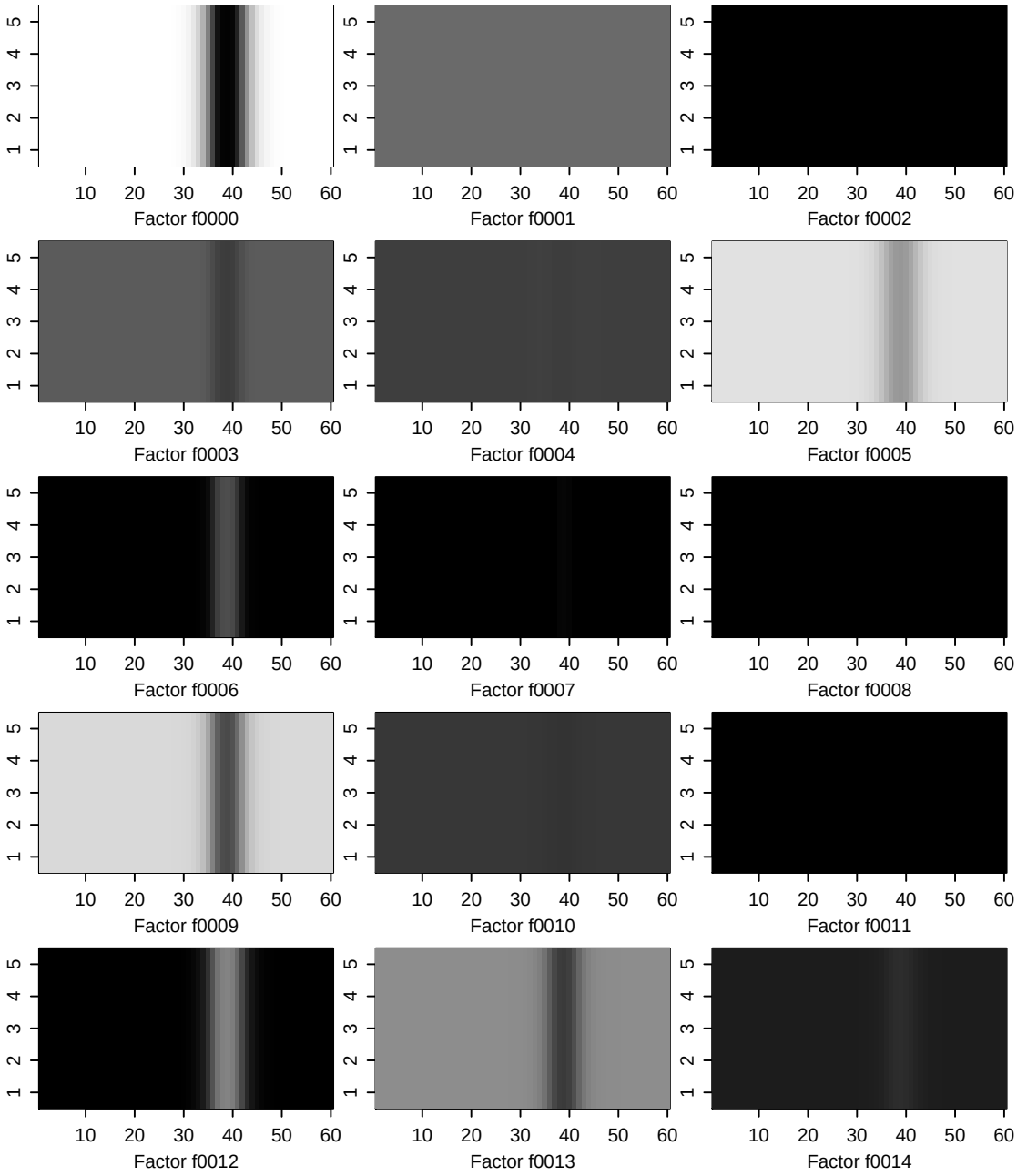


FIGURE 5.4: Greyscale images of factor concentrations from a transsys program with 0.2222 density. The objective score is -9.24 and below the threshold (designated with the dashed line in figure 5.1.

Finaly a transsys program from the high part of the edge density range was selected. In figure 5.5 the factor concentration heterogeneity of the transsys program that obtained the highest objective score after optimisation is illustrated. Most of the factors in this transsys program exhibit the "stripy lattice" phenomenon. The edge density for this transsys program is $d = 0.266$

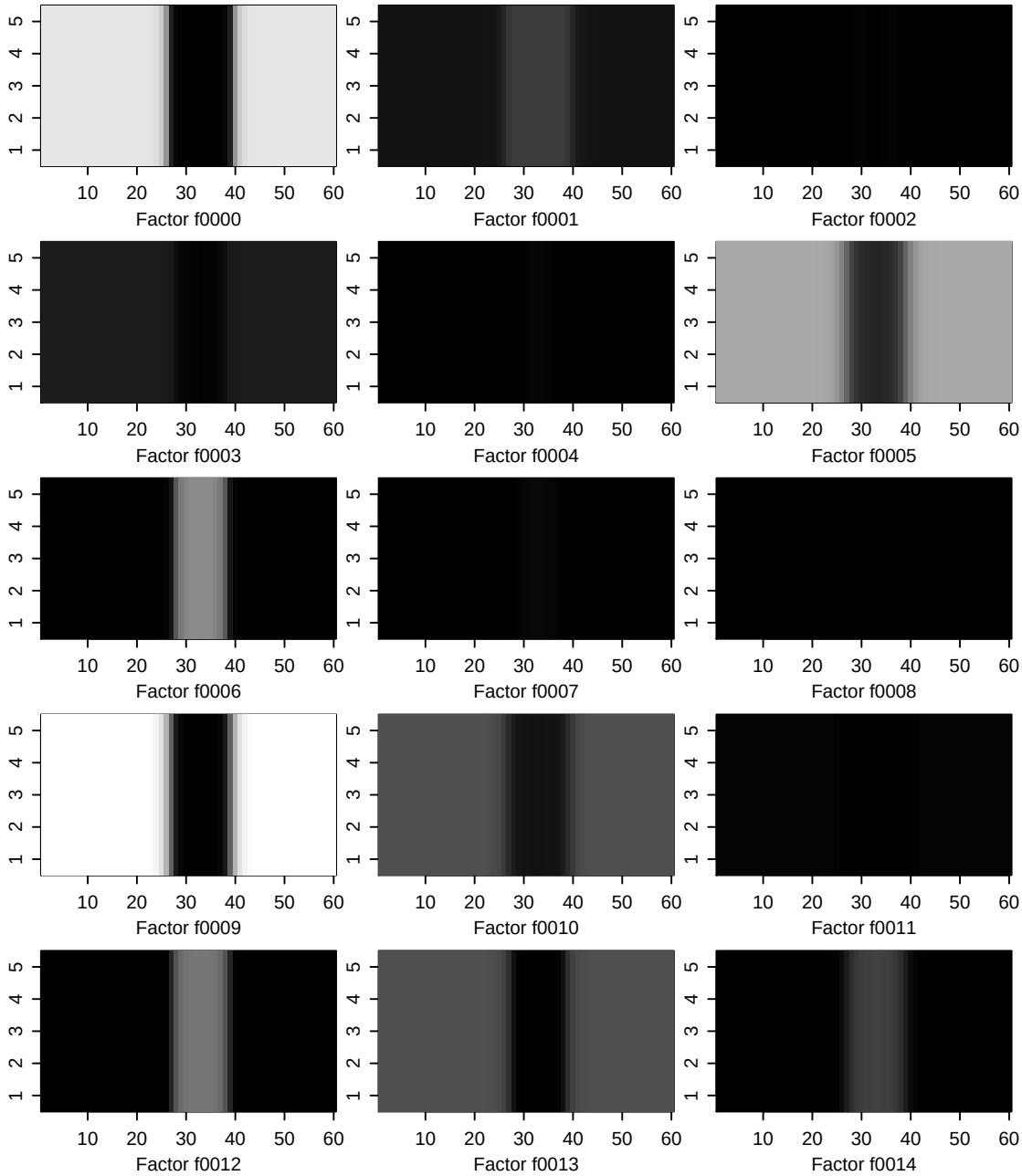


FIGURE 5.5: Greyscale images of factor concentrations from a transsys program with 0.266 density. The objective score is -16.69 and well below the threshold (designated with the dashed line in figure 5.1. The highest scoring transsys program in the density experiment has most of its factors in a heterogeneous state, exhibiting the "stripy lattice" phenomenon.

5.2 Global Network Properties

The relationships of global topological network properties with the capacity of GRNs to have low objective scores have been consecutively studied. This study is based on results generated by using the reference control parameter settings (section 4.1.1) with the only deviation from these settings being the number of different network topologies which has increased from 15 to 20 and all the parameters were kept intact. Connecting static network topological properties of GRNs with dynamical properties the networks exhibit, is an ongoing target in network biology (Fox and Hill, 2001; Kuo, Banzhaf, and Leier, 2006) and to address that question an experiment was designed as follows.

A population of random networks has been generated, comprising 20 ER and 20 PL network topologies –here the number of different network topologies has been increased compared to the reference set from 15 to 20 to obtain a larger sample of the network topologies space– this constituted the only deviation from the reference parameters set for this experiment. The total population of transsys program consisting of 2 generation mechanisms $\times$ 20 network topologies $\times$ 30 parametrisations equals to 1200 transsys programs. The network topological properties included in the study comprise the clustering coefficient, the diameter, the total number of cycles and the average cycle length of a network. These topological properties, introduced in section 2.2.3.1, were calculated for each network topology using tools and algorithms described in section 4.2. Correlation studies of each measurement against the objective score of the transsys program after optimisation were conducted and the results are presented and discussed here.

The network generation mechanism, either the ER or the PL generation process did not have any significant impact on the transsys program objective scores. The notched boxplots or the ER and the PL networks (figure 5.6) show an overlap on the notches of the boxes between ER and PL generated networks. Overlap between notches suggests that there no significant difference between the medians of the two objective score distributions.

The Wilcoxon rank sum test has been used to corroborate further the finding, it is a non-parametric test and has been chosen as the distributions of the objective score values are not normal. The Wilcoxon test checks for a location shift between the two distributions and returns the Wilcoxon statistic W and the associated p-value. The Wilcoxon test results for ER and PL obtained objective scores are: $W = 187442$, p-value $= 0.2151$. A p-value of 0.215 suggests that there is no

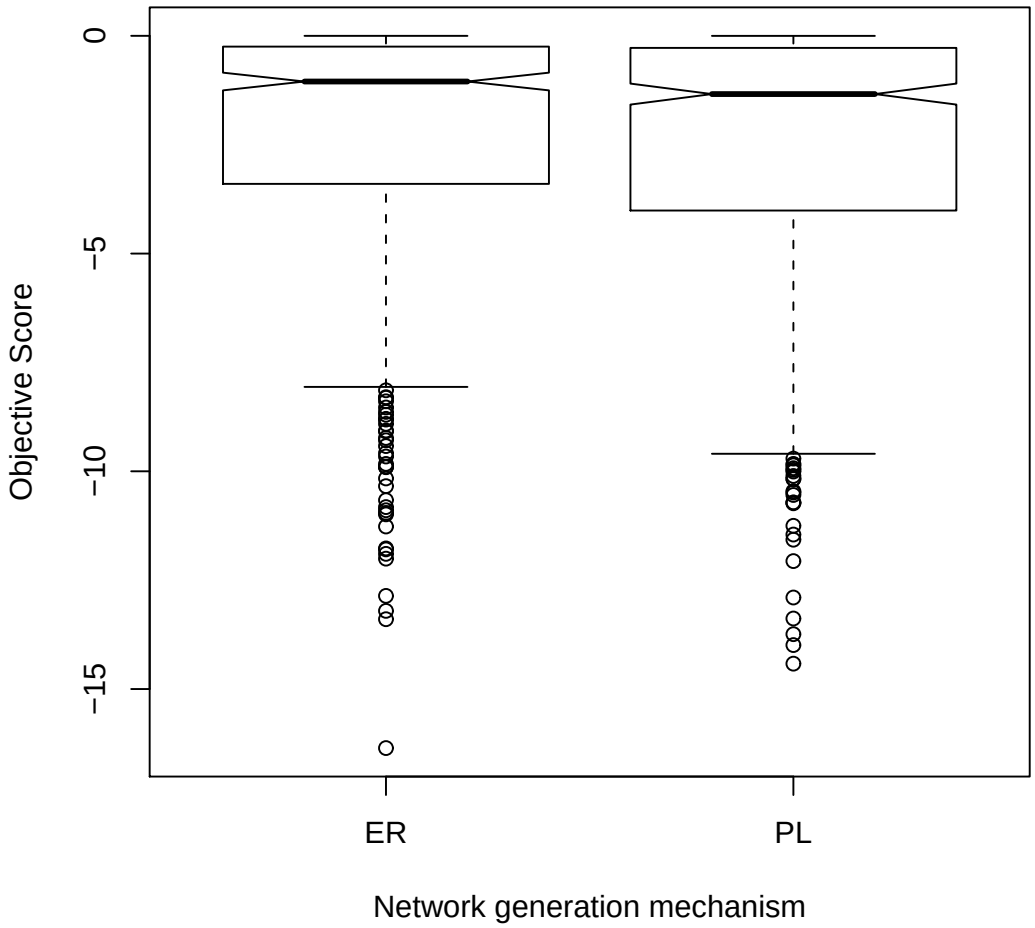


FIGURE 5.6: Notched boxplots of objective scores of transsys programs from the reference parameters set after optimisation. The networks have been generated by an ER and a PL process. No significant difference is observed between the objective score medians of the two network generation procedures.

statistical significance in the difference of the objective score distributions locations between ER and PL generated networks. Suggesting that the network generation process does not affect significantly the objective score of transsys programs after optimisation.

The clustering coefficient is a measure for the degree of cliqueness and the density of triangles. Correlation studies of clustering coefficient vs. the objective score found no correlation between those two. The Spearman rank correlation coefficient ρ has been found very low ($\rho = -0.01$) and the p-value of the correlation very high (p-value $= 0.868$). Therefore, as is it also illustrated in the scatter-plot of

figure 5.7, no correlation has been observed between a transsys program objective score and its clustering coefficient.

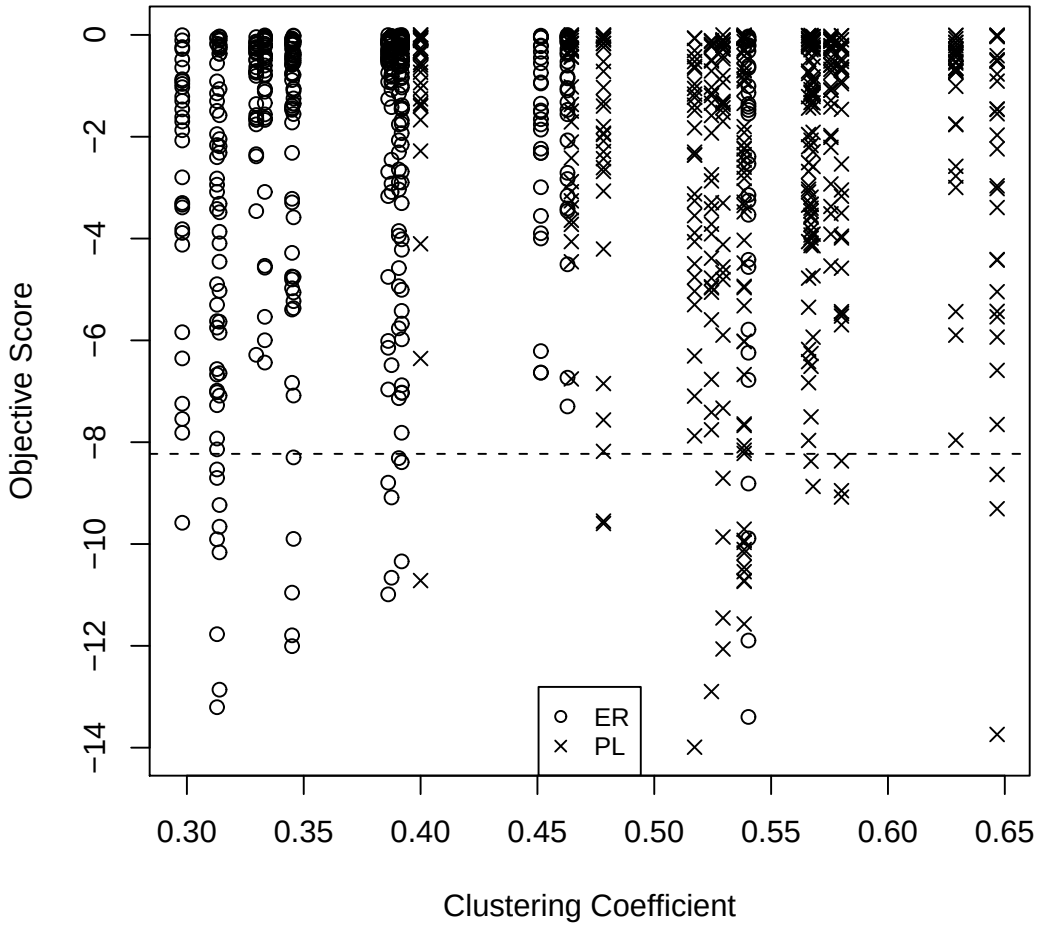


FIGURE 5.7: Correlation scatter-plot of transsys program objective scores after optimisation against the network clustering coefficient. For the combined reference experimental set no association is observed clustering coefficient and objective score after optimisation. The Spearman ρ is -0.01 and a p-value = 0.882 does not support any asociation between clustering coefficient and objective score. The dashed line illustrates the operational threshold for the "stripy lattice" phenomenon as introduced in section 4.1.3.

The contrary finding holds for the next global network property in the analysis, the network diameter. As diameter takes only discrete values the diameter measures have been grouped to each diameter level and thus notched boxplots of diameters and transsys program objective scores after optimisation has been prepared and presented in figure 5.8. The objective score is weakly correlated with the diameter

–Spearman rank correlation coefficient $\rho = 0.12$– and a low p-value ($\approx 10^{-4}$) support the dependency.

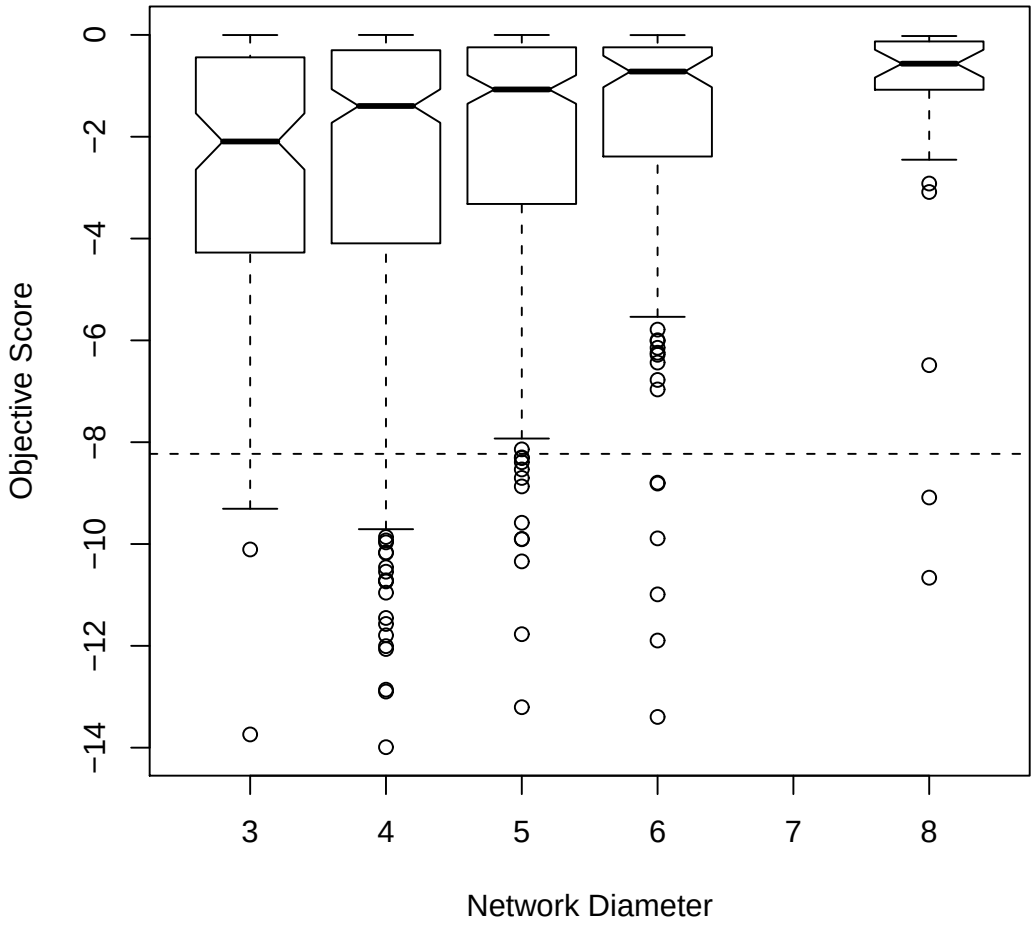


FIGURE 5.8: Notched boxplots of transsys program objective scores. Each boxplot contains objective scores of transsys programs which have the same network diameter –thus each boxplot contains different number of transsys programs. The networks with smaller diameter have lower median objective score and more transsys programs under the operational threshold for the "stripy lattice" phenomenon. The Spearman correlation $\rho = 0.12$ with a p-value of 2.72e-04, supports association between small network diameters and low objective scores.

Figure 5.8 illustrates a trend that networks with smaller diameter have lower objective scores and thus higher propensity to generate heterogeneity in lattices and not in the well stirred reactor, or exhibiting the "stripy lattice" phenomenon. Smaller diameters as well as higher clustering coefficients are characteristic of the small world networks described in (Watts and Strogatz, 1998) and biological networks

are among other networks that share the small world characteristics according to (Milo, Itzkovitz, Kashtan, Levitt, Shen-Orr, Ayzenshtat, Sheffer, and Alon, 2004). Although there has been reported from empirically generated networks that a relatively high clustering coefficient is a characteristic of biological networks the findings of the analysis here did not reveal any significant relationship between clustering coefficient and transsys program objective scores. The correlations of two of the small world properties, (i.e. the clustering coefficient and the diameter) with transsys programs objective score suggest that, only the diameter and not the clustering coefficient of a GRN is suitable to be a predictor for generating spatial patterns on lattices. The last argument can also be supported by the fact that the clustering coefficient is a well defined measure only for undirected graphs (and is not unambiguously defined for directed (Fagiolo, 2007)) and undirected graphs are not adequate representations of GRNs.s

Cycles in biological networks have been studied theoretically by R. Thomas (reviewed in section 2.2.3.1). To study potential associations between the number of cycles in a network and the objective score after optimisation a correlation analysis has been conducted and the scatter-plot in figure 5.9 illustrates the findings.

No correlation has been retrieved between the total number of directed cycles in a network and the objective score value of a transsys program after optimisation. Both the Spearman rank correlation coefficient ρ and the p-value do not suggest any significant correlation. The number of cycles per se as an aggregate measure can not reflect the dynamical property of "stripy lattice" that is measured by the objective score value.

Investigating further the potential effect of cycles on the dynamics of gene expression separate correlation studies have been conducted to explore associations of the number of positive and negative cycles -separately- on the transsys objective scores after optimisation. The effect of positive cycles on cell differentiation and multistationarity has been formally studied by (Thomas and D'Ari, 1990) and discussed in section 2.2.3.1. Figure 5.10 shows correlation scatter-plots of positive and negative, no correlation can be observed between the number of positive cycles and the objective scores (p-value $= 0.845$) neither between negative cycles number and objective scores (p-value $= 0.625$). These results can not support any role for the number of cycles in generating gene expression differentiation in lattices.

A comparable result to the number of cycles has been obtained from correlation studies of the average cycle length measure. Again the correlation scatter-plot of the average length of cycles on a network against the objective score of transsys

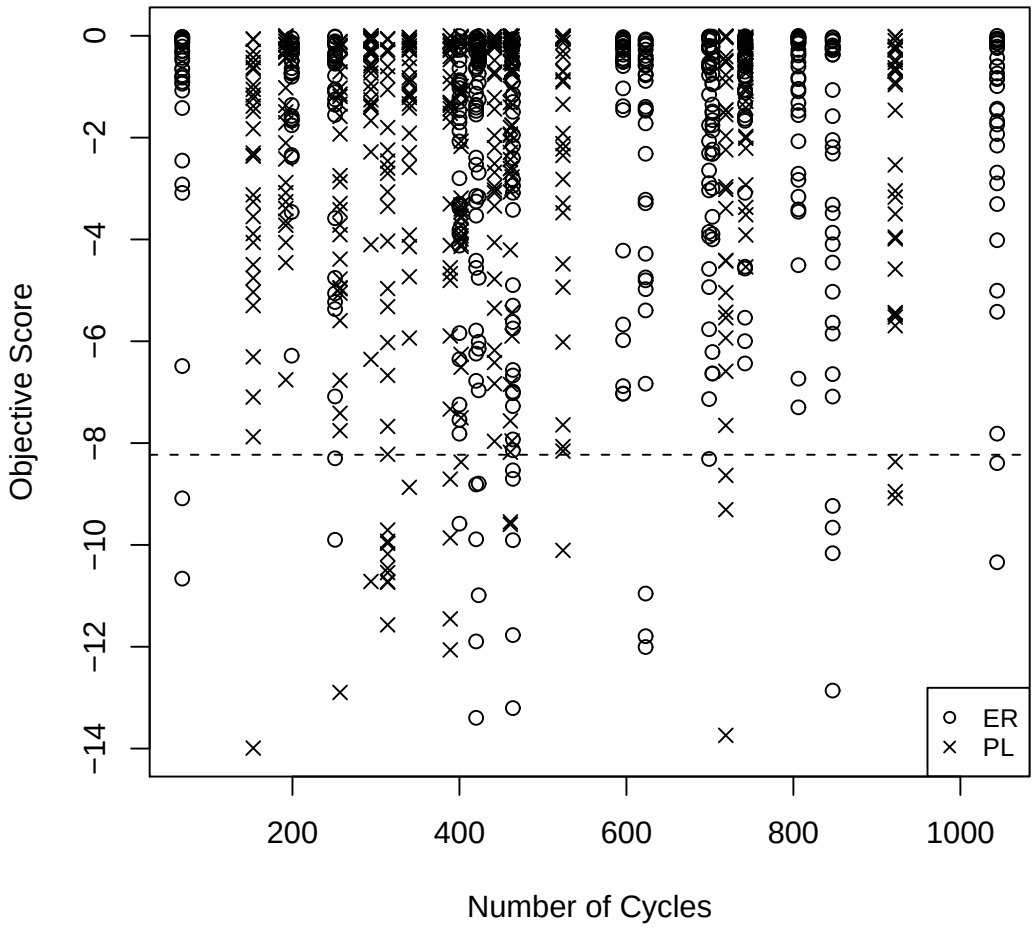


FIGURE 5.9: Correlation scatter-plot of the total number of cycles per network against transsys program objective scores after optimisation. No amount of correlation has been found (p-value = 0.844 and Spearman $\rho = 0.006$). The dashed line represents the operational threshold for the "stripy lattice" phenomenon as defined in section 4.1.3

programs after optimisation saw no correlation, as it is illustrated in figure 5.11 and the Spearman ρ rank correlation coefficient and p-values are very low and very high respectively.

The two aggregate network cycle measures that were studied, the number of cycles and the average cycle length of the network, do not constitute adequate predictors of a GRN's objective score. Both the cycle based measures studied can not be associated with the dynamical property of "stripy lattice" of several GRNs.

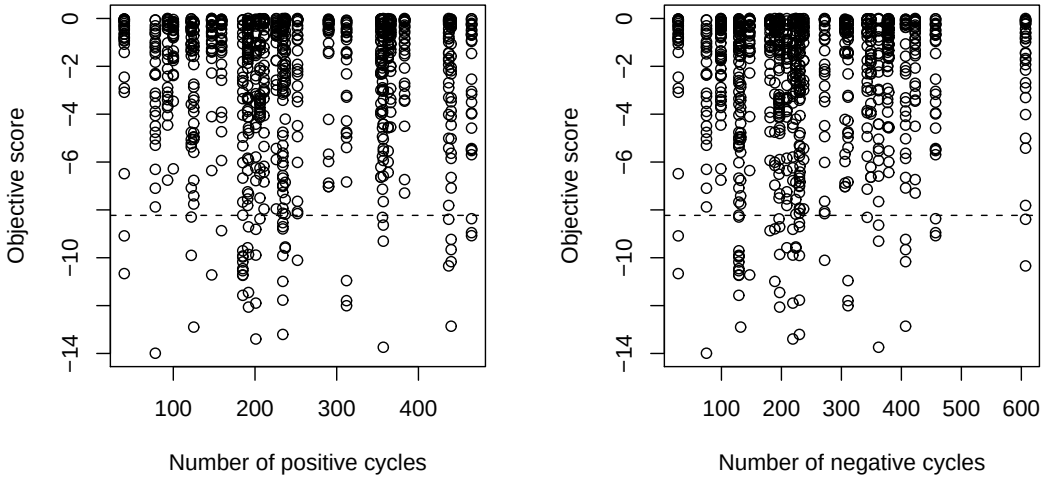


FIGURE 5.10: Correlation scatter plots of number of positive (left) and negative (right) cycles against the objective score after optimisation for each transsys program of the reference parameter set. No correlation has been found (p-value = 0.845 for the positive and p-value = 0.625 for the negative cycles correlations respectively). The dashed line depicts the operational threshold for the "stripy lattice" phenomenon as defined in section 4.1.3

5.3 Individual Elements Properties

The same dataset described in the previous section (section 5.2) was used to study the impact of individual element network properties. For every transsys program generated for the study of the global network properties the reference experiment of a full set of single network element deletion was conducted. Single element deletion experiments are designed to assess the impact of individual network elements (genes and regulatory interactions) on the objective score and the methodology is described in section 3.6.1. The difference in the objective score from the original transsys program is correlated against network properties of the deleted element.

5.3.1 Gene properties

Correlation studies of individual vertex (gene) network properties against the objective score difference of the single gene knock-out transsys program from the wild-type have been carried out. The studied measures include gene centrality measures and cycle related measures these are the degree, the closeness, the betweenness and the eigenvector centrality as well as the number of cycles a gene

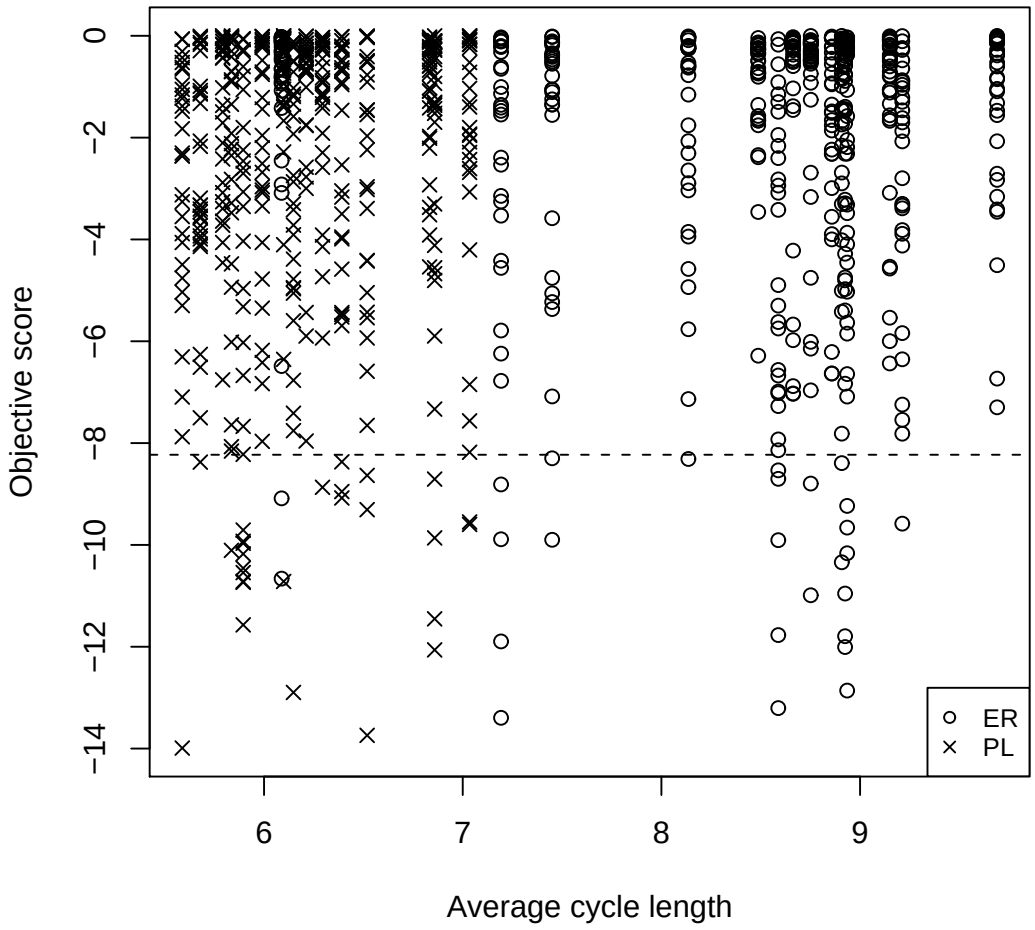


FIGURE 5.11: Correlation scatter-plot of the average length of all the cycles in a network against the objective score after optimisation. No significant correlation can be reported as the p-value $= 0.317$ and the Spearman $\rho = 0.033$. The dashed line represent the threshold for the "stripy lattice" behaviour.

is a member of (introduced in section 2.2.3.3). Each transsys program in the set generated in the section 5.2 after optimisation has entered the single gene knock-out process as described in section 3.6.1. Then correlations of the difference in the objective score of the single gene knock-out from the wild-type transsys program against network related measures of the knocked-out gene were examined and presented here.

Gene centrality measures are correlated with the objective score loss due to single gene knock-out in the majority of the transsys programs that exhibit a relatively low objective score. Transsys programs with a substantially low objective score which is defined as less than the maximum information content that a single

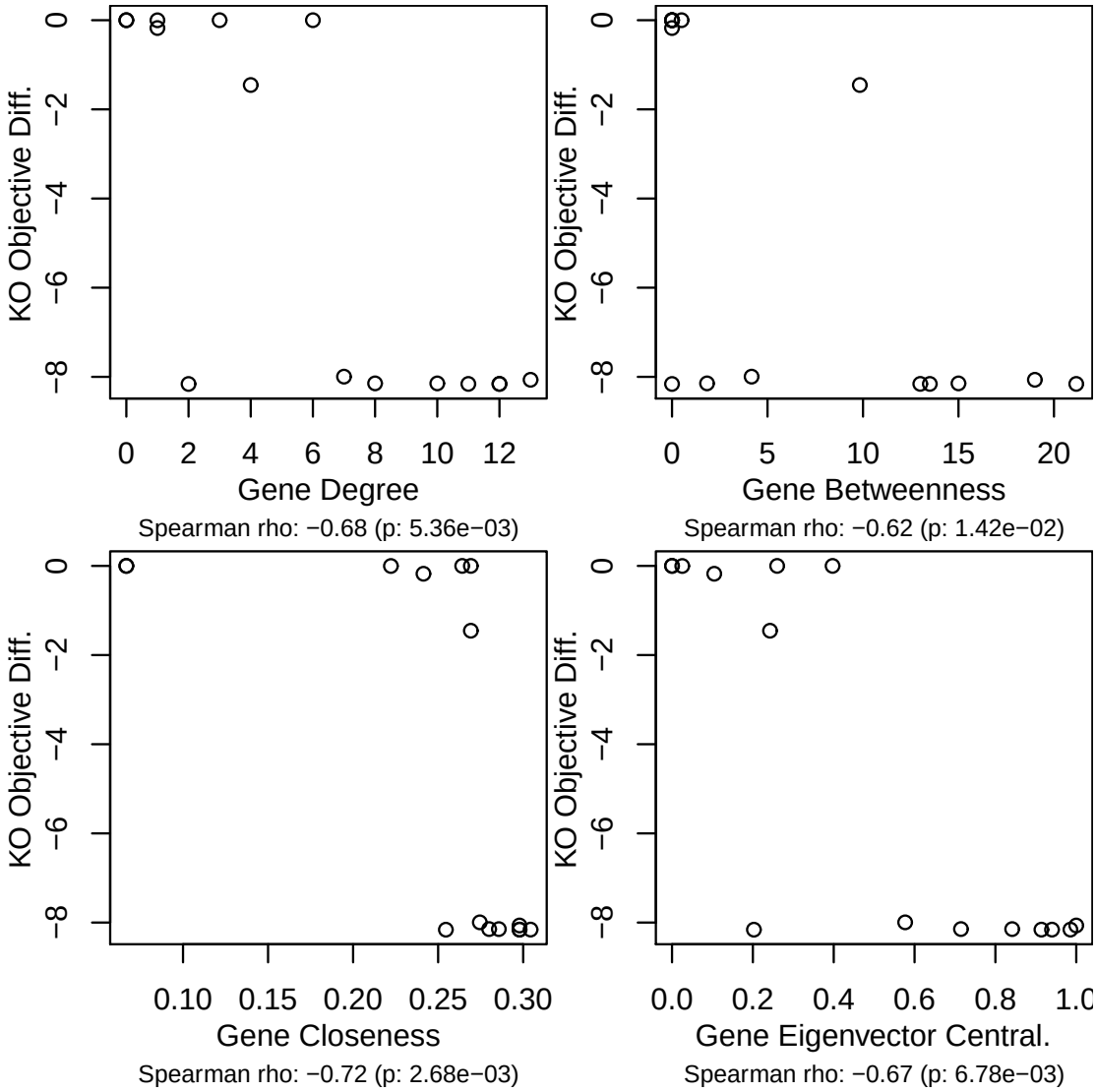


FIGURE 5.12: Scatter-plots of the objective score difference of the single gene knock-out from the wild-type transsys program against gene network centrality measures. The plots are drawn from a transsys program that has objective score below the operational threshold depicted in figure 5.7.

factor can have in a lattice of 300 cells, that is $-\log_2 300 \approx -8.228$ and has been introduced as an operational threshold for the "stripy lattice" phenomenon in section 4.1.3, were checked for gene network measures correlations. Out of the 1200 transsys programs, a set of 77 had a substantially low objective score lower or equal to -8.228. Out of this selected transsys programs, the correlation plot of a characteristic example is presented in figure 5.12.

All the centrality measures are significantly correlated with the objective score loss, indicating that centrality measures of a gene on a network are reliable predictors of the contribution of a particular gene to gene expression heterogeneity on spatially organised systems. A second selected transsys program is also appearing to have

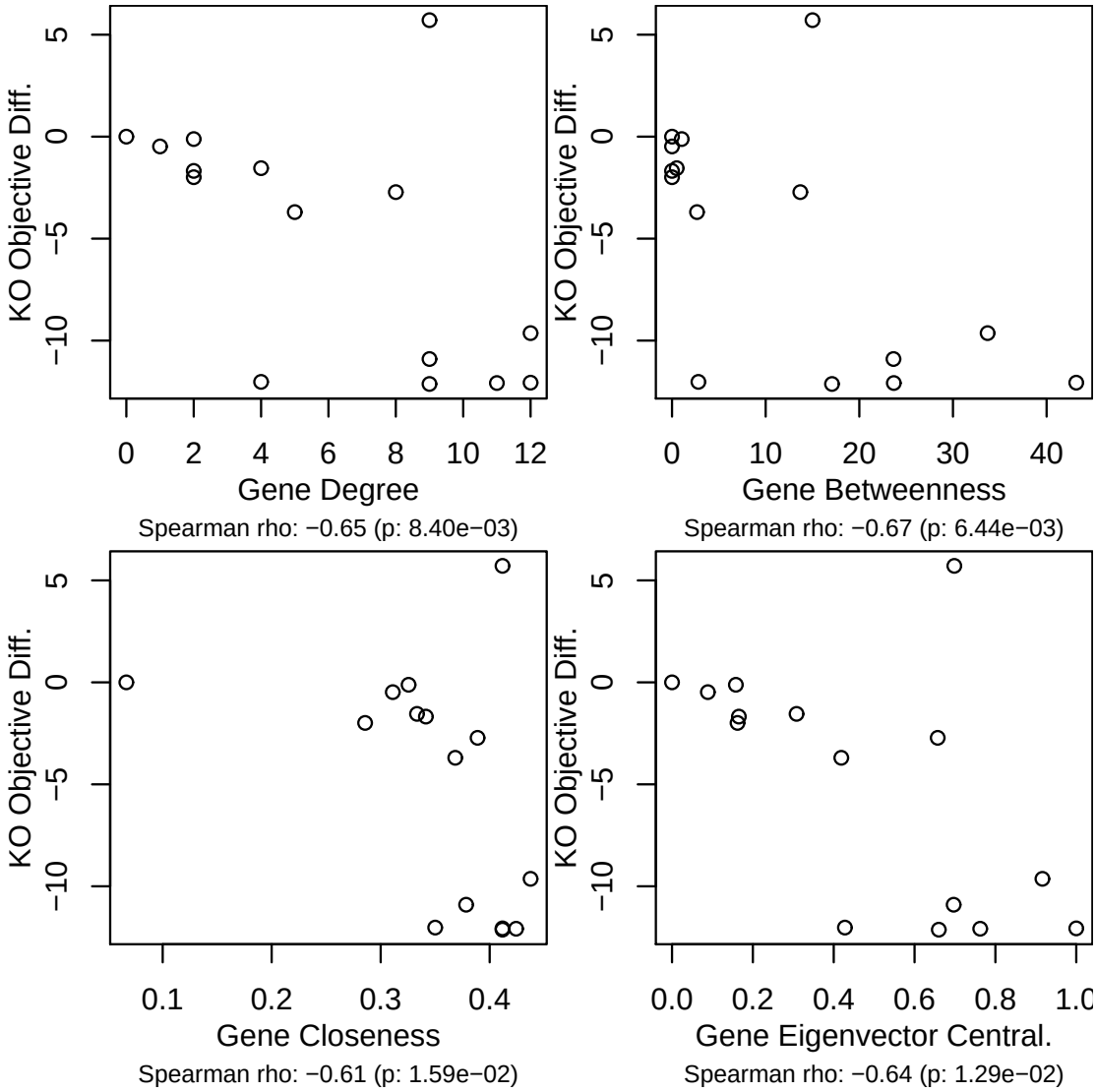


FIGURE 5.13: Scatter-plots of the objective score difference of the single gene knock-out from the wild-type transsys program against gene network centrality measures. The plots are drawn from a transsys program that has objective score below the operational threshold depicted in figure 5.7.

the centrality measures correlated as it is depicted in figure **??**. The two transsys programs that have been selected for the plots in figures 5.12 and **??** are the two transsys programs with the lower objective score of all the 1200 transsys programs used in this experiment. Gene centralities have been brought to interest in the analysis of GRNs relatively recently (Koschützki and Schreiber, 2004,0) and the results of this section are in line with the findings of this previous work. More interestingly gene centralities have been found to be positively correlated with the rate of evolutionary change as well as the gene expression variability in networks of yeast transcription factors (Jovelin and Phillips, 2009). In the experiment of this thesis centralities have been correlated with gene expression

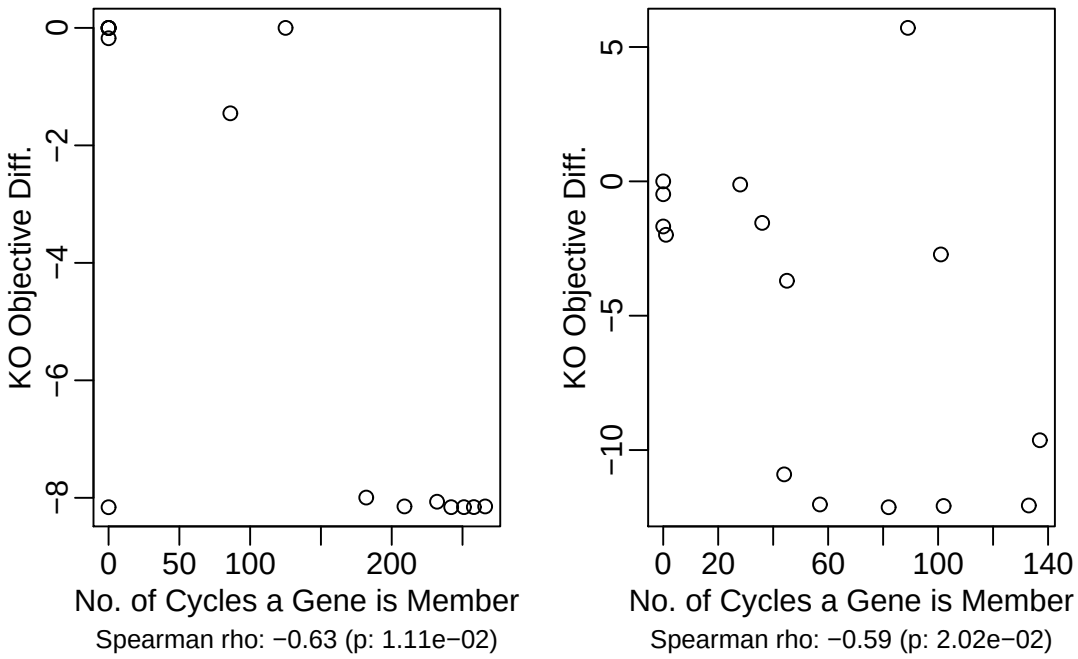


FIGURE 5.14: Scatter-plot of the objective score difference of the single gene knock-out from two wild-type transsys program against the number of cycles a gene is a member of. The plot is drawn from the same two transsys programs used for figures 5.12 and 5.13, that have objective scores below the operational threshold depicted in figure 5.7. The number of cycles a gene is a memeber of is correlated with the objective score loss.

heterogeneity in networks that generate heterogeneity, finding close to the gene expression variability correlation that the (Jovelin and Phillips, 2009) study has proposed.

The number of cycles that a gene is a member of, correlates with the objective score loss of the single gene knock-out. Figure 5.14 illustrates the scatter-plot of the objective score differences of the single gene knock-outs from the wild-type transsys program against the number of cycles a gene is a member of. The Spearman correlation coefficient ρ suggests a negative correlation ($\rho = -0.63$) between number of cycles and the objective score loss and the p-value supports the statistical significance of this correlation (p-value ≈ 0.01). Comparable results have been obtained from all the collection of networks that exhibit a substantially low objective score (as defined two paragraphs above) with the correlation to be either statistically significant (p-value < 0.05) or marginally statistically significant ($0.05 \leq p$-value < 0.1). No statistically significant correlation has been observed for networks that do not exhibit substantially low objective score and thus do not exhibit the "stripy lattice" property. The results of the cycles studies provide

some indications that the number of cycles although found to be of no significance as a global network measure (as discussed in section 5.2), it plays a role as an individual gene measure in terms of the number of cycles a gene is a member of. The more cycles a gene is member of the higher the effect of the knock-out of this gene on the transsys program objective score and consequently on the level of spatial heterogeneity.

5.3.2 Regulatory interaction properties

The next part of the single element deletion experiments has been designed to assess the effects of single regulatory interaction (edge) deletion on a transsys program objective score and study potential relationships between the objective score difference and network edge measures. The edge network measures that have been taken into account are the edge betweenness and the number of cycles an edge participates in. A full set of single edge deletion experiments as introduced in section 3.6.1 has been carried out for the whole population of transsys programs used in the previous studies of this chapter as described in section 2.2.3.3. Then correlations of the difference in the objective score of the single edge mutants from the original transsys program against network related measures of the deleted edge have been examined.

Both the network measures discussed here –the edge betweenness and the number of cycles an edge participates in (as defined in section 2.2.3.3)– have a statistically significant negative correlation with the objective score difference between an edge reduced and the original transsys program. This finding holds for the majority of the 77 transsys programs that showed a substantially low objective score (as defined in section 5.3.1). Figure 5.15 depicts the scatter-plots of the objective score differences against the edge betweenness and the participation in cycles from a characteristic transsys program among the 77 selected. The findings indicate an individual edge centrality (edge betweenness) measure to be statistically significantly correlated to the objective score loss in a similar fashion with the individual gene measures studied in the previous section (Spearman ρ rank correlation coefficient -0.33, p value ≈ 0.024). Edge betweenness plays a significant role in the information flow of gene and protein interaction networks, this role has been assessed in recent studies (Missiuro, Liu, Zou, Ross, Zhao, Liu, and Ge, 2009) by using an information flow measure. Here a measure of a dynamical property, that is the potential of generating gene expression heterogeneity in spatial organised

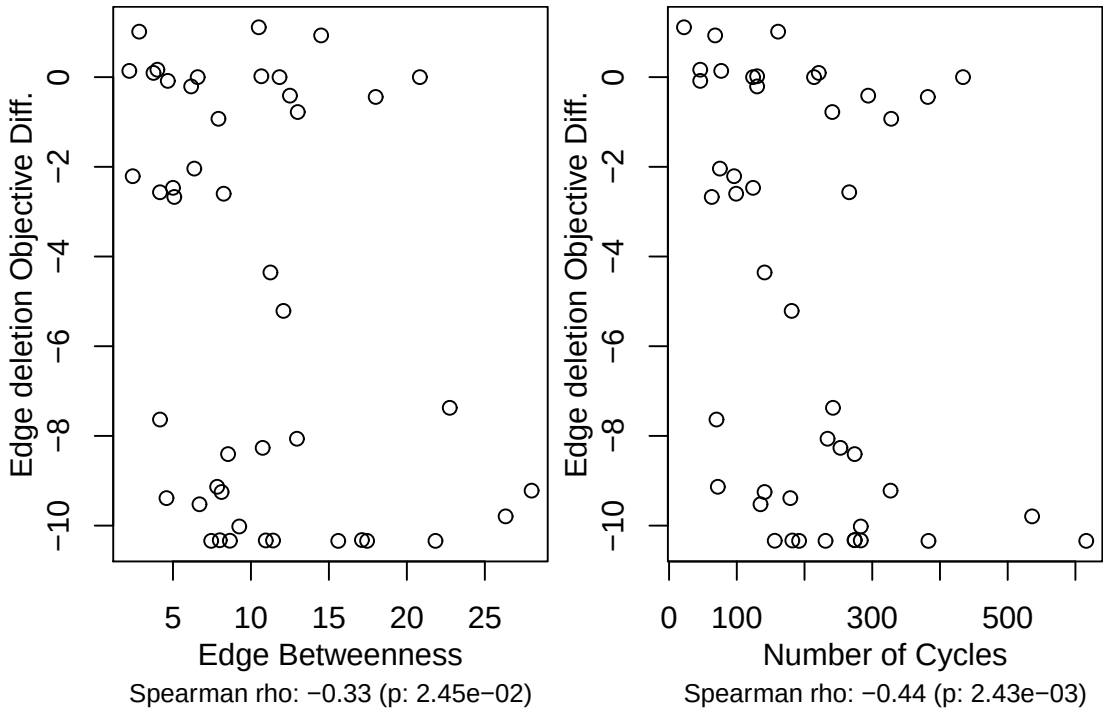


FIGURE 5.15: Correlation plots of edge related network topological properties vs. the objective score difference from the wild-type trannsys program owing to edge deletion. The plots are drawn from a transsys program that has objective score below the operational threshold depicted in figure 5.7.

systems, has been associated to the betweenness measure of individual gene regulatory interactions. Statistically significant correlation between the number of cycles an edge participates in and the objective score loss (figure 5.15) further corroborates the finding of the previous section that cycle measures of individual network elements can be used to check for network dynamical properties.

For every regulatory interaction the dynamical parameters were also available, each interaction holds two dynamical parameters the $a_{\max}$ which is equivalent to the maximum regulatory strength the interaction can exert and the α_{spec} which designates the binding specificity of a transcription factor with its DNA binding site. For each edge deletion the relationship with each interaction's $a_{\max}$ and α_{spec} was studied by correlation scatter-plots of these dynamical parameters against the objective score loss (figure 5.16 illustrates this for one of the 77 selected networks with substantially low objective score).

No significant correlation has been observed for the $a_{\max}$, however there has been a small number of transsys programs that have exhibit a marginally statistically significant correlation (data not shown). Instead, for the majority of the 77 transsys program with a substantially low objective score (as defined in section 5.3.1) the

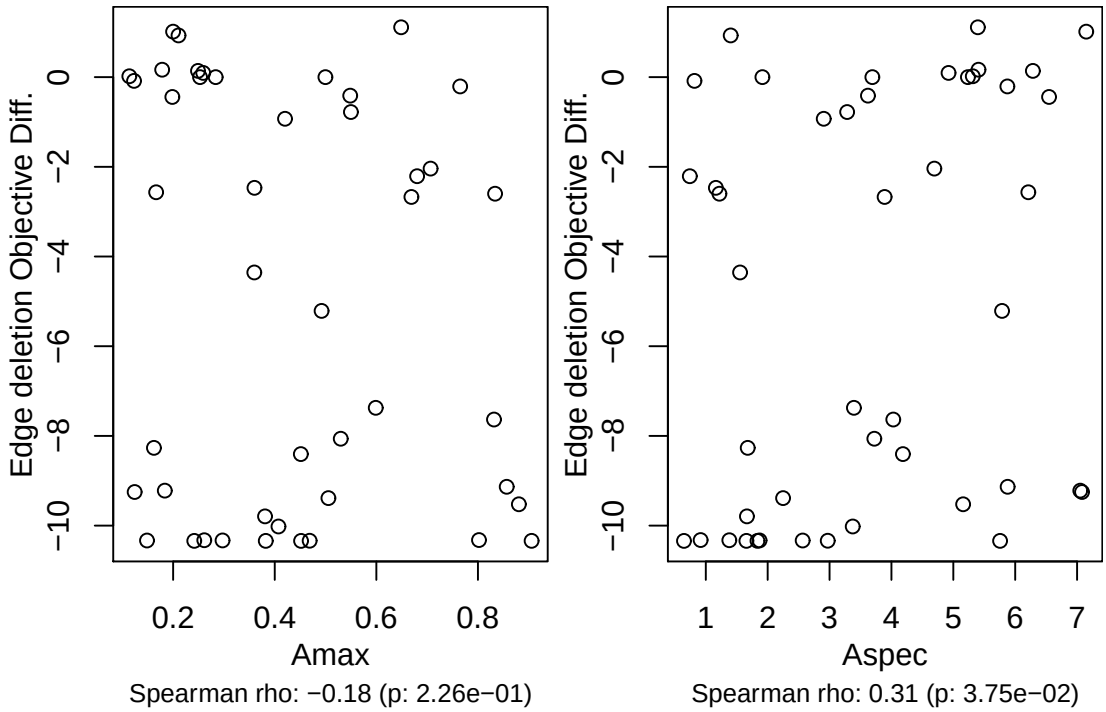


FIGURE 5.16: Correlation plots of single edge dynamical properties ($a_{\max}$ and α_{spec}) vs. the objective score difference from the wild-type transsys program owing to edge deletion. The plots are drawn from a transsys program that has objective score below the operational threshold depicted in figure 5.7.

α_{spec} dynamical parameter is positively correlated with the objective score difference between the edge reduced transsys program from the original. This finding suggest that the optimiser is exploiting the α_{spec} as a means to compensate for unwanted interactions. The highest the α_{spec} value for a regulatory interaction the lowest its regulatory strength and consequently the optimiser has the option to increase the α_{spec} in an attempt to eliminate interactions that have a negative effect on lower objective scores.

It is worthwhile to report that no trace of significant correlation has been found between the devised measure to study the size-3 motifs, the size-3 motif profile differences, and the objective score differences for both the genes as well as the regulatory interactions deletion experiments. Therefore, the studies in this thesis could not associate local network characteristics such as the size-3 motifs with the "stripy lattice" phenomenon in GRNs.

All the results presented in this chapter sections (i.e. 5.2 and 5.3) are newly generated and extended data from the experiments presented in (Bouyioukos and Kim, 2009), however comparable results have been obtained using networks with higher number of genes (25) and edges (75). Results from this analysis further statistically corroborate the analysis of the last two sectors by providing larger

samples for genes (25 instead of 15) and regulatory interactions (75 instead of 45). Indicative correlation scatter-plots from a selected transsys program with larger number of genes and regulatory interactions can be found in the appendix C.

Chapter 6

GRNs and Initial Conditions

Initial conditions of biological systems, in terms of factor concentration variability, affect gene expression in a number of ways. GRNs organise gene expression dynamics in a twofold way. On one hand there are GRNs (mostly the ones that are involved in signal transduction pathways) that have a significant degree of sensitivity to initial conditions and are transforming initial conditions variability to signals that elicit a response (in terms of differential gene expression) to variable initial conditions. On the other hand there exist GRNs (predominantly the ones involved in developmental processes) that are required to organised a robust response, that is a stable gene expression pattern, to initial conditions variability. This chapter investigates the latter property of GRNs by conducting an experiment which calculates transsys programs objective scores obtained by different initial reactor states, as a proxy to different initial conditions.

Every objective score calculation of a transsys program, as a result of the objective function evaluation, is determined by the transsys program and the simulator control parameters only (section 3.2.1). Here all the simulator control parameters are kept constant apart from the initial state, thus the objective score is determined only by the initial reactor state and the transsys program. The objective function equation (eq. 3.7) can be expressed as a mapping from the (initial reactor states $\times$ transsys programs) domain to $\mathbb{R}$.

In this chapter the contribution of these two domains, the initial reactor state domain and the transsys program domain will be examined. The analysis of the experiments results aims to explain the influence of a set of different initial reactor states and transsys programs on the variability of the objective score.

6.1 Experimental Setting

The transsys program population that was used for the initial conditions experiment was the one that was returned from the reference experimental procedure and equals to 2 network generation mechanisms × 15 topologies × 30 parametrisations = 900 different transsys programs. Each transsys program after exiting the optimisation procedure has its objective score evaluated starting from 100 different initial factor concentration reactor states, thus the whole data set analysed in this chapter consists of 900 transsys programs × 100 evaluations from different initial reactor states = 90000 objective function evaluations. The data set was grouped according to 4 different groupings: 100 initial states groups each containing 900 objective scores, 2 network generation groups with 45000 objective scores each, 30 topology groups each one with 3000 objective scores and 900 transsys program groups containing 100 objective scores each. Each of these grouping is considered as a contributing component of the objective score variability and each component was treated as containing categorical data. A statistical approach based in exploratory data analysis and standard analysis of variance techniques has been followed to determined which component (or grouping) is able to capture more of the observed variability of the objective score and thus has the greater impact in determining this score.

6.2 Initial Reactor States

The core motivation in the design of this chapter's experiment was to assess and quantify the effects of either the initial rector state or the transsys program on the objective scores. To study the effect of the initial reactor state the transsys objective scores that have been calculated starting from an identical initial reactor state have been grouped together, thus 100 different initial reactor states groups have been formed each one containing 2 network generation mechanisms × 15 topologies × 30 parametrisations = 900 objective scores. Exploratory analysis of these effects has been applied including boxplots generation and calculation of descriptive statistics of the mean and variance of initial state group means. A standard analysis of variance approach followed to quantify the impact of initial states on the variability of the objective scores.

Exploratory analysis of the initial reactor state grouping data has been carried out by generating the boxplots for all the 100 groups and plot them together. Figure 6.1 illustrates all the boxplots of each initial reactor state group.

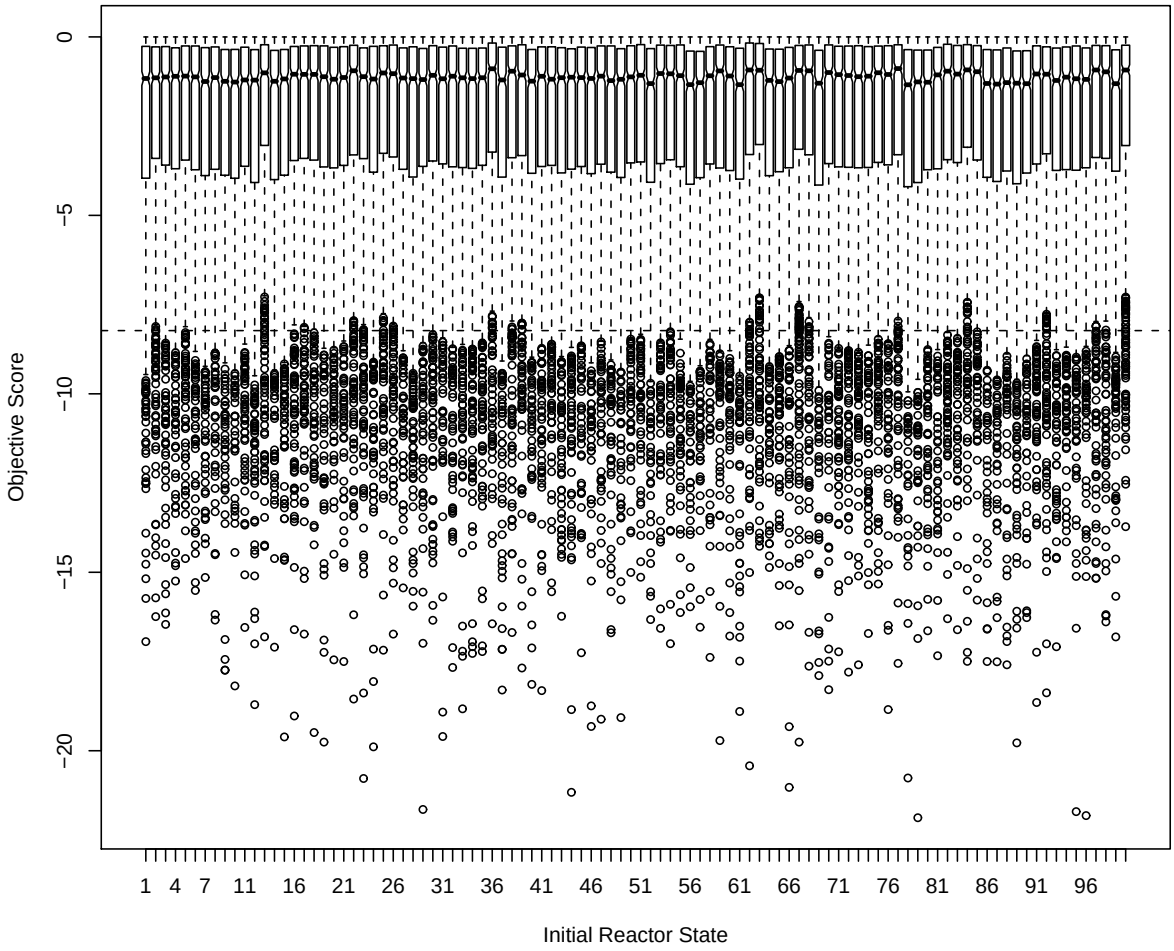


FIGURE 6.1: Notched boxplots of transsys program objective scores. Labels on the x-axis designate the different initial reactor factor concentration states. Each boxplot contains all the objective scores from 900 different transsys program grouped by the same intiial reactor state. The doted line represent the operational threshold for the "stripy lattice" phenomenon, as introduced in section 4.1.3.

Boxplots in figure 6.1 indicate that groups of objective scores based to initial reactor states share relatively little differences in terms of the median objective score, suggesting that objective scores are not possibly affected by the initial state groups. Also the boxes (representing the inter-quartile range) appeared to have comparable sizes among all the initial states further suggesting a relatively limited impact of the initial reactor state on the objective scores.

Further exploratory analysis of the data supports the boxplots results discussed above. The mean and the variance of the means of each boxplot of objective score evaluations grouped by different initial states have been calculated. The mean of the initial state groups means $\mu_{\text{initStates}}$ is -2.419 and the variance of the means $\sigma^2_{\text{initStates}} = 0.016$. Compared to the overall variance of the objective scores ($\sigma^2_{\text{total}} = 8.824$) the variance of the initial reactor states means is two orders of magnitude lower and also two orders of magnitude lower than the mean of means, suggesting that grouping according to the initial reactor states is not able to capture a substantial amount of the objective score variability.

Consequent analysis of variance of the initial reactor state groupings has been conducted. The initial state grouping has been treated as a categorical variable of 100 levels and the results from the analysis of variance of a linear model of the grouping are as follows:

```
aov(formula = lmInitState)
Terms:
                factorInitState Residuals
Sum of Squares           1467.7  792718.8
Deg. of Freedom              99     89900
Residual standard error: 2.969475
```

From the analysis of variance results the within each initial state group variability (or the Mean Square Error (MSE)) can be calculated. The MSE is a measure of the variability of the original data that the model under consideration has captured. For the initial reactor state grouping is MSE = 792718.8/89900 $\approx$ 8.817. In the context of calculating MSE from the rest of the groupings (aggregate results of the statistics calculated in this chapter can be found in table 6.1), the MSE of initial reactor state grouping has the highest value among all the rest of MSEs from other groupings (network generation, topology, transsys program).

The ANOVA table provides indications of the significance of the grouping as a potential predictor of the objective score and is as follows:

```
Analysis of Variance Table
Response: objScore
                  Df Sum Sq Mean Sq F value    Pr(>F)
factorInitState    99   1468      15  1.6813 2.634e-05 ***
Residuals       89900 792719       9
```

The p-value of the initial state model (the "Pr(¿F)" column) indicates statistical significance, however the relatively large number of the degrees of freedom (Df = 99) diminishes the significance of this finding.

Together the last two findings indicate that the initial state although that determines the objective score evaluations (according to the ANOVA table), does not effectively capture the of variability of the objective scores (according to the boxplots figure 6.1 and the calculation of the MSE) and therefore can be chosen arbitrarily, that is that any choice of the initial reactor state can be arbitrary, will have equivalent effects in the objective scores and will generate statistically comparable results.

6.3 Effects of Network Generation Mechanism

The effect of each of the two network generation mechanisms has been studied by grouping the objective scores obtained from the ER process and the PL process of generating random networks in two separate groups. Exploratory analysis of the two groups by generating the respective boxplots has revealed possible difference between the means of the objective score distributions of the ER and PL networks, as illustrated in figure 6.2.

The boxplots indicate a lower objective score median derived from PL networks than the ER as the notches between the two boxes do not overlap. The latter was not the case for the experiment reported in section 5.2. The objective scores boxplots in figure 5.6 did not imply any significant difference in the ER and PL medians. Further corroboration of the boxplot finding comes by performing a Wilcoxon rank sum test. The Wilcoxon test compares the location parameters of two distributions and applying it on the ER and PL grouping has revealed a statistical significant difference: $W = 1083968746$, p-value $< 2.2e-16$. The low p-value of the Wilcoxon test indicates that the location shift in the distribution is significantly different. Furthermore, from the boxplots figure the median of the objective scores from PL networks is lower and the median of the ER networks, however ER networks have more transsys programs that encoded for lower objective scores. Overall, the results favours the hypothesis that PL networks are more robust to initial reactor states variation than ER networks.

In addition to the comparisons, the role of network generation mechanism as a contributing factor to the objective score variability has been studied. The analysis

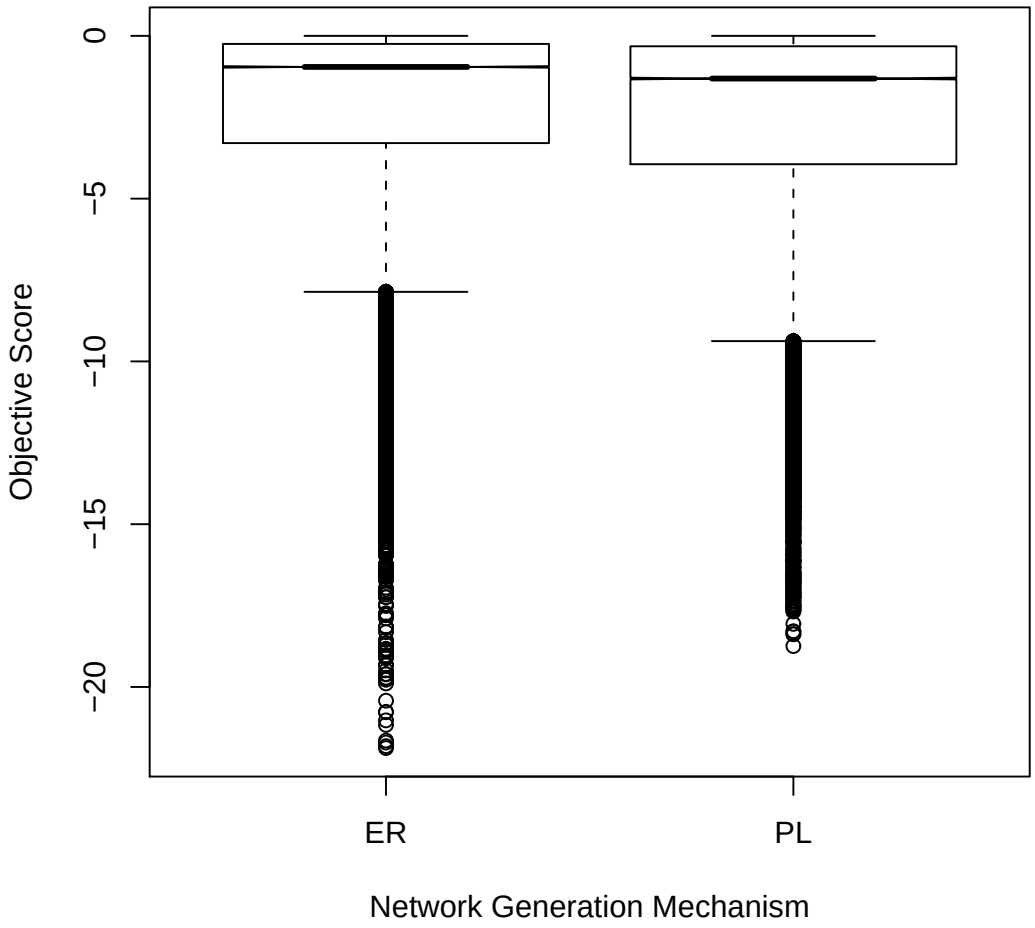


FIGURE 6.2: Notched boxplots of transsys program objective scores. The network generation mechanism (either ER or PL as introduced in section 3.4) was used to populate the two boxplots. A significant difference in the median is observed in the figure and it is also supported by a Wilcoxon test in the text. (Each boxplot contains 45000 scores and thus the boxplot notches are so narrow that rendered almost invisible in the figure).

of variance output of the ER – PL grouping has been conducted, from where the mean square error can be calculated. The mean square error (or within groups variability) is the ratio of residuals sum of squares over the degrees of freedom thus MSE = 791806.7/89998 = 8.798, and it is the second larger MSE of the study according to the aggregate table 6.1.

```
aov(formula = lmNetgen)
Terms:
                 factorNetGen Residuals
```

```
Sum of Squares        2379.8  791806.7
Deg. of Freedom            1     89998
Residual standard error: 2.966150
```

Grouping the transsys program objective scores evaluation according to the network generation mechanism revealed possible differences of the robustness of the PL networks with regard of the initial reactor state. The mean squared error (or the within group variation) when the two groups network generator groups are considered was significantly lower than the mean of squares (or the between group variation). The results of the ANOVA table suggested the network generation mechanism as a potential, yet weak, predictor for the objective score evaluation from multiple initial reactor states.

```
Analysis of Variance Table
Response: objScore
                Df Sum Sq Mean Sq F value    Pr(>F)
factorNetGen     1   2380    2380  270.49 < 2.2e-16 ***
Residuals    89998 791807       9
```

6.4 Network Topologies

The next level of grouping of the transsys program objective scores was carried out according to their respective network topology. For each of the two network generators 15 different topologies have been produced, thus resulting to 30 different groups each containing 300 objective scores from each distinct topology. Exploratory analysis of the data by the boxplots of figures 6.3 & 6.4 provides an initial picture of relatively little variability of the objective scores with regard to the network topologies. This was anticipated as different topologies have derived from random sampling of the space of all the potential topologies and they have not been subject to any alteration through the optimisation process.

Continuing the exploratory mode of the analysis, the mean and the variance of the means of objective scores from each topology group have been calculated. The mean of means of the objective score from the 30 topology groups $\mu_{\text{topologies}}$ is -2.419 and the variance $\sigma^2_{\text{topologies}} = 1.281$. Both the mean and the variance of means are in the same order of magnitude and the variance of the means is substantially lower than the total variance ($\sigma^2_{\text{total}} = 8.824$) indicating that an amount

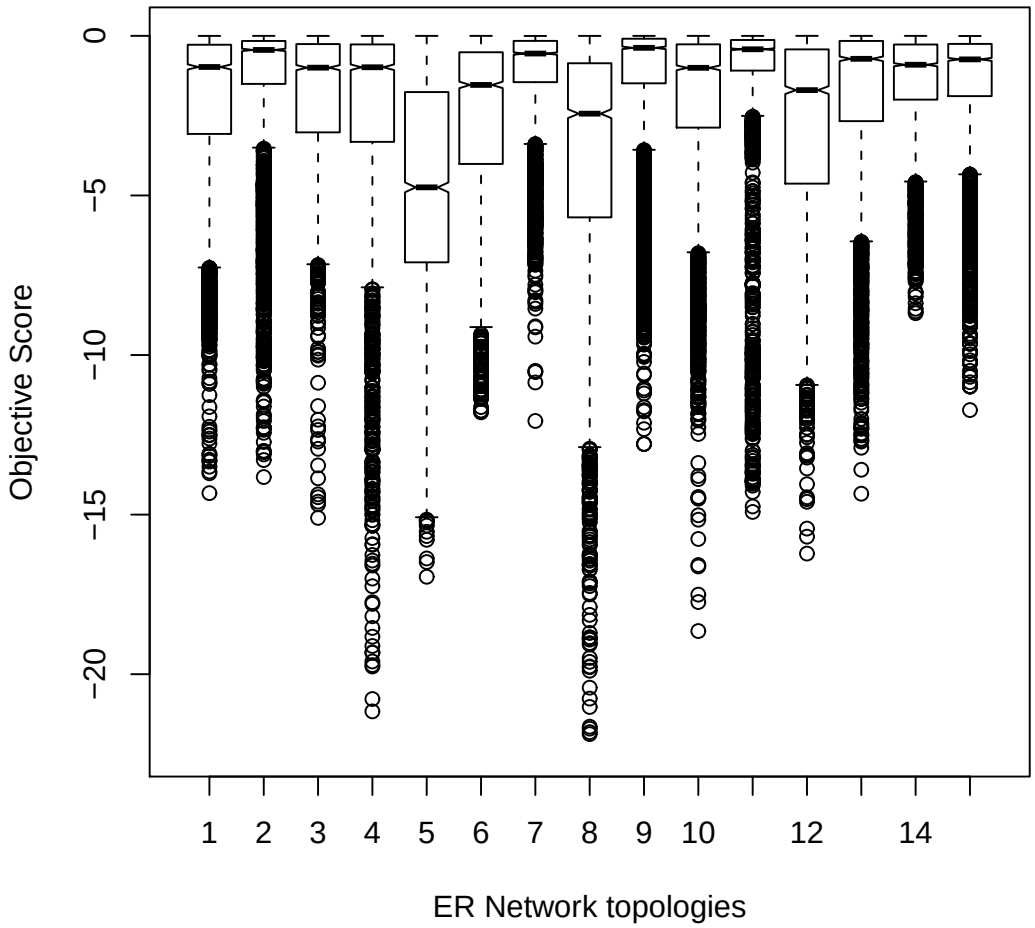


FIGURE 6.3: Notched boxplots of transsys programs objective scores generated according to the Erdös-Rényi model. The objective scores are grouped according to the 15 different different network topologies that have been generated for the ER networks.

of variability of the objective scores was captured when scores were grouped according to different network topologies.

The impact of the network topology in the objective score is a central objective of this thesis (as stated in section 1.6). To further investigate potential relationships of the network topologies with the objective score evaluations, a standard analysis of variance approach was used. The grouping of 30 topologies was treated as a categorical variable of 30 levels and the analysis of variances of the linear model results were as follows:

```
aov(formula = lmTopology)
```

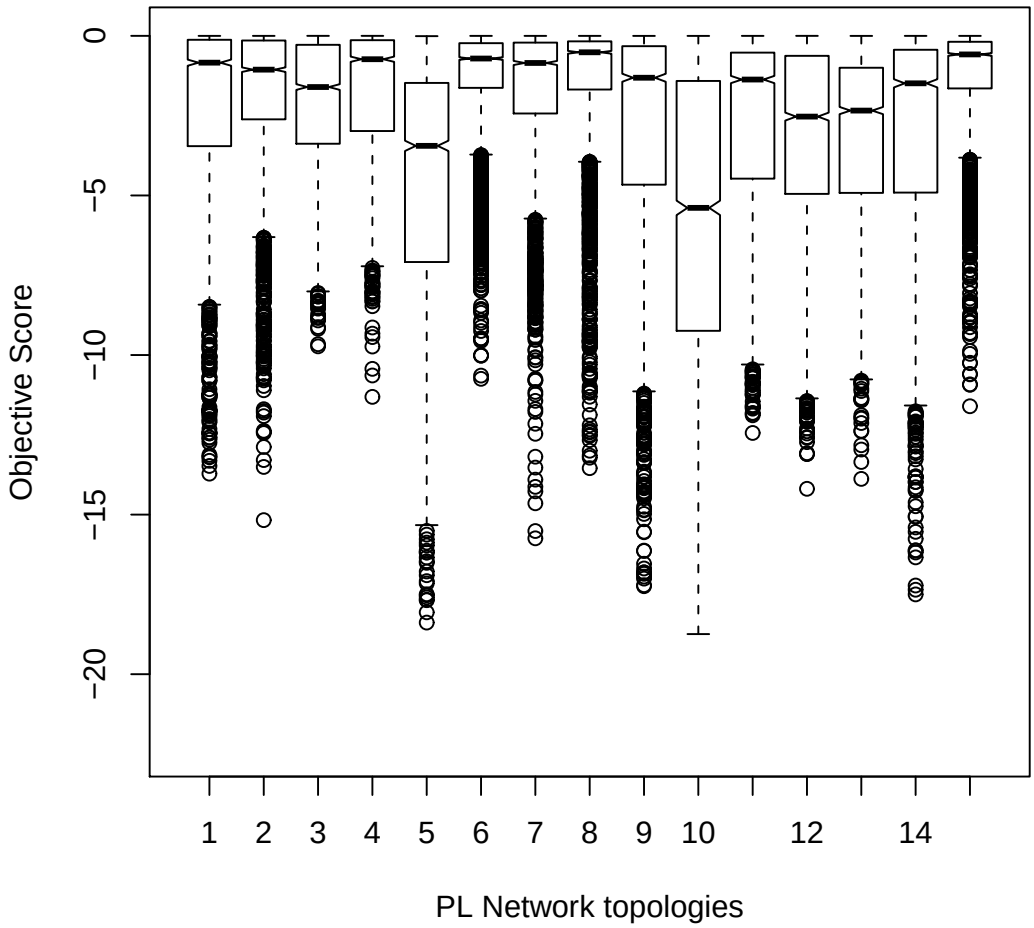


FIGURE 6.4: Notched boxplots of transsys programs objective scores generated by a process that results to networks with power-law degree distribution. The objective scores are grouped according to the 15 different network topologies that have been generated for the PL networks.

```
Terms:
                  factorNetwork Residuals
Sum of Squares         111527.3  682659.2
Deg. of Freedom              29     89970
```

From the analysis of variance results the mean square error (MSE) was calculated. MSE = 682659.2/89970 = 7.588 the mean square error (or the within group variation) is less than the MSE calculated for the network generation grouping and indicates that grouping by topology can capture more of the variability of the objective score that the network generation mechanism.

Accordingly, to asses the significance of the grouping by topologies the ANOVA table has been calculated:

```
Analysis of Variance Table
Response: objScore
                 Df Sum Sq Mean Sq F value    Pr(>F)
factorNetwork    29 111527    3846  506.85 < 2.2e-16 ***
Residuals     89970 682659       8
```

Topology significantly determines the objective scores, however the relatively large amount of the degrees of freedom (29) abates the highly statistical significance of the association between topology and objective score that the low p-value indicates.

6.5 Transsys Programs

The last level of grouping is based on different transsys programs. As described in section 6.1 for each network topology 30 transsys programs with different initial dynamical parameter setting have been generated. Therefore the grouping that is used to analyse the data is the one based on individual transsys programs, this grouping consists of 2 network generation mechanisms $\times$ 15 topologies $\times$ 30 initial parametrisations = 900 different transsys program groups each containing 100 objective scores obtained from the respective different initial reactor states.

The boxplots of transsys program objective scores for each individual transsys program will facilitate the exploratory analysis of the data. The transsys program boxplots derived from the same network topology are plotted together in the same figure. Form the exploratory analysis of the data the variability that is due to the different initial parametrisations of each transsys program can be studied, the full complement of the 900 transsys program groups boxplots divided into 30 figures can be found in the appendix D.

Here, transsys program boxplots from two indicative topologies one representing transsys programs with the highest objective score variability and one with the lowest are presented. The variability has been quantified using the coefficient of variability (that is the ratio of the standard deviation over the mean). The coefficient of variability has been calculated for each different topology and all the topologies then have sorted in descending order. The topology with the highest

coefficient of variability is illustrated in figure 6.5 (it is the 5^{th} from the ER generation mechanism) and the one with the lowest in figure 6.6 (the 11^{th} from the ER mechanism).

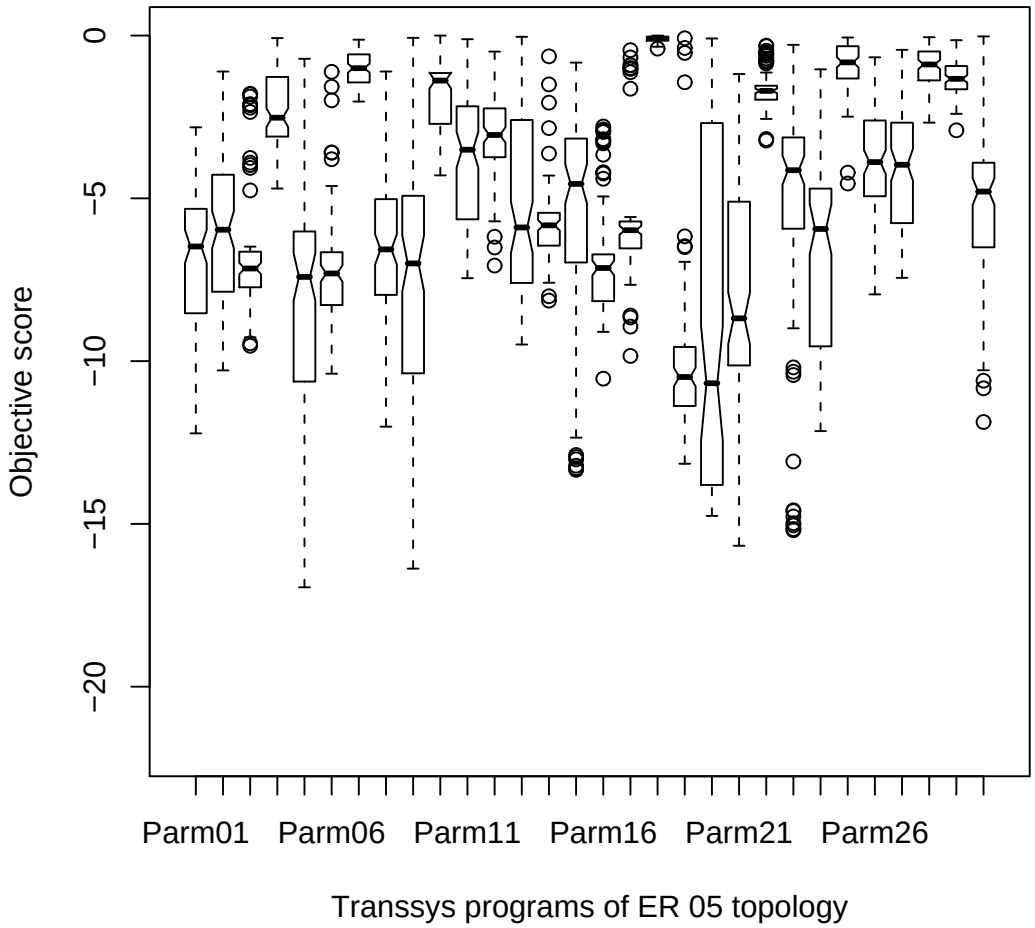


FIGURE 6.5: Notched boxplots of transsys programs objective scores grouped by the same initial dynamical parameters. This is a selected topology which exhibits relatively large variation for each set of dynamical parameters, most of the parametrisations have resulted to a median objective score lower than -6 and only a couple have median objective score around zero.

Figure 6.5 depicts boxplots of objective scores from the transsys programs of the network topology with the higher coefficient of variability. The degree of variation of the objective scores is the highest among all the other network topologies, and a couple of transsys programs are reaching median objective score lower than -10. Boxplot analysis offers an initial demonstration that the transsys program accounts for a significant amount of objective score variability.

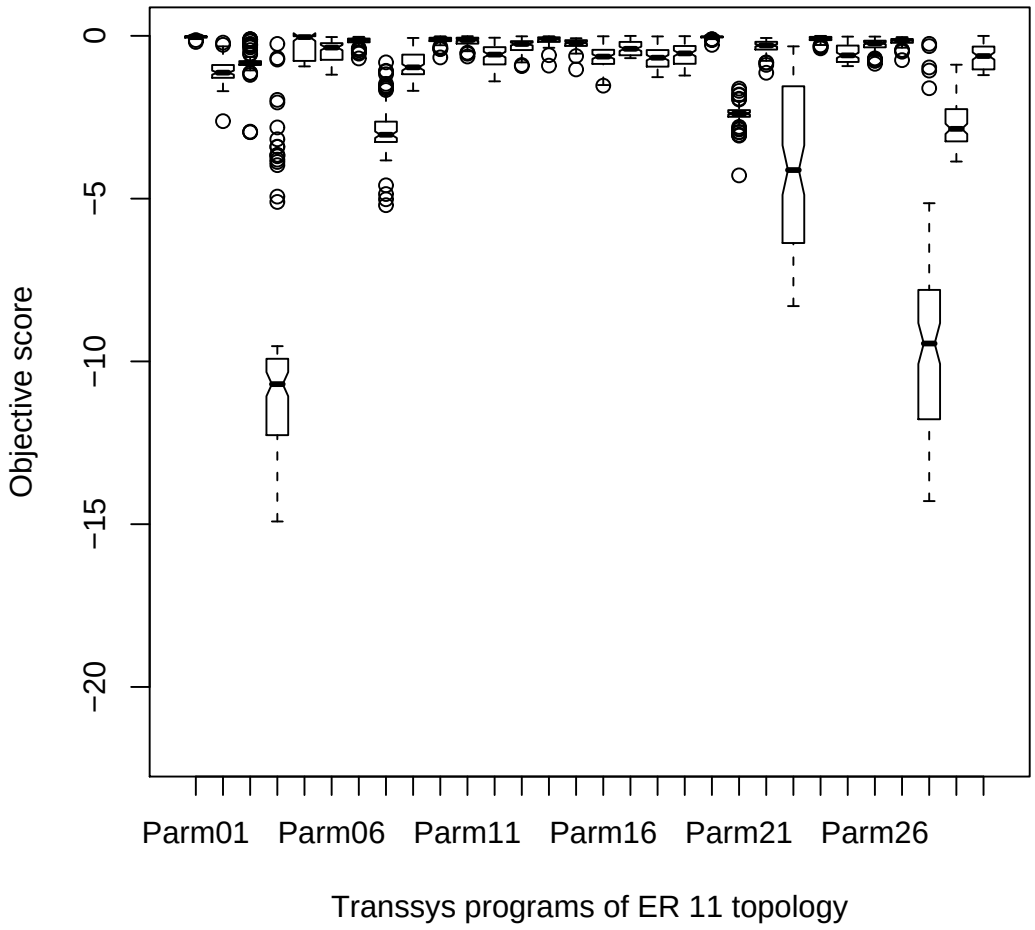


FIGURE 6.6: Notched boxplots of objective scores evaluations from different initial conditions grouped by the same initial dynamical parameters. This is a selected topology which exhibits relatively small variation for each set of dynamical parameters, a few parametrisations have generated a low median objective score whereas most are mainly concentrated around zero.

Figure 6.6 illustrates boxplots of objective scores of transsys programs derived from the topology with the lowest coefficient of variability. Most of the transsys programs have objective scores medians concentrated close and relatively little lower than zero. However still a couple of transsys programs have generated relatively low objective score median, which provides an additional indication that the transsys program grouping can capture a substantial amount of the objective score variability.

Moving on with the exploratory analysis the mean and the variance of the means from each boxplot of objective scores that was grouped according to different

transsys program were calculated. The mean of the means of the transsys programs groups μ_{tp} was -2.419 and the variance of the medians $\sigma^2_{\text{tp}} = 6.604$. Both these descriptive statistics are in the same order of magnitude however the variance is almost 3 times higher than the mean value indicating a relatively large degree of variability at the individual transsys program groups, and compared to the total objective score variance ($\sigma^2_{\text{total}} = 8.824$) is relatively close. Grouping the objective scores on the transsys program level captures the highest (compared to the topology, section 6.4, and the initial reactor state, section 6.2, groupings) variability of the objective score evaluations.

Boxplots of transsys programs from two indicative topologies have been presented here. The full complement of all the data-set comprise 30 images and 900 boxplots one for each parametrisation can be found in the appendix D, the collection of all the Erdös-Rényi networks are in appendix D.1.1 and for networks with power law degree distribution in the appendix D.1.2.

Comparison of the figure 6.5 from a topology that show substantial amount of variability in objective score evaluations from different transsys programs with the boxplots of objective scores evaluations from a topology that do not show substantial variability in figure 6.6 indicates that the network topology is a significant factor that determines the objective score. An observation that comes in line with the results of the network grouping section 6.4.

The analysis of variances between and within groups also reveals an equivalent to the exploratory analysis result.

```
   aov(formula = lmTranssys)
Terms:
                factorTranssysProgram Residuals
Sum of Squares                593786.5  200400.0
Deg. of Freedom                    899     89100
Residual standard error: 1.499719
```

The mean square error MSE = 200400/89100 $\approx$ 2.249 is the lowest than the rest of MSEs calculated for the rest of the grouping (consult th aggregate table 6.1 for comparisons), suggesting that the transsys program grouping is able to capture the biggest portion of the variability, compared to the rest of the groupings, of the objective scores.

Furthermore the ANOVA table indicates that the transsys programs groups are a significant predictor of the objective score evaluations variability:

```
Analysis of Variance Table
Response: objScore
                          Df Sum Sq Mean Sq F value    Pr(>F)
factorTranssysProgram    899 593787     660  293.66 < 2.2e-16 ***
Residuals              89100 200400       2
```

The grouping is a statistical significant predictor of the objective score, however the very large number of degrees of freedom (899) diminishes the statistically significant role of transsys program as a predictor that the low p-value suggests.

6.6 Discussion

To summarise the findings of the multiple initial reactor state experiment and to compare the results of each individual grouping with the rest a table collecting the statistics calculated from the exploratory and the analysis of variances from each grouping is presented:

Grouping	Mean of means	Variance of means	MSE	D.f.	p-values
Initial reactor states	-2.419	0.016	8.818	99	2.634×10^{-05}
Network mechanisms	-2.419	0.052	8.798	1	$< 2.2 \times 10^{-16}$
Network topologies	-2.419	1.281	7.587	29	$< 2.2 \times 10^{-16}$
Transsys programs	-2.419	6.604	2.249	899	$< 2.2 \times 10^{-16}$
Whole data	-2.419	8.824 (total variance)	N/A	N/A	N/A

TABLE 6.1: Summarising table of all the statistics from the exploratory and the analysis of variances of different groupings

The table summarises what has been shown in the individual sections of this chapter that the transsys program grouping has been found as the component that captures the highest variability of the objective scores. This is explicable by the fact that the transsys program is subject to optimisation, every transsys program entered this analysis has its dynamical parameters optimised to minimise the objective score. Therefore the transsys programs, consisting of the topology and the dynamical parameters, that used in this experiment are not a unbiased,

random sampling of all the population of possible transsys programs. In addition the hierarchical structure of the transsys program generation, that each transsys program belongs together with other in a network topology group and each network topology belongs with others to a network generation mechanism group is determining that higher levels of grouping will be able to capture less objective score variability, for instance the network generation mechanism grouping, by construction, will capture less objective score variability than the network topology grouping. Finally, regarding the initial reactor factor concentration states and in a biological analogy, initial variability of factor concentrations along a tissue, although is a necessary condition for the generation of heterogeneity (as it was suggested by the ANOVA table of the initial reactor state grouping), it can be chosen arbitrarily and different initial states will have limited effects on the statistics of the "stripy lattice" properties that the developmental GRN will generate.

Chapter 7

Exploring Robustness

Robustness as discussed in the literature review chapter (section 2.2.4) is a property that permeates all the levels of biological organisation and in the context of GRNs studies encompasses more than one aspect. In chapter 6 the role of transsys programs and initial reactor states in determining the objective score variability has been examined and indications of GRN robustness against randomly generated initial reactor factor concentration states have been pointed out. The course of the current chapter is built upon the results and the findings of the previous chapter (chapter 6), and is focused on further investigations of the topological properties of transsys programs. Topological properties of a selection of transsys programs that is compiled for robustness of their objective scores from different initial reactor states were identified and network elements pruning experiments were carried out to further study the GRNs topological robustness to sequential network element deletion.

7.1 Experimental Setting and Analysis

The robustness of particular transsys programs objective scores to different initial reactor states has been motivated by the observed variability of the transsys program groupings (results in section 6.5, full set of results in the appendix D and summary table 6.1), results that indicate that the differences in transsys programs is the predominant factor to describe the variability of the objective score. In this chapter the experiment is based in the evaluation of how robust is the low objective score generated by different transsys programs gainst network perturbations. The transsys program objective score robustness has been quantified based on two

criteria: relatively low objective score and relatively small variation of the series of different objective scores obtained from different initial reactor states for each transsys program. A selection process has been devised in order to distinguish transsys programs that exhibit a robust behaviour of the objective scores from different initial reactor states. The selection criteria were based in two descriptive statistic measures the median of the objective scores and the Median Absolute Deviation (MAD). The MAD is a measure of dispersion that is more robust compared to the inter-quartile range (IQR) and it is defined as the median of the absolute deviations from the data's median.

The process for selecting the networks that have a robust –i.e. relatively small MAD and relatively low median of objective score evaluations– introduces two arbitrarily selected thresholds for the two selection criteria. The median should be less than or equal to -6, so only transsys programs with objective scores low enough to exhibit spatial heterogeneity will be selected; and the MAD should be less than or equal to 2, so that transsys programs that have a low dispersal of their objective scores due to differences in initial conditions will be selected. Figure 7.1 graphically represents the two thresholds and the selected transsys programs.

These two arbitrarily chosen thresholds were used to set up a proxy measure of the robustness of the objective score. Transsys programs that fulfil the two selection criteria have been chosen from the total population of transsys programs. 67 transsys programs (the ones outside the dashed line area of figure 7.1) have been obtained by this selection procedure out of the 900 of the total population of transsys programs generated by the reference control parameters set. To satisfy the objective to characterise the network topology of these transsys programs, a set of network topological properties have been extracted from the selected transsys programs and their statistics were compared with topological properties of the total transsys programs population.

7.1.1 Pruning Networks Analyses

The topological robustness of the transsys programs was examined by applying the cumulative network elements approaches described in section 3.6.2. The approach consists of two experiments one aims to reduce the network by knocking-out genes the second by deleting regulatory interactions, both in a cumulative fashion. There is a threshold when these cumulative reduction operations stop and it is set at the 50% of the wild type transsys program objective score. The 50% threshold has

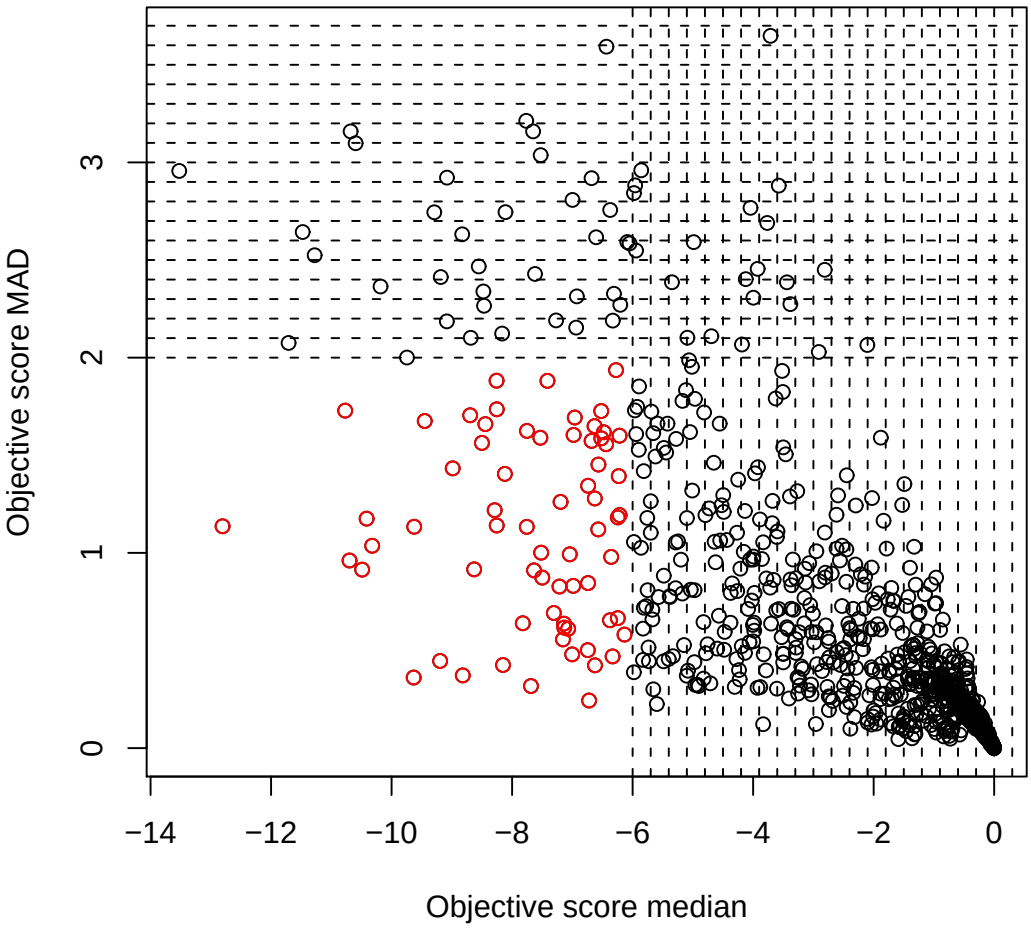


FIGURE 7.1: Plot of the transsys program objective scores medians against the median absolute deviation each point represents an individual transsys program. Medians and MADs were calculated from transsys programs objective scores from random initial reactor states (as generated in chapter 6). The area outside the dashed lines was the selected one. The positions of the particular transsys programs are indicated in red.

been chosen as it is the mean of the objective score and most of the reduction operations have either a minute effect or a very severe effect that brings the the objective score close to zero, thus a 50% threshold constitutes a safe choice to asses the topological robustness. Comparing pruned network elements from the 67 selected transsys programs with the rest of the transsys programs population might reveal potential relationship between two aspects of GRN robustness, the robustness to initial reactor state and robustness to network pruning.

7.2 Topological Properties of Robust GRNs

Out of all the global network topological measures that were introduced in section 5.2 to study topological properties, the clustering coefficient and the diameter were the network measures significantly differentiated (based on the boxplots notch overlaping indication) between the selected transsys program and the total population.

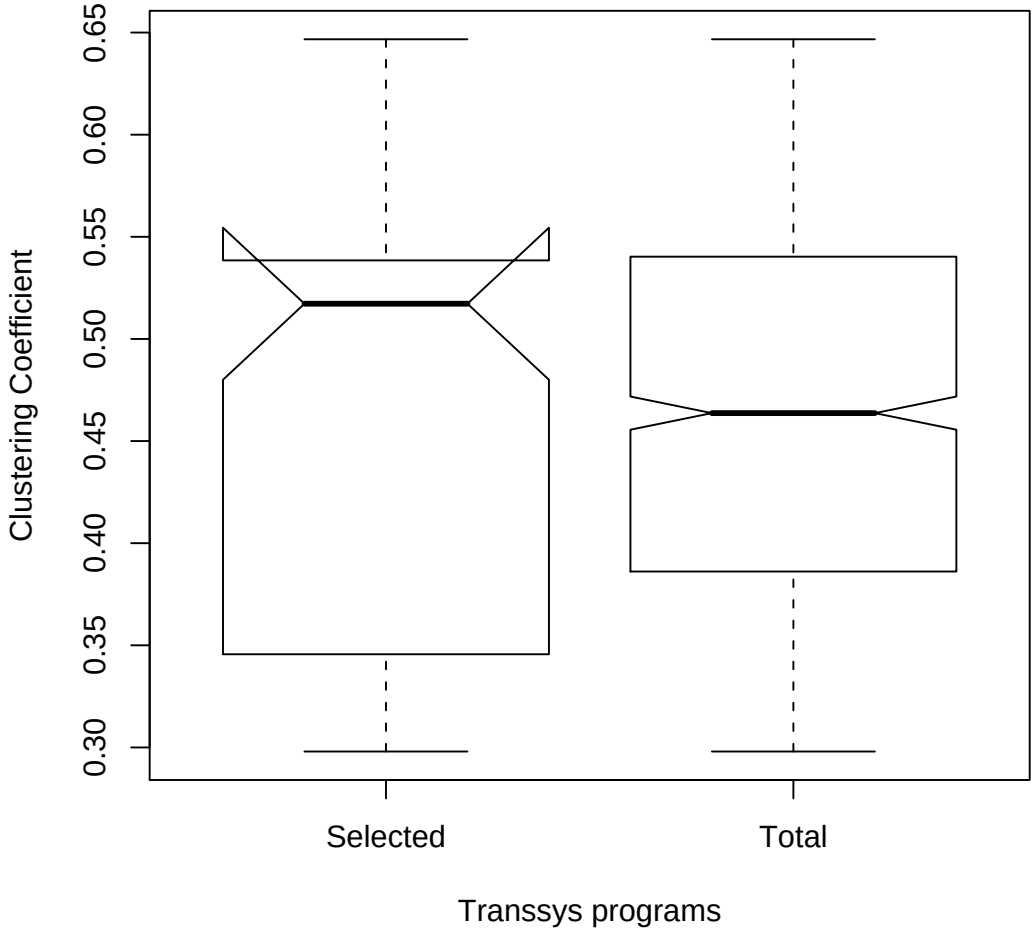


FIGURE 7.2: Notched boxplots of the clustering coefficients of the robust selected transsys programs (red points in figure 7.1) and the whole transsys program population.

Figure 7.2 illustrates that the median of the clustering coefficients of the selected transsys programs is higher (with no notch overlapping) from the total population of transsys programs indicating that higher clustering coefficients might be

beneficial for transsys programs to exhibit objective score robustness to different initial reactor states. However further support of this argument by a Wilcoxon rank sum test show no statistical significance in the location shift between the selected transsys programs and the total population of transsys programs clustering coefficient distributions, $W = 30390$, p-value $= 0.9135$. The p-value clearly indicates that the null hypothesis that the location shift between the clustering coefficient distributions of the selected and the total transsys program populations is zero is accepted.

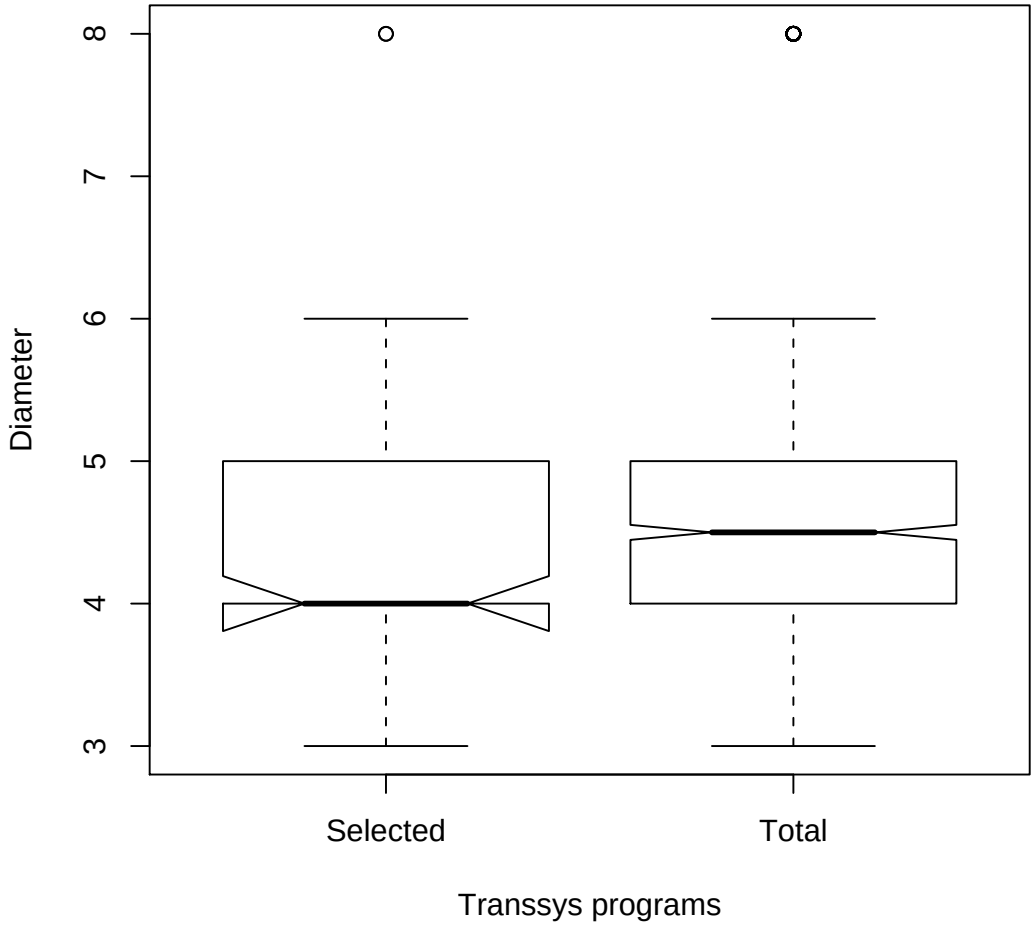


FIGURE 7.3: Notched boxplots of the network diameter of the robust selected transsys programs (red points in figure 7.1) and the whole transsys program population.

Boxplots of the diameter, in figure 7.3, of selected transsys program with objective scores that were robust to initial reactor state variability and the total transsys program population reveal also a potentially significant difference in the objective

score medians. The notches do not overlap suggesting a lower diameter of the selected compared to the total transsys program population. This observation is marginally statistically supported as the Wilcoxon rank sum test has returned $W = 26610$, p-value $= 0.09227$. The 0.09 p-value ($0.05 <= p$-value $<= 0.1$) indicates marginal statistical significance for a distribution location shift between the diameters of selected and the total population of transsys programs.

The clustering coefficient and diameter results come as a complementary argument to previous finding that networks with small diameter tend to have lower objective scores (results discussed in section 5.2). Both these findings support that "small world" networks is a potential topological characteristic of gene regulatory networks, as has been reported from previous studies (Almaas, 2007; Watts and Strogatz, 1998). The results of both the experiments in this section combined with the ones in section 5.2 are suggesting that the small diameter is a characteristic of networks with capacity to generate heterogeneity an thus these networks share the "small world" property that numerous biological networks have (examples in (Watts and Strogatz, 1998)).

7.3 Single Element Deletions and Robustness

The single element network properties and the objective score loss due to the single element deletion were studied next under the light of the objective score robustness selection. The Spearman ρ rank correlation coefficients between single element deletions and network element topological properties were retrieved from all the transsys programs generated with the reference parameters set. The date set used here is the same with the one used for the individual element property analysis presented in section 2.2.3.3. Using the transsys program selection procedure described in the experimental settings section 7.1, an aggregate statistical analysis has been carried out for the Spearman ρ rank correlation coefficients. The correlation coefficients of the degree of the knocked-out gene against the knock-out objective score loss were collected for the selected transsys programs and the total population and the two distributions are depicted as boxplots in figure 7.4.

As demonstrated in figure 7.4, for the selected for objective score robustness transsys programs gene degree is associated more with the objective score loss compared to the association of the total population. Equivalent behaviour of the rank correlation coefficient is observed for the number of cycles that a gene is a

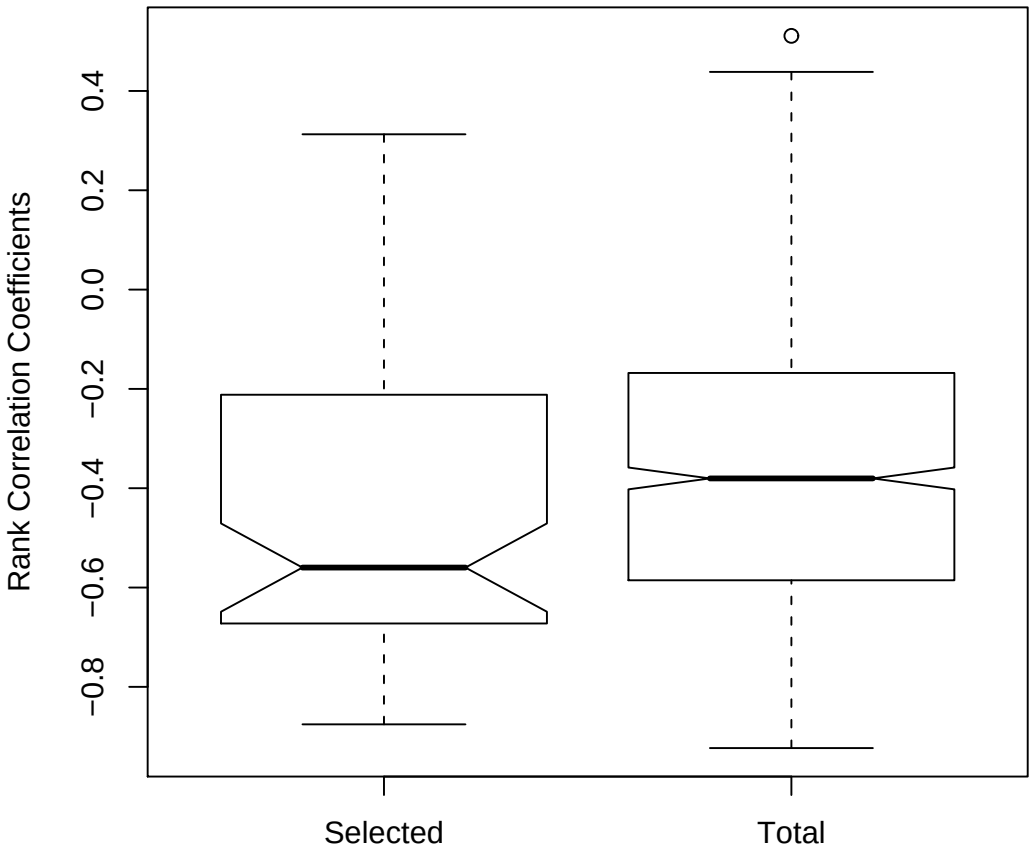


FIGURE 7.4: Notched boxplots of Spearman ρ rank correlation coefficient between the degree of a knocked-out gene and the objective score loss due to this knock-out. Data are from individual gene deletion experiments for each gene in the selected robust transsys programs and the total population.

member of, for the closeness and for the eigenvector centrality individual gene properties, with the respective boxplots are in the appendix E.

Consequently for the selected transsys programs four individual network properties: the degree, the number of cycles a gene is member of, the closeness and the eigenvector centrality, have been found to exhibit stronger correlation coefficients than the total transsys program population. These result is in line with the observation in section 2.2.3.3 that the majority of the transsys programs with a substantially low objective score have their individual network element properties correlated with the objective score difference.

7.4 Topological Robustness

GRNs robustness in terms of GRNs topology is a crucial property of network that organise phenomena of cell differentiation as tolerance to structural mutations is essential for the GRN to organise accurately phenomena of cell differentiation and pattern formation. In addition, examine the elimination of unwanted network elements is a methodology that can lead to the identification of minimal core GRN topologies that are able to exhibit the "stripy lattice" property. To study robustness of the GRNs topology, a sequential element deletion experiment, as described in the experimental setting section 3.6.2 was conducted and the results are reported and discussed here.

To demonstrate the application of the 50% objective score threshold in cumulative gene pruning and regulatory interaction pruning experiment respectively two figures where generated. Figure 7.5 illustrates the objective scores traces of a transsys program from the reference as the cumulative gene knock-out experiment proceeds. In each step a gene gets knocked- out and a reduced transsys program with less genes is generated.

In the example of figure 7.5 the reduced transsys program that the experiment will return is the one that is the outcome after the knock-out of all the genes below the 50% objective score threshold (indicated by the dotted line).

Figure 7.6 illustrates objective score traces of a transsys program from the reference set over cumulative regulatory interactions deletion steps. In each step a regulatory interaction is removed and a pruned transsys program with less edges is generated.

In the example of figure 7.6 the pruned transsys program that is returned is the one that is the outcome of the removal of all the regulatory interactions below the 50% objective core threshold (indicated by the dotted line).

The cumulative element deletion operation in combination with the set up of the 50% threshold eliminate a certain amount of network elements (genes or regulatory interactions). This number of pruned network elements is used as a proxy to the GRNs robustness to cumulative pruning, the more elements get pruned the more robust a GRN is. The hypothesis is that transsys programs that exhibit robust objective score in a series of random initial factor concentration states (selected in section 7.1) will be robust to cumulative element deletion as well. The motivation behind that is to investigate any potential connections between two different aspects of GRN robustness. Robustness of objective scores against

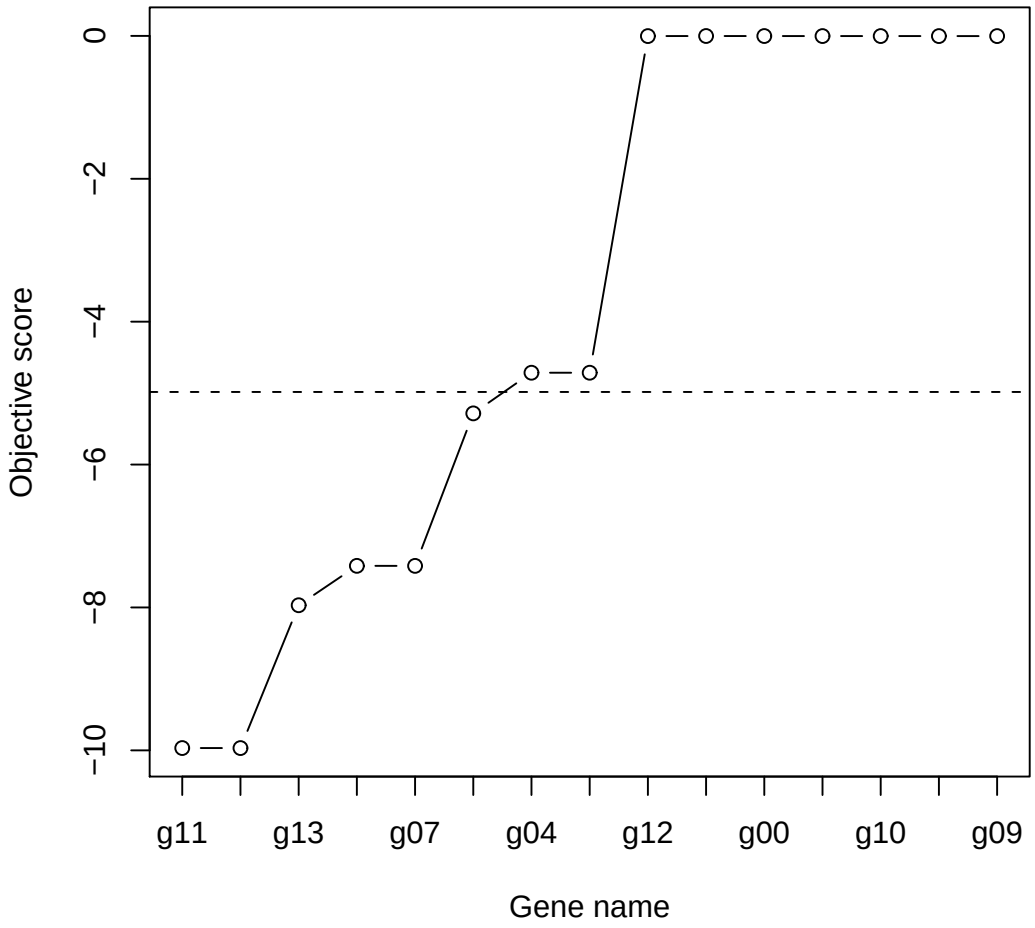


FIGURE 7.5: Plot of the objective score of a transsys program after applying cumulative gene knock-out operations. The horizontal axis shows the gene name that is been knocked out, the horizontal dashed line the 50% objective score cutoff. In this case 6 genes needed to be deleted for the objective score to reduced to half of the original transsys program score.

random initial conditions and objective score robustness to cumulative network elements pruning.

Notched boxplots of the total number of pruned genes from the transsys programs that have been selected for objective score robustness in different initial reactor states and the total population are illustrated in figure 7.7

There is no observable overlapping to the boxplot notches, indicating that the median of the selected transsys programs number of cumulatively pruned genes is lower than the total transsys program population. Accordingly the Wilcoxon rank

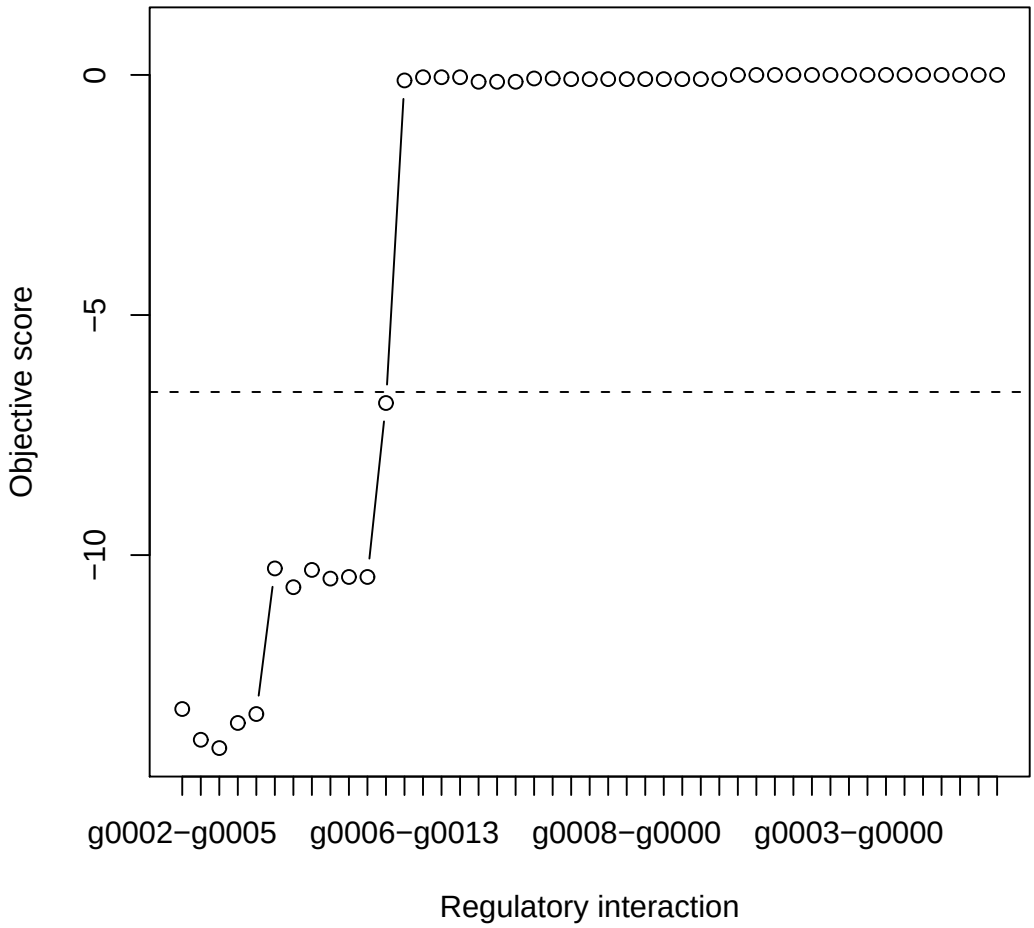


FIGURE 7.6: Plot of the objective score of a transsys program after applying cumulative regulatory interaction removal operations. The horizontal axis shows the regulatory interaction identifier (as a pair of genes that the interaction connects) that is been removed, the horizontal dotted line the 50% objective score cutoff. In this case 12 regulatory interactions needed to be removed before the objective score is reduced to half of the original transsys program score.

sum test supports a statistically significant difference in the vertices distribution location between the selected and the total distribution of transsys program pruned genes: $W = 19823$, p-value $= 2.492\text{e} - 06$.

Therefore, as it is depicted in the notched boxplots of figure 7.7 and supported by the Wilcoxon test, the selected transsys programs can tolerate significantly less cumulative gene knock- outs before they reach the 50% of the objective score of the original transsys program. There is no indication that the selected ones are more robust to the cumulative pruning procedure than the rest of the population.

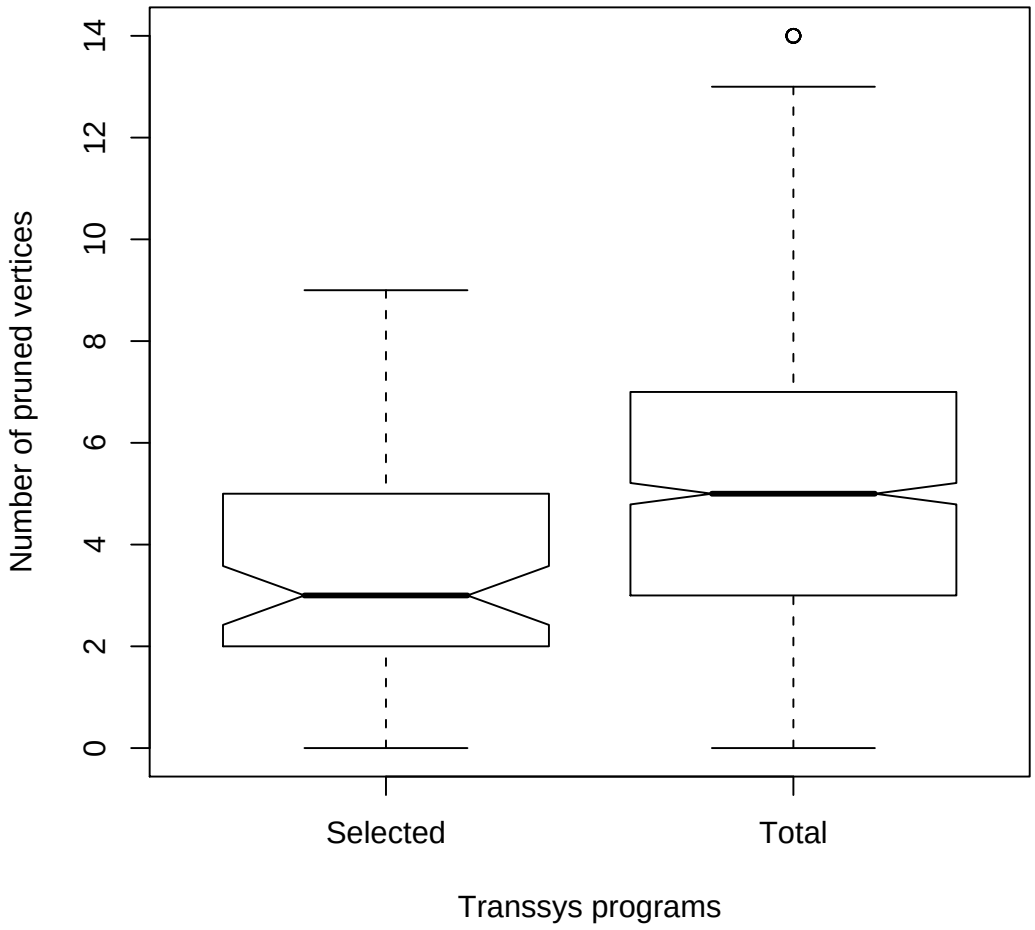


FIGURE 7.7: Notched boxplots of the number of pruned vertices before the objective score of the pruned transsys program reaches the 50% of the wild type. The selected transsys programs are the ones that passed the robustness selection process in section 7.1 and the total the full set of transsys programs that underwent the cumulative gene pruning procedure.

Similarly notched boxplots of the number of pruned edges (regulatory interactions) from the transsys programs that have been selected for objective score robustness in different initial reactor states and the total transsys program population are presented in figure 7.8

Again, the boxplot notches do not overlap and the median of the selected transsys program cumulatively deleted regulatory interactions is lower than the total transsys program population. Accordingly the Wilcoxon rank correlation test supports a statistically significant difference in the edges distribution location between the

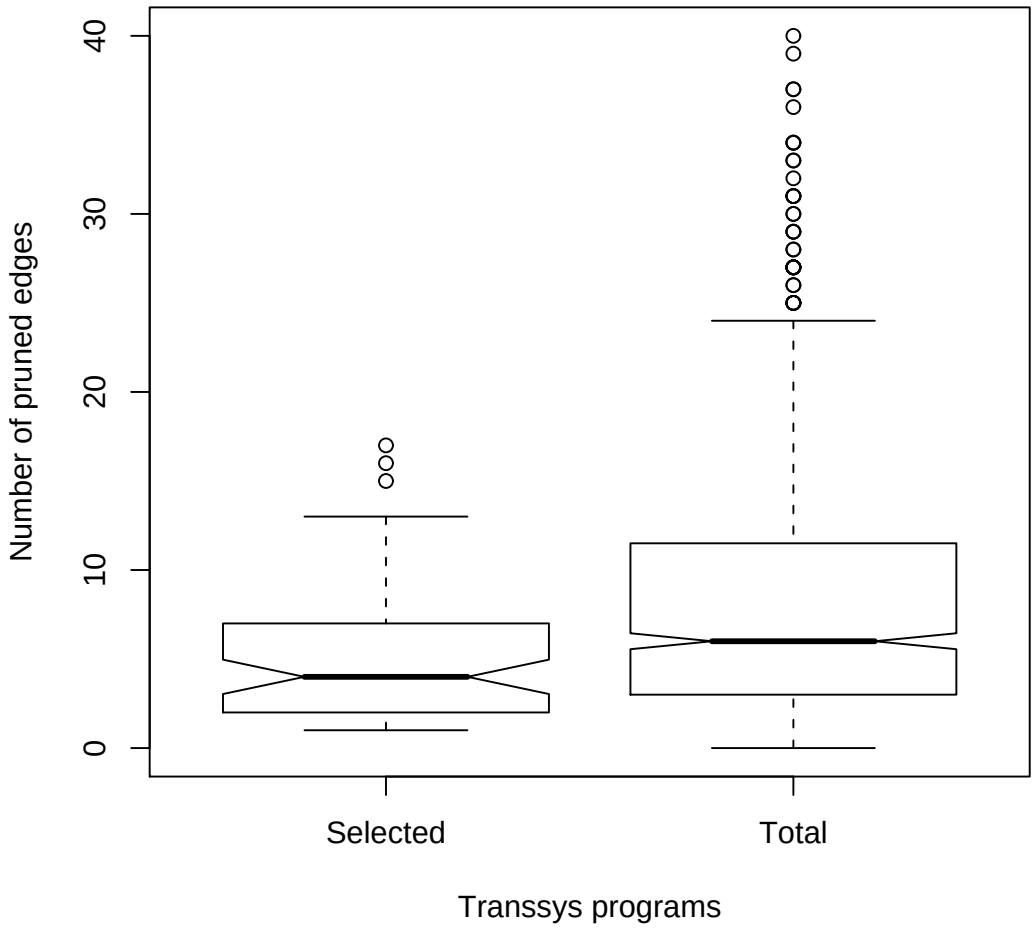


FIGURE 7.8: Notched boxplots of the number of pruned edges before the objective score of the pruned transsys program reaches the 50% of the wild type. The selected transsys programs are the ones that passed the robustness selection process in section 7.1 and the total the full set of transsys programs that underwent the cumulative edge pruning procedure.

selected and the total distribution of transsys program pruned genes with a low p-value: $W = 22635$, p-value $= 0.0006339$.

The selected transsys programs have significantly less genes and regulatory interactions pruned before the objective score reaches the 50% threshold than the total transsys program population. Both the gene and the regulatory interactions robustness to pruning investigations were unsuccessful in revealing any association between one aspect of robustness, that is the objective score robustness for different initial reactor states, and the second aspect of robustness being robustness to cumulative elements pruning.

Topological properties of GRNs selected to exhibit the "stripy lattice" property and have robust objective score to initial condition variability have been investigated in this chapter. The network diameter of the selected GRNs has been found lower than the diameter of the rest of the GRN population, suggesting that the small world phenomenon is associated both with the "stripy lattice" property as well as to the robustness to initial condition variability on GRNs. Further studies of topological properties of the robust GRNs have failed to identify links between the initial condition robustness and robustness to cumulative network pruning. The experiments however took in to account a single description of robustness to cumulative pruning, the number of network elements removed before the objective score reaches the 50% of the original transsys program objective score and might have overlooked others. Therefore the results of the topological robustness, given the current experimental design and analysis can only be characterised inconclusive.

Chapter 8

Conclusions - Outlook

This thesis aimed at providing a computational framework to simulate a biologically relevant phenomenon, that is to generate gene expression heterogeneity which is higher on spatialy extended system (a lattice) than in a background model (a well stirred reactor). The spatial heterogeneity mechanism was based on a reaction-diffusion system were transsys provided the mechanism for the reaction part and the spatial structure (the 2-dimensional orthogonal lattice) the diffusion component of the system. The system was able to reproduce the patterns that are characteristic of reaction-diffusion systems and can be classified at the genearl category of Turing-Meinhardt patterns. The predominant pattern that was observed in the lattices was the "stripy lattice" pattern, as given the size of the lattices in this work (relatively short height compared to width) any spatial heterogeneity will be observed only in the forms of stripes. Section 4.1.3 provides illustrative examples of "stripy lattice "spatial heterogeneity. For the development of the background experiment, to represent systems that lack any notion of spatial organisation, a null model was developed in the form of a well stirred reactor. It provides a concept for testing GRNs with negative control model in addition to the main model and to my knowledge is a negative control appeared for the first time in this work and it can be used as a background experiment for studies on spatial extended systems. In addition, the thesis aimed at devising a measure to quantify spatial gene expression heterogeneity and score GRNs accordingly, and studying and characteristic network topological properties of GRNs in terms of their capacity to generate the "stripy lattice" property (second aim

point as introduced in section 1.6). I presented the development of a computational framework able to capture this property and devised an objective score to quantify it, in accordance with previous studies of gene expression on lattices (Bignone, 1993; Keränen, 2004). I use this score as the objective of an optimisation approach trying to find GRN parameters such that the heterogeneity measure will be minimised. Low objective scores are connected with higher gene expression heterogeneity on a lattice compared to a well stirred reactor (chapter 3), thus low objective scores are associated with the "stripy lattice" phenomenon, as demonstrated in section 4.1.3. I hypothesise that by random sampling the topology space of random networks that share common features with biological relevant networks –like the edge density and the number of vertices– and trying to optimise for the dynamical parameters of the network, GRNs with topologies that favour lower objective scores will be more amenable to optimisation and thus will be distinguished from topologies that do not have the potential to generate heterogeneity. The findings presented in chapter 6 as well as the random sampling benchmark of the optimiser (section 4.1.2) provide evidences that the optimisation, although a simplistic one, is able to discriminate network topologies with higher propensity to generate heterogeneity in spatially organised systems. Therefore I considered the first two objectives, the reproduction of the spatial heterogeneity phenomena and the development of a measure for the quantification of these phenomena as accomplished, (aims have been formally introduced in section 1.6).

In addressing the 3^{rd} central aim of the the thesis chapter 5, section 1.6) is studying associations (using correlation studies) between GRN topological properties and the objective score. The network density has been found to correlate significantly with the transsys program objective score after optimisation, suggesting that there is a certain amount of regulatory interactions, found to be between 4 and 8 per gene in this study, for a GRN to have the capacity to generate "stripy lattice" paterns. Global network topological properties including average clustering coefficient, number of network cycles and average cycle length, as well as average path length, were found to have no correlation with the transsys program objective scores after optimisation. However, small network diameter correlates with low objective scores suggesting that the small world phenomenon in GRNs pertains to a biologically relevant property such as gene expression heterogeneity, as represented by the low objective score. As far as for the second characteristic of small world networks, the clustering coefficient, no significant correlation has been detected in any of the experiments. This indication together with the

fact that the average clustering coefficient is not unambiguously defined for directed networks, designates the diameter, and not the clustering coefficient, as a characteristic measure for the small world phenomenon in GRNs.

Chapter 6, aligns with the 4^{th} of the theis central aims and studies the impact of a set of factors –including network generation mechanism, network topology, network dynamical parametrisation and initial reactor state– on the objective score. Primarily using analysis of variance, the variability of the scores of already optimised transsys programs is studied with regard to each individual factor. The findings indicate that the transsys program primarily and the GRN toology as well as the dynamical parameers secondly are more significant determining factors of the objective score than the initial reactor state or the network topology alone. In fact the results illustrated in figure 6.1 indicate that the initial reactor state can be arbitrarily chosen without significantly affecting the statistics of the objective scores of GRNs.

Finally, addressing the last aim of the thesis (section 1.6) chapter 7 studies aspects of GRN topological robustness as well as GRNs objective score robustness. The experiments attempt to find associations between two aspects of robustness: robustness to different initial reactor states and robustness to cumulative network element deletions. For GRNs selected for higher robustness to initial reactor states, the diameter of selected GRNs is marginally significantly lower than the total GRN population. Thus in line with chapter's 5 finding that GRNs with smaller diameter have lower objective score here GRNs selected for robustly low objective score have lower diameter. However, studies to relate robustness of GRNs to cumulative pruning of genes or regulatory interactions with the robustness to initial reactor state did not bring up any conclusive results.

Summarising the findings, this thesis has studied GRNs of which the topology was generated with a random and unbiased way and investigated topological properties of GRNs with regard to their ability to generate heterogeneity. The GRN topology was randomly sampled from the space of all potential GRN topologies. Therefore the sampled topologies were not subject to any bias introduced by either biological processes (i.e. evolution), or experimental processes, such as bias incurred by methods that detect some regulatory interactions better than others or bias introduced by the focus of researchers in genes and regulatory interactions which are members of an interesting biological processes. GRN topologies were sampled from topologies that share a generic biological property that is the ratio of edges to

nodes (i.e the network density) and that is close to one observed in biological processes, moreover the power-law network generation mechanism has been proposed to be a representative generation process of biological networks (Almaas, 2007). The parameter optimisation of GNRs whose topology was randomly sampled from an unbiased space of network topologies has provided insights to topological determinants for GRNs that exhibit low objective score such as the small diameter. In addition, low objective scores is associated to individual network elements and correlated with their centrality measures. Furthermore, an indication that the initial reactor state does not have significant effect on the statistics of the objective scores of GRNs was established, however the fact that the transsys program was found the most significant factor determining the objective score, signifies that GRN dynamics is not determined solely by the topology of the network but from the dynamical parameters settings as well. The nested structure of the transsys program generation approach –the fact that every transsys program group is a subgroup of a topology group– enables the level of transsys programs groups to capture more of the variability of the objective score and as a result makes the topology vs. dynamical parameters relationship infeasible to elucidate with the current experimental procedure. Similar summarising can be made for the robustness studies, no striking indication was found that might connect robustness to initial reactor states with robustness to cumulative pruning, however the results are inconclusive and a potential connection between the two aspects of robustness can not be ruled out.

The work in the context of this thesis has generated an array of tools and experimental procedures to study topological properties of GRNs in spatially extended systems, and have conducted a significant set of experiments to study this relationship. Pattern formation mechanisms have been proposed and study analyticaly by Alan Turing and later by Hans Meinhardt and Alfred Gierer (Gierer and Meinhardt, 1972; Turing, 1952) where the model was a set of partial differential equations with a fast diffusing inhibitor and a slow diffusing activaror that generate spatial heterogeneity of concentrations. However here a mechanism that involves gene regulatory networks and that shows that these systems have the capacity to organise heterogeneity through optimisation of their dynamical parameters by a random local search. Thus the generation of Turing/Meinhardt patterns can be reproduced by the mechanisms described in this thesis. The topologies of GRNs capable of pattern formation has been also studied elsewhere (Salazar-Ciudad et al., 2000; Salazar-Ciudad, Newman, and Sol, 2001), where a set of potential minimal networks that are able to generate differential gene expression along a

string of connected cells has been described. This thesis studies gene expression heterogeneity on a 2-dimensional structure and has proposed a set of global and local topological properties that identifying GRNs rather than specific topologies. In addition to that this work has provided a contribution toward setting a framework in which GRNs can be assessed for their capacity to generate spatial gene expression heterogeneity by studying the distributions of robustness to initial reactor states (section 7.1 and in particular figure 7.1) of an unbiased sample of topologies. Setting the quantiles (figure 7.1) at which GRNs objective score is robust against initial reactor states one is able to position a given GRN and compare its capacity to generate spatial heterogeneity with the unbiased sample of GRNs. However, limitations of this approach lay in the fact that not any given GRN can be the input of this procedure as only networks that can be represented by directed graphs, that is networks that do not take in to account synergistic or antagonistic effects between the regulators, can be analysed in the framework developed in this thesis. A promising extension of this work towards specifying sets of topologies would be to initiate the optimisation experiment with a fully connected network and gradually delete the edge with the lowest contribution to the objective score until any further deletion will have a detrimental effect to the objective score. Starting from a set of different initial conditions a set of topologies can be retrieved and potentially some common topological properties can be identified.

Moreover, the work on the robustness of the *D. melanogaster* segment polarity GRN (von Dassow et al., 2000), has identified a core network topology which is robust both in terms of random edge deletions as well as initial conditions perturbations. In this work, the initial conditions robustness, was proposed to be a property of developmental GRNs for buffering against developmental noise. The results of chapter 6 that the "stripy lattice" property can be observed, on average, with any arbitrarily chosen set of random initial reactor states, and that there exist transsys programs which exhibit robustness to the initial reactor state choice are very close to the robustness to developmental noise that (von Dassow et al., 2000) have examined. In the context of the robustness studies and in the inquest of finding relationships between different aspects of robustness research focused on the identification of a core topology common for a significant number of the pruned networks, will be a promising one. Furthermore, it would be motivating for a different experimental endeavour to elucidate the topology vs. dynamical parameters relation with regards to GRNs' dynamical properties, a relationship

that is tried to get resolved in the wet biological experimental level (Kim, Shay, O'Shea, and Regev, 2009).

In addition an interesting branching of the research conducted here will be the study of GRNs not in a static spatial structure like the lattice presented here but in a dynamic one, like a collection of cells in a 2-dimensional structure where cell division is allowed. More interestingly, the cell division events of this growing structure should be under the control of the factors of the regulatory network. This structure embedded in environments where there is a source of variation (nutrient or light gradient, space antagonism) can facilitate the topological studies of GRNs in an evo-devo context and the impact of the shape of the growing structure on gene expression dynamics. Parts of the computational infrastructure that has developed for this work can be used towards this goal, but an addition design of a novel computational framework is required for conducting these studies.

Finally, the findings of this work have contributed in identifying network topological properties of GRNs that encode for the "stripy lattice" property, a property that can be connected with higher level biological properties such as cell differentiation and pattern formation. Furthermore, the process of describing a non-specific biological phenomenon and model it in computational terms, quantify the characteristic property of this phenomenon and associate this property with features of the system that generate this phenomenon, constitutes a contribution towards a means of measuring success of the grand challenge to build a complete reactive model ((Mareé, Panfilov, and Hogeweg, 1999) for a classical example) of a biological organism as described by D. Harel as the "Grand Challenge" in Systems Biology (more in (Harel, 2005)).

Appendices

Appendix A

transsys Language Lexical Elements

Technical details of lexical elements of the transsys language useful for the understanding of the syntax of a transsys program.

Every transsys program starts with the keyword `transsys` followed by the name of the transsys program. All gene and factor declarations should be in the body of the transsys program, enclosed in curly braces. A transsys program contain no statements is accepted by the transsys language and can be like the following:

```
transsys abc
{
}
```

A valid transsys language syntax to specify two factors in the transsys program that was introduced above can be the following:

```
transsys demo
{
  factor FactorA
  {
    decay: 0.3;
    diffusibility: 0.1;
  }
```

```
  factor FactorB
  {
    decay: 0.3;
    diffusibility: 0.2;
  }
}
```

In the context of this work every factor declaration explicitly states the decay rate and the diffusibility expressions as real value numbers. Here diffusibility expressions vary and for instance `FactorB` diffuses more quickly than `FactorA`.

A gene's syntactic structure including the promoter and the product block and specifying a constitutive, an activate and a repress promoter element is presented. Gene `geneA` encodes for the `FactorA` of the transsys program specified in the previous paragraph, including `geneA` in the transsys program `demo` will give the following:

```
transsys demo
{
  factor FactorA
  {
    decay: 0.3;
    diffusibility: 0.1;
  }

  factor FactorB
  {
    decay: 0.3;
    diffusibility: 0.2;
  }

  gene geneA
  {
    promoter
    {
      constitutive: 0.1;
      FactorB: activate(1.0, 0.5);
      FactorA: repress(1.0, 0.1);
```

```
    }
    product
    {
      default: FactorA;
    }
  }
}
```

Note that in this example the product of the gene regulates its own expression, this represents a loop in network terms and an auto(self)-regulation in dynamical systems terms.

The transsys dynamical parameters are real valued parameters that specify the diffusibility expression and the decay rate of factors and the constitutive expression, the α_{spec} and the a_{max} (for either activation or repression) of the promoter block of a gene. The parameters are highlighted in the part of a transsys program illustrated in figure A.1

```
...
factor f03
  {decay:  0.2;
  diffusibility:  0.1;}
gene g03
  {promoter
    {constitutive:  0.1;
    f10:  repress(3.6, 0.1);
    f05:  repress(2.6, 0.4);
    f03:  activate(2.2, 0.2);
    f02:  repress(6.3, 0.4);
    f09:  activate(2.2, 0.5);}
  product
    {default:  f03;}
}
...
```

FIGURE A.1: A part of a transsys program, with all its dynamical parameters highlighted in blue.

Appendix B

Patterns on Lattices

Grey-scale images of factor concentrations from lattices of various sizes are presented here. These are factor concentration values that have been simulated from transsys programs that have been optimised using the reference parameter sets. Factor concentrations from transsys programs that exhibit spatial heterogeneity on lattices have been selected for illustration purposes here.

B.1 Lattices from the Reference Parameters Set

The size of the reactors (both the lattice and the control) was 60 cells width and 5 cells height (according to the control parameter settings) and all the rest of control parameters were kept equal to the reference set. A characteristic pattern of stripes of width few (or more) cells with higher factor concentration that the neighbouring cells has been observed to several factors from transsys programs with low objective score after optimisation (figure B.1). Other factors from the same transsys programs exhibited the reverse pattern, that is a stripe of width of few cells that has lower factor concentrations than the neighbouring cells (figure B.1).

B.2 Elongated Lattices

The pattern that is described on the reference parameter settings lattices consisting of stripes of cells where the concentration of a factor is relatively lower (or higher) than the neighbouring cells exhibits a characteristic periodicity. This

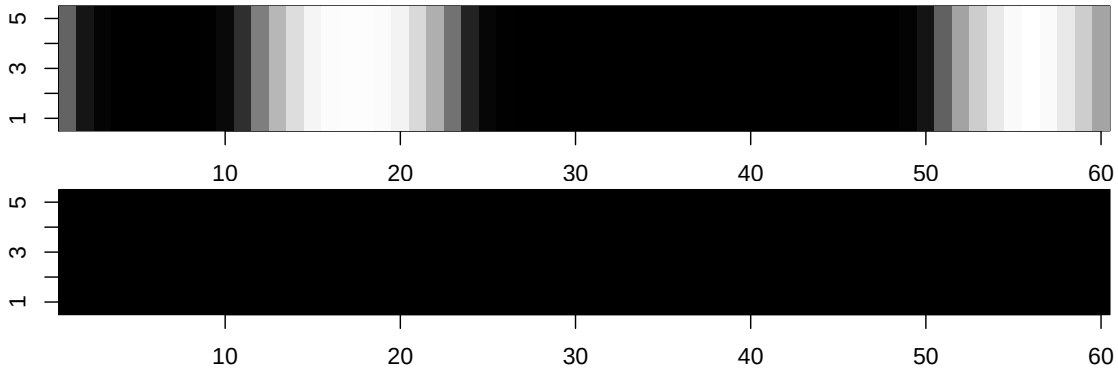


FIGURE B.1: Grey-scale images of factor concentrations of one factor on a lattice after optimisation (top) and a well stirred reactor (bottom). Concentration values range from ≈ 0 (black) to ≈ 0.27 (white). The information content of the factor on the lattice is ≈ 1.3 bits and for the WSR ≈ 0 bits

periodicity can be measured from the distance between two peaks (or valleys) of factor concentration, or the characteristic scale of the stripes. To illustrate a case of this characteristic scale of the pattern described above lattices that have been elongated with regard to their width have been subject to optimisation using the reference values of the control parameter sets. The only difference was the width which has been increased 5 times, so the reactors illustrated in this section are of height 5 and width 300 cells.

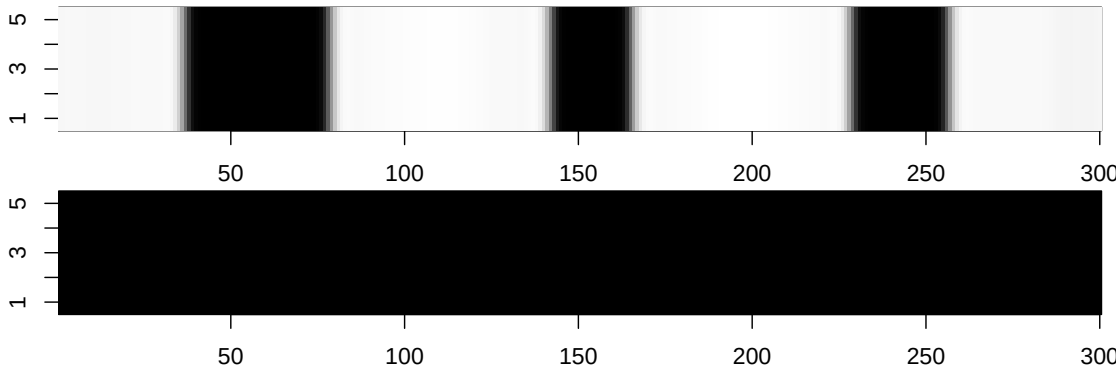


FIGURE B.2: Grey-scale images of factor concentrations of one factor on an elongated lattice after optimisation (top) and a well stirred reactor (bottom). Concentration values range from ≈ 0 (black) to ≈ 1.0 (white). The information content of the factor on the lattice is ≈ 0.48 bits and for the WSR ≈ 0 bits

Note that the greyscale images of figure B.2 are not scaled (i.e. the cells are not squares as in figure B.1). Therefore each cell on the elongated lattice plots a width of 1/5 of its height.

B.3 Squared Lattices

The pattern of high (or low) factor concentration stripes described in the two previous sections is a characteristic pattern of elongated lattices (i.e. lattices where the width is a multiple of the height). Here I illustrate the presence of a different characteristic pattern that appears in lattices that are square, (have equal width with height). The factor concentration grey-scale images that are presented here are generated from transsys programs after optimisation with the reference values of the control parameter sets and reactor sizes of width 30 and height 30 cells.

Note that a non stripy pattern appears in figure B.3, as the lattice is not rectangular anymore but has equal number of cells in its height and width.

reactor is initialised following an identical process like a lattice reactor, the factor concentrations on the ICR are updated with the same update function like a lattice apart from diffusion. By eliminating diffusion the factor concentration is determined solely by the GRN structure and cells do not exchange gene products with their neighbours. Gene expression heterogeneity on the lattice from simulations using the an isolated cells reactor as a control experiment are comparable with results obtained from the well stirred reactor as a control.

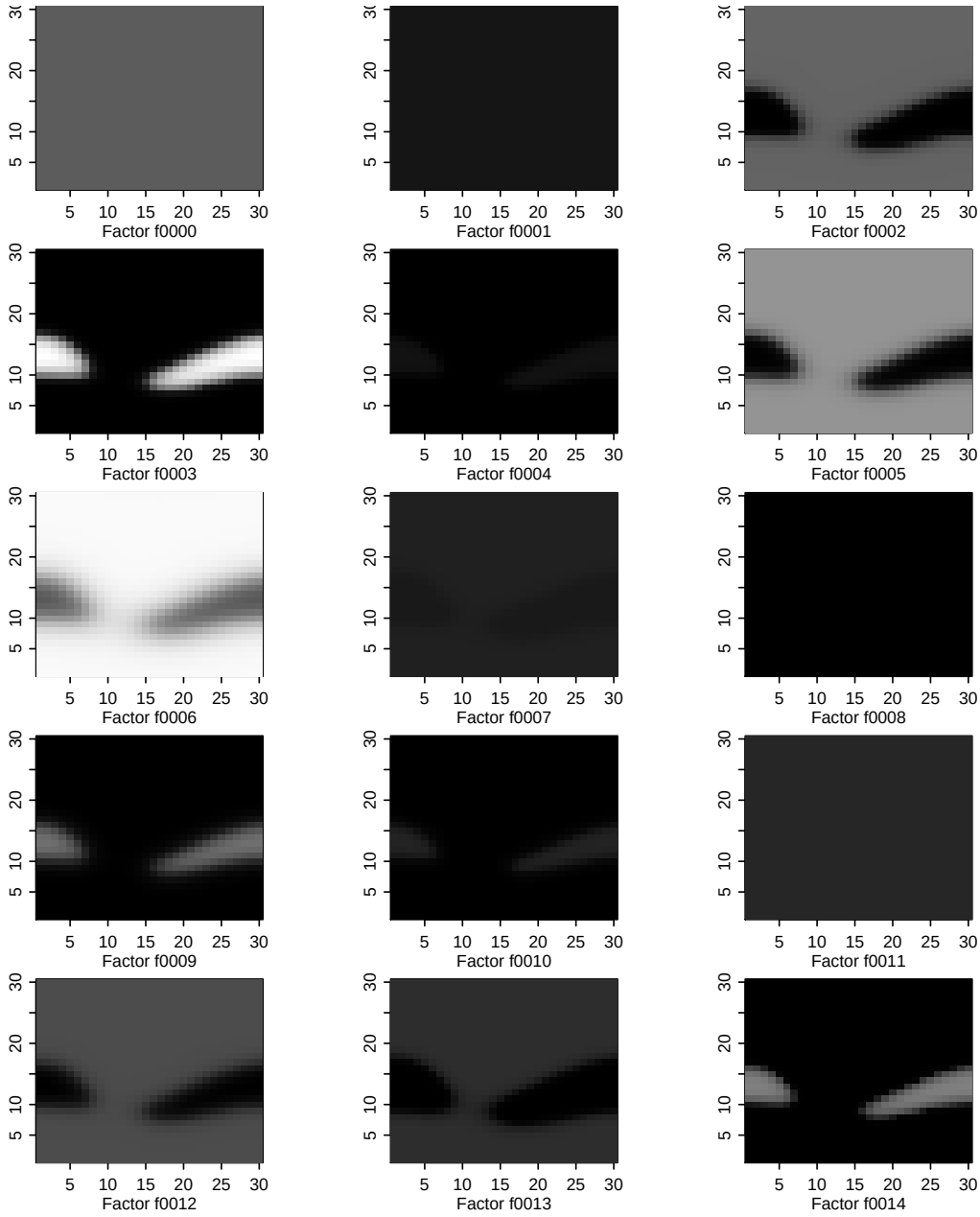


FIGURE B.3: Grey-scale images of factor concentrations of all the factors on a square lattice after optimisation. Concentration values range form ≈ 0 (black) to ≈ 2.4 (white). The information content measure for this particular transsys program was ≈ 13.9 bits

Appendix C

Large GRN individual elements results

Gene knock-out and edge deletion results are presented here, the results are equivalent to the ones presented in section 5.3, however refer to a larger transsys program with 25 genes and 75 edges and therefore the statistics are more robust owning to larger number of samples. The results presented here are all generated by following the reference control parameters set, apart from the network size.

C.1 Gene Properties

After performing a single gene knock-out experiment for every gene in a transsys program, the objective score difference from the wild type one has been correlated with the gene centralities and the number of cycles that the gene is a member of. Results from a representative transsys program which has exhibited a stripe pattern are presented in figure C.1

The correlation plots in figure C.1 illustrate results comparable to the plots presented in the individual network elements analysis. The transsys program analysed here has a higher number of genes (25 instead of 15) and thus the corresponding p-values are lower. The statistics though hold for an increased number of genes fact that corroborates the results of section 5.3.

C.2 Regulatory Interaction Properties

After performing an individual regulatory interaction deletion for every regulatory interaction in a transsys program the difference in the objective score of the wild type transsys program from each mutant has been calculated. The objective score difference due to individual regulatory interaction deletion is then correlated with two network properties of the edge that represents the regulatory interaction in the graph and two transsys program properties the $a_{\max}$ and the α_{spec}, the results are presented in figure C.2

Here again more regulatory interaction (75) compared to the results presented in section 5.3 provide further statistical corroboration of the results regarding regulatory interaction deletions. Figure C.2 illustrates correlation between the edge related measures and some weak correlation with the transsys program dynamical parameters.

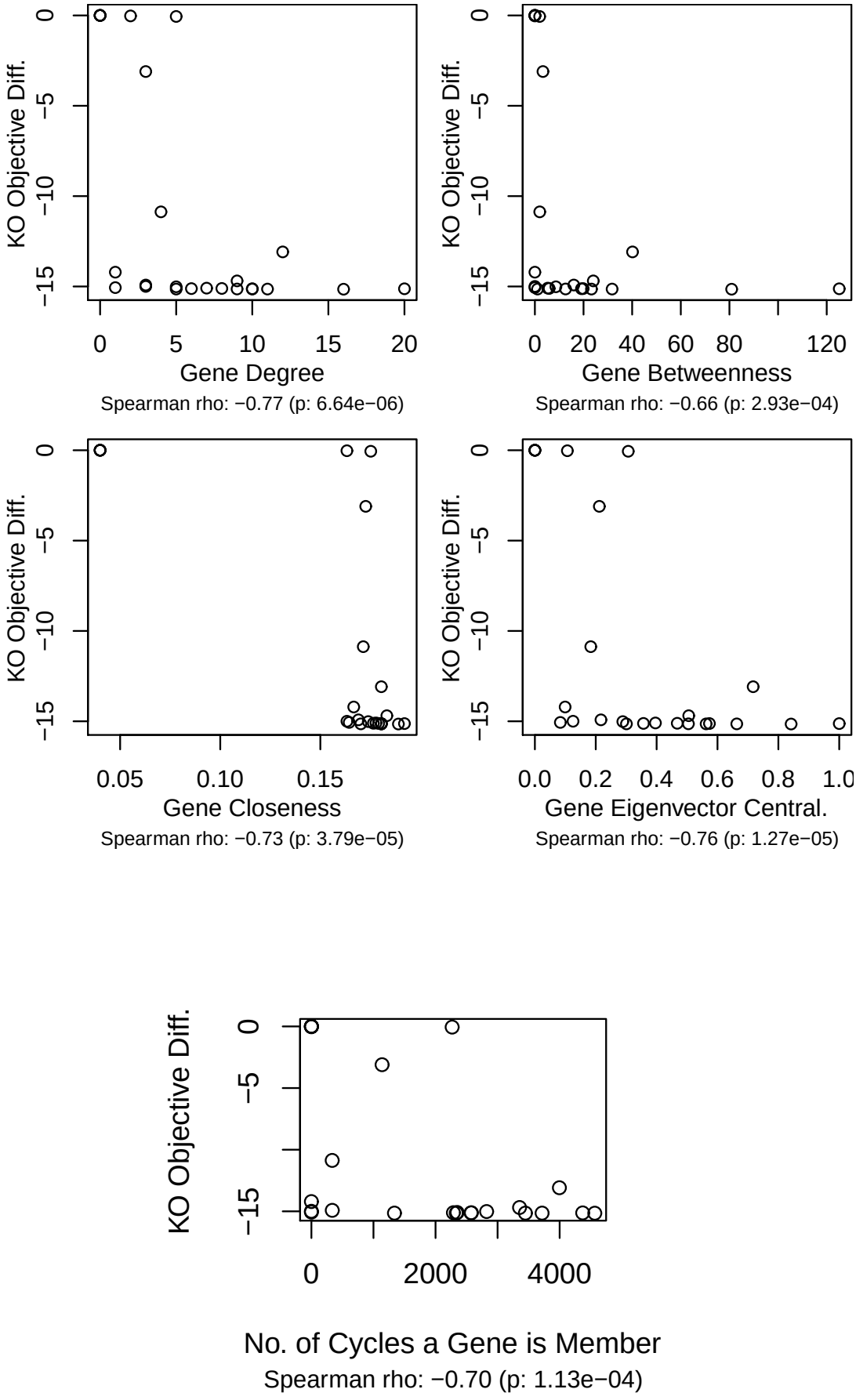


FIGURE C.1: Correlation plots of the difference in the objective score of the single gene mutant transsys program vs. centrality measures of the knocked-out gene (top). The same difference vs. the number of cycles the knocked-out gene is a member of.

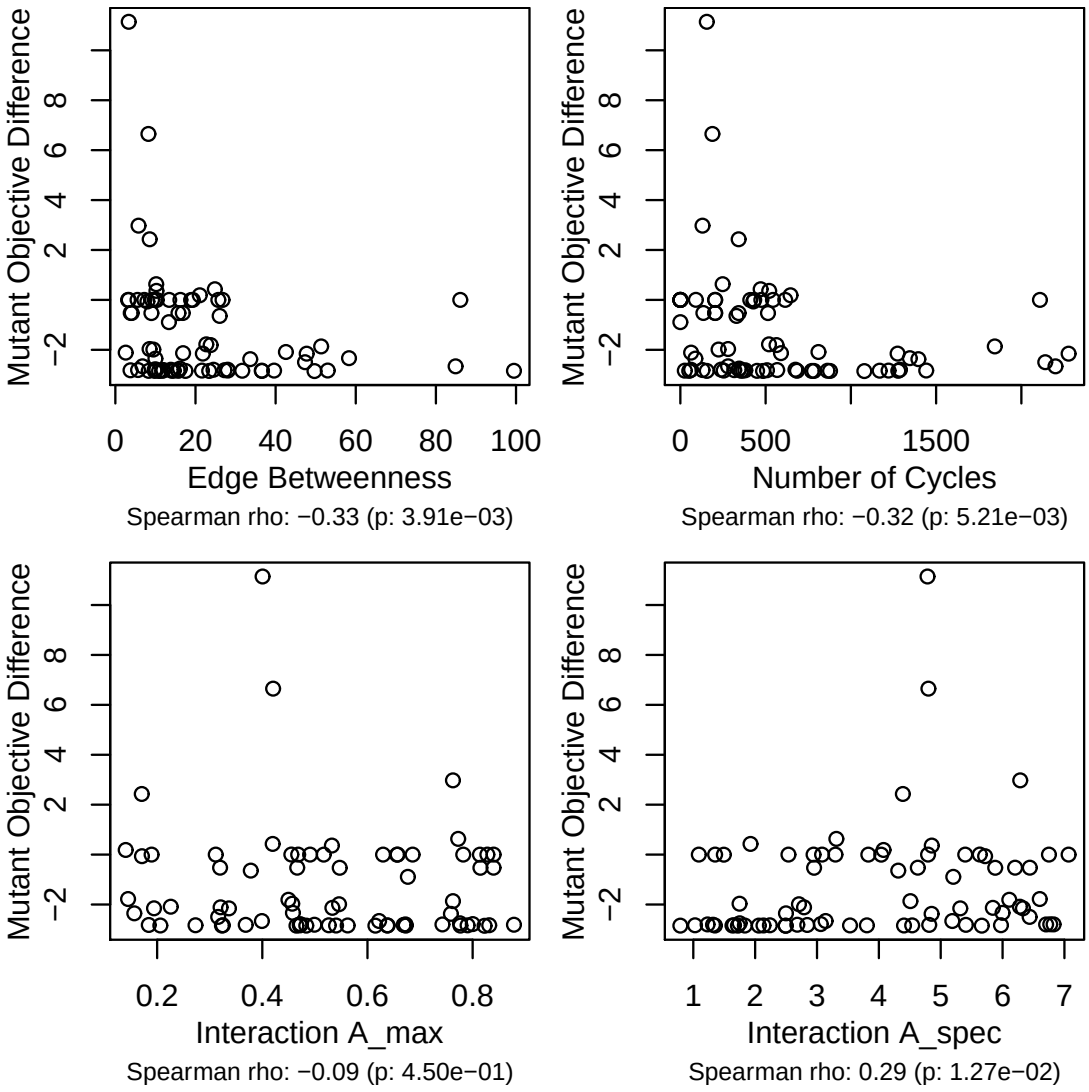


FIGURE C.2: Correlation plots of the difference in the objective score of an individual regulatory interaction deletion mutant transsys program vs. edge network measures of the deleted regulatory interaction (top). The same difference vs. measures pertaining to dynamical parameters of the transsys program.

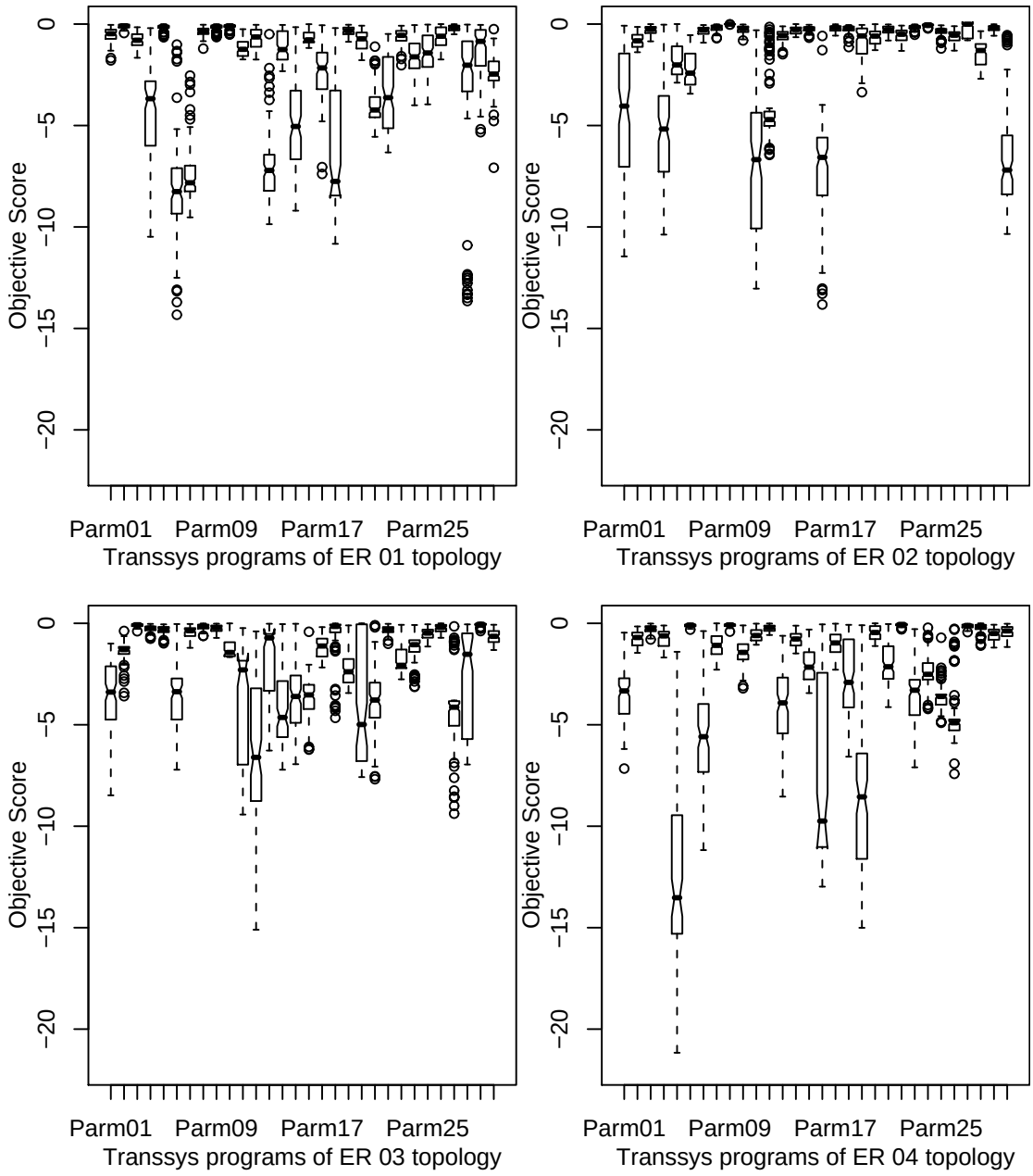
Appendix D

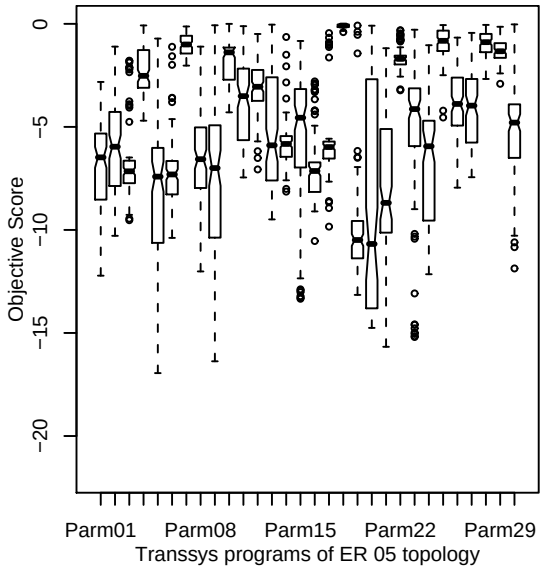
Initial Reactor State Experiment Results

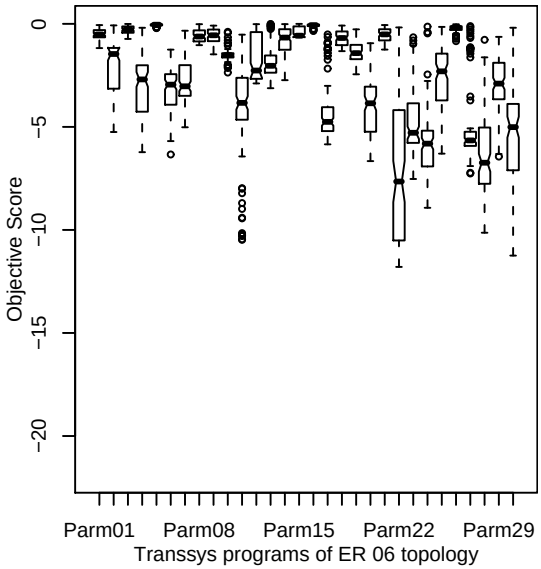
D.1 Transsys Program Parametrisations Boxplots

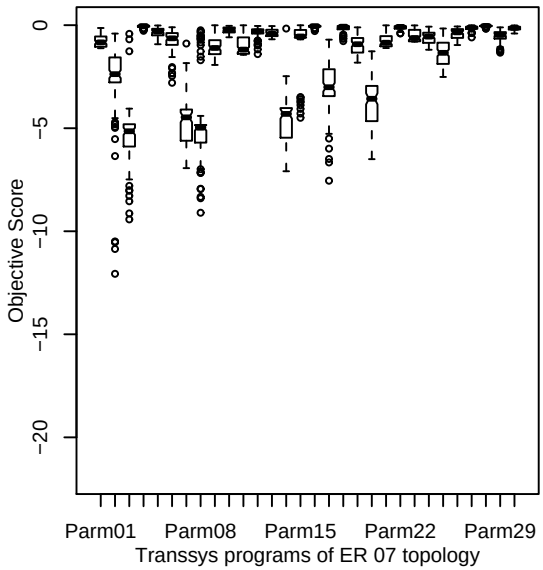
The full set of boxplots for each transsys program that has been generated from 30 different initial reactor states is presented here. It is the complement of the analysis presented and discussed at section 6.5.

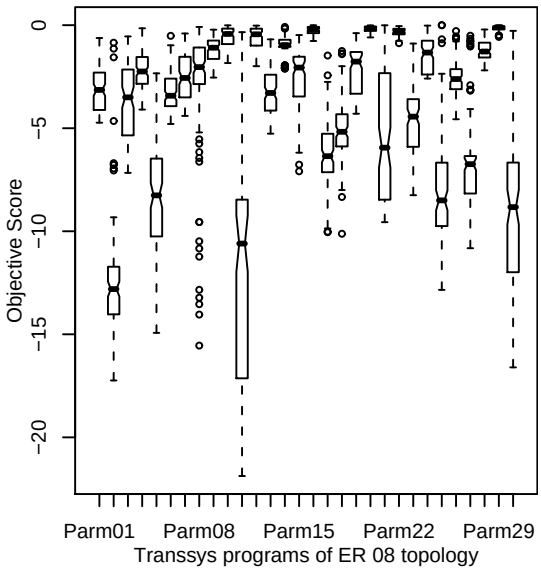
D.1.1 Erdös-Rényi networks boxplots

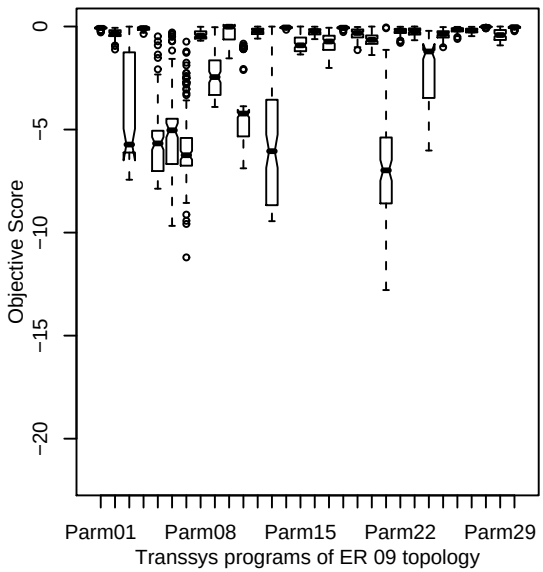


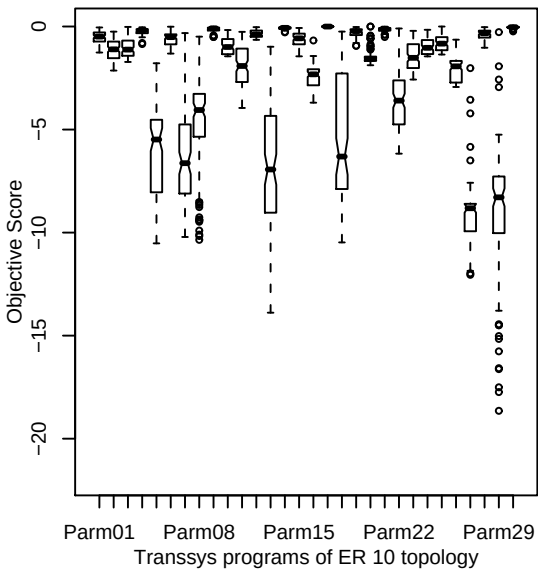

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 05 topology

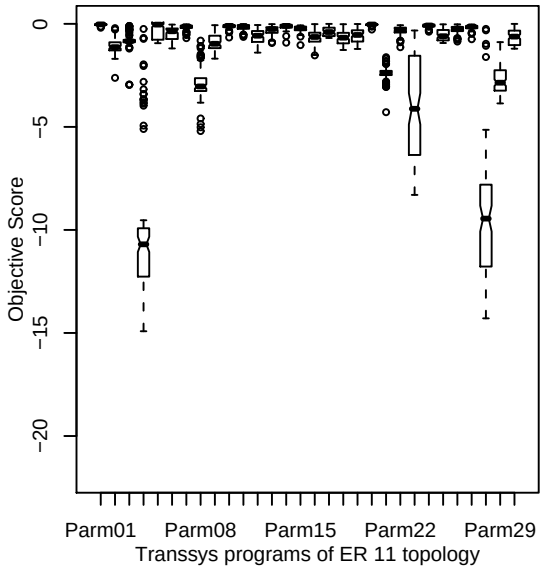

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 06 topology

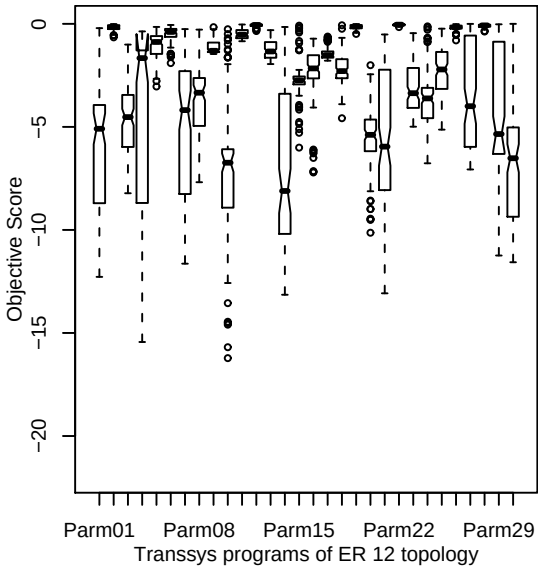

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 07 topology

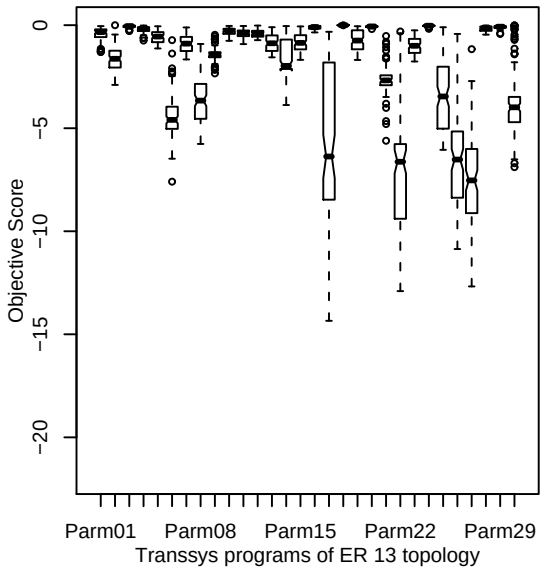

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 08 topology

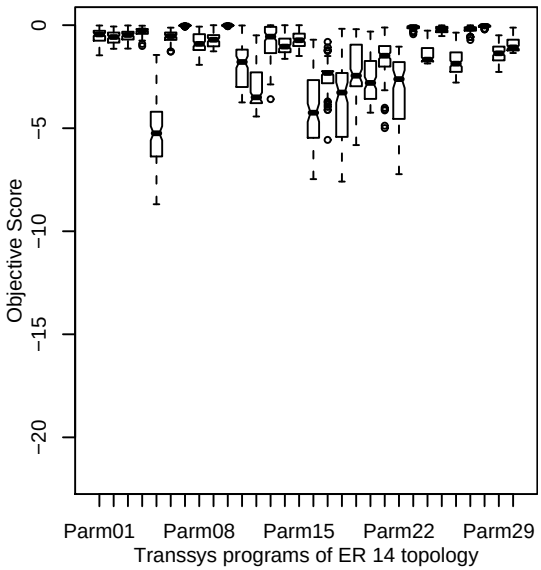

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 09 topology

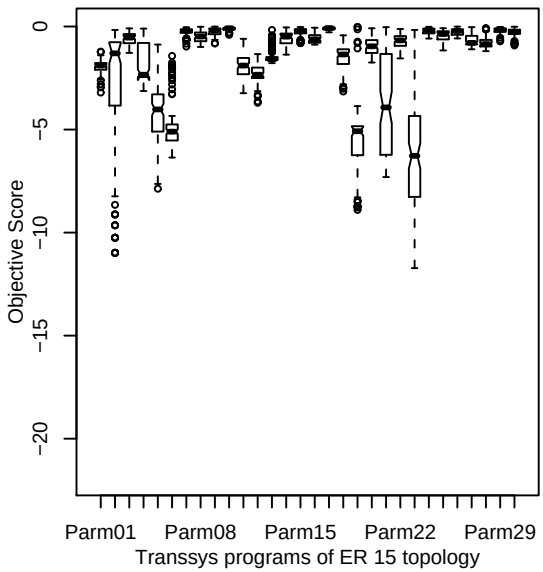

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 10 topology


Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 11 topology

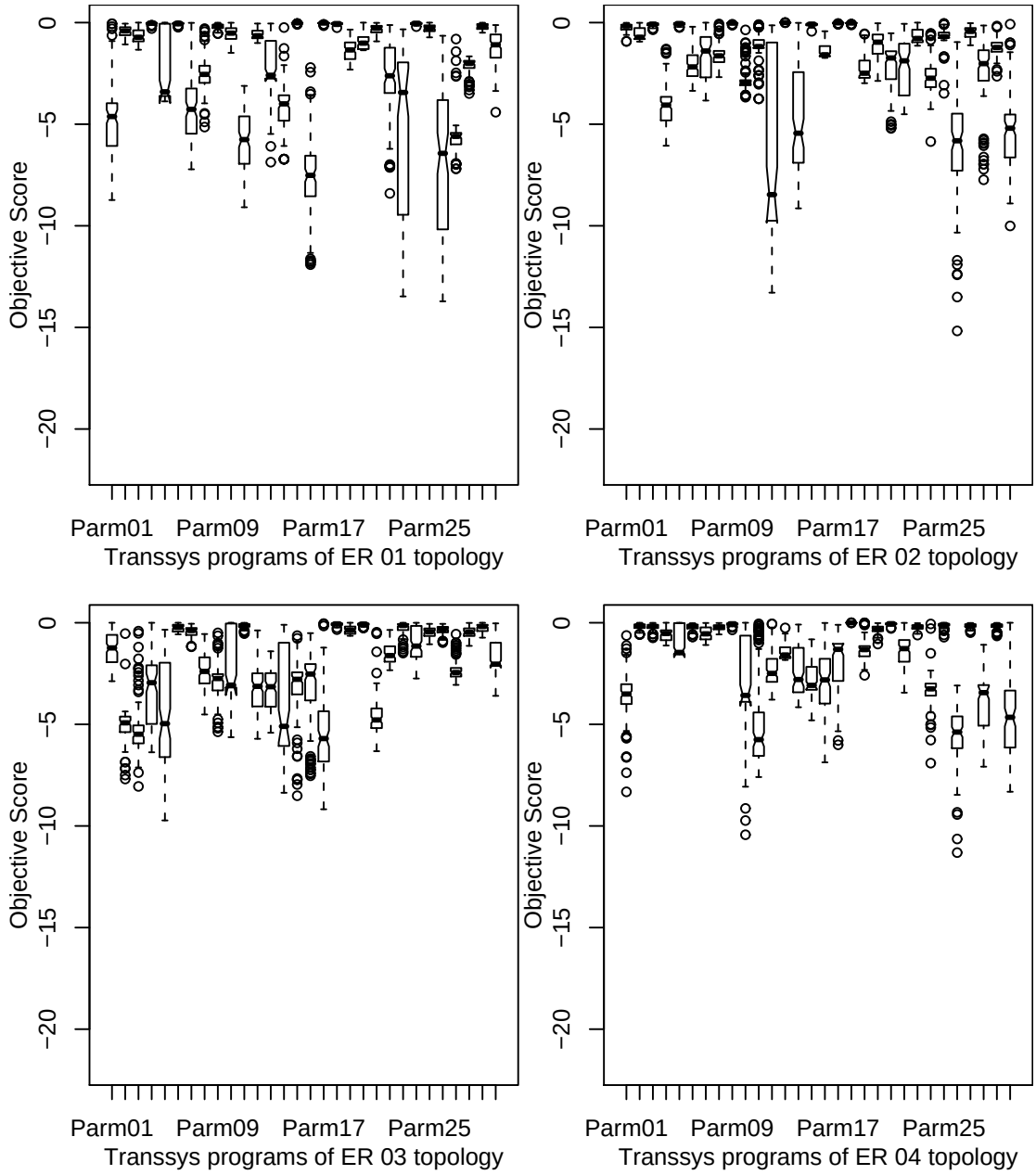

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 12 topology

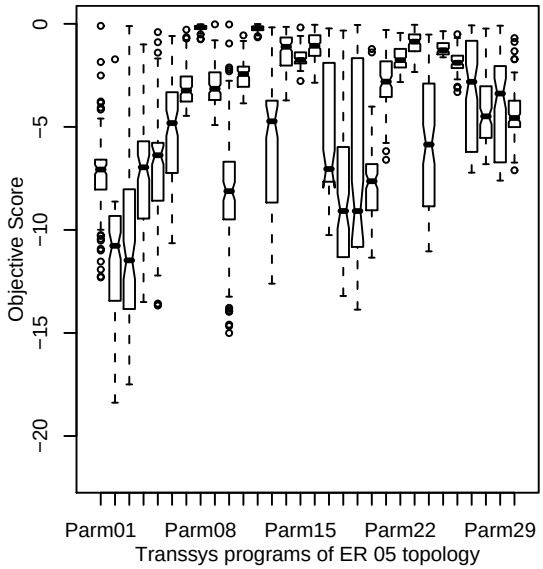

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 13 topology

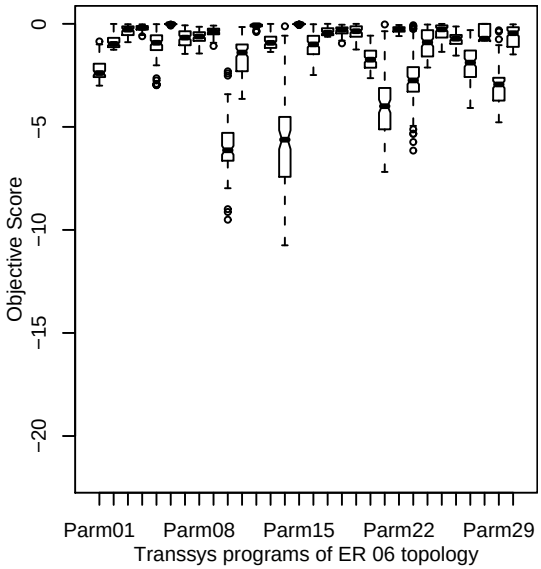

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 14 topology

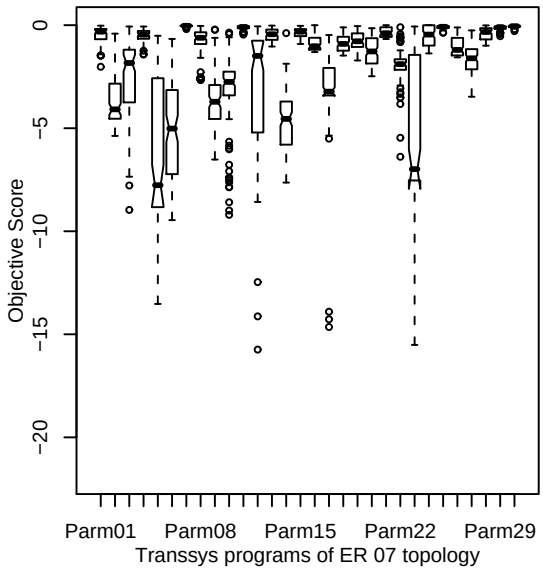

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 15 topology

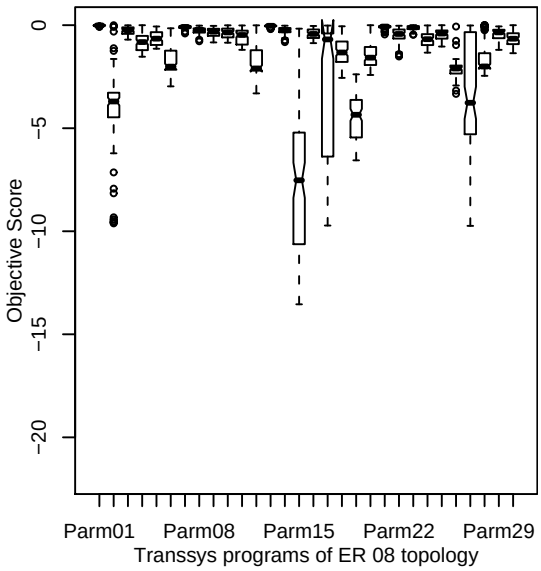
D.1.2 Networks with Power-law degree distribution boxplots

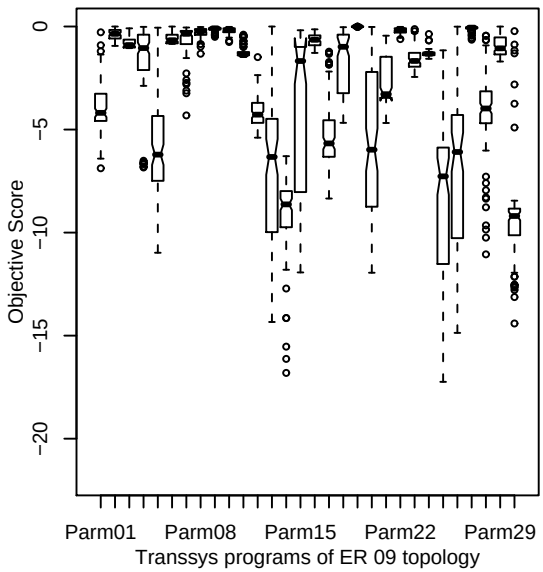


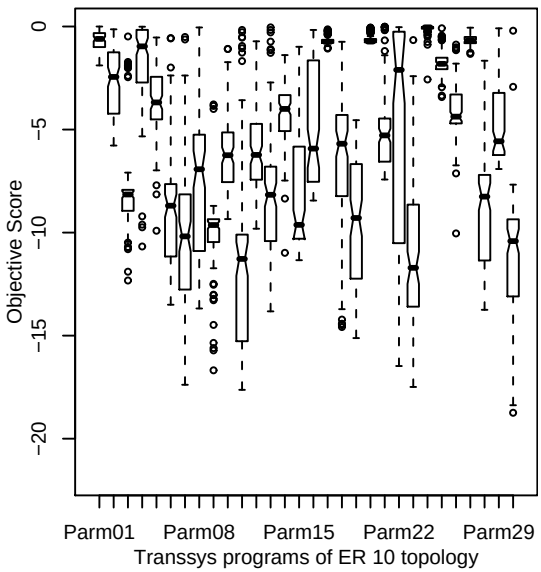

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 05 topology

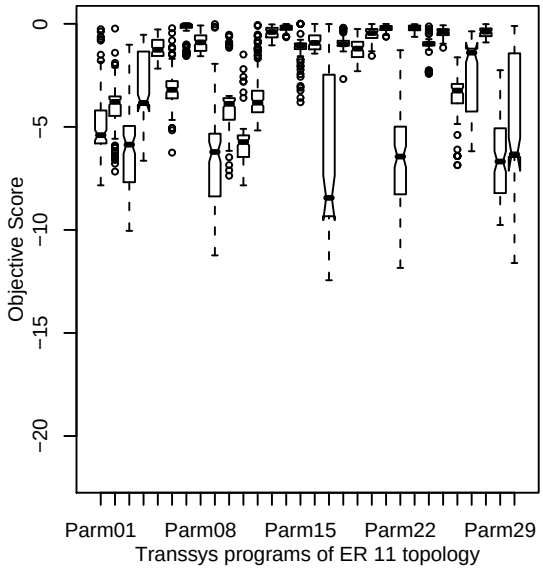

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 06 topology

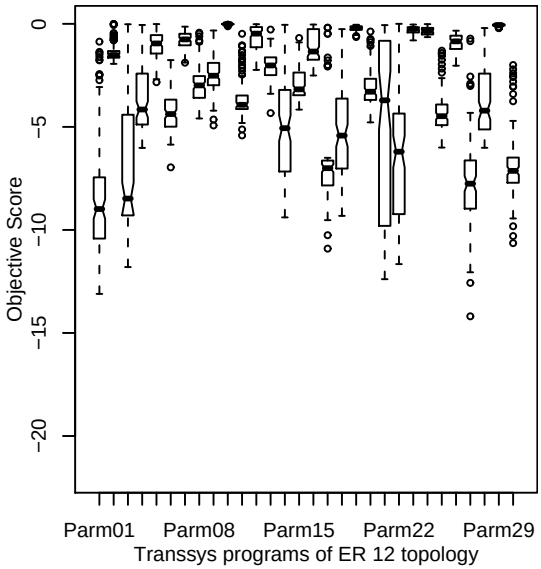

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 07 topology

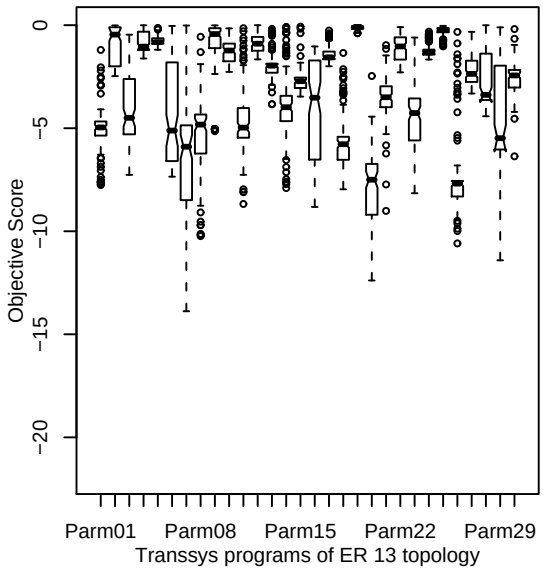

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 08 topology

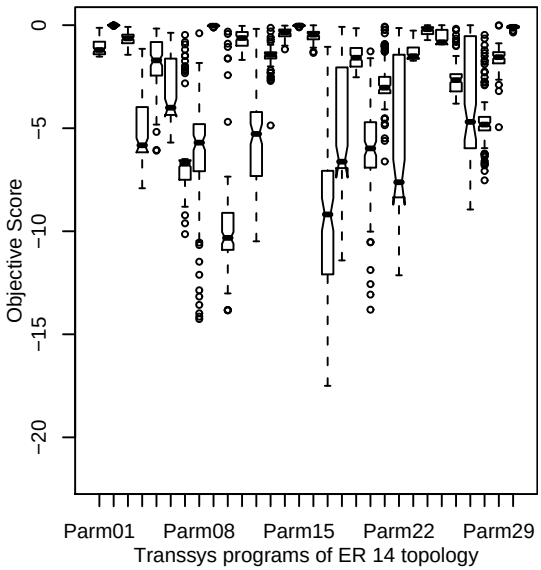

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 09 topology

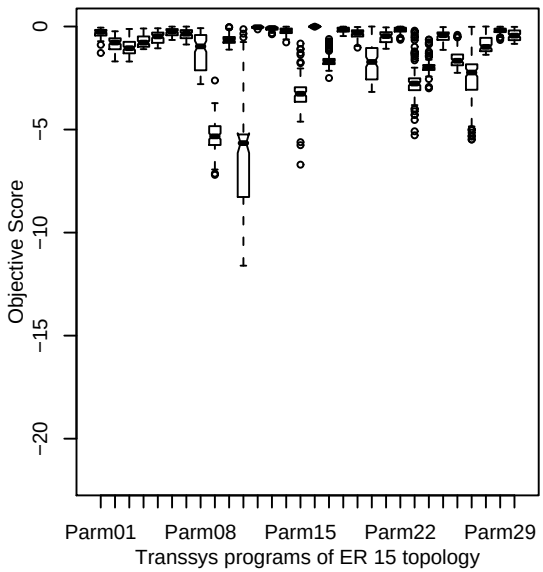

Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 10 topology


Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 11 topology


Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 12 topology


Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 13 topology


Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 14 topology


Objective Score
0
−5
−10
−15
−20
Parm01
Parm08
Parm15
Parm22
Parm29
Transsys programs of ER 15 topology

Appendix E

Robustness Studies

Additional figures from the robustness studies chapter can be found in this appendix. Refer to every figure's caption for more details.

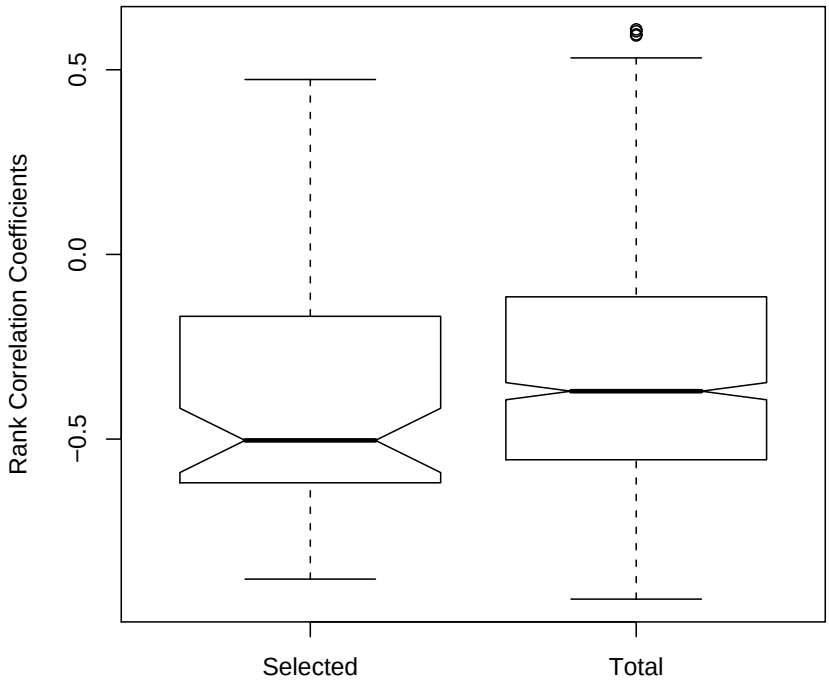


FIGURE E.1: Notched boxplot of the Spearman ρ rank correlation coefficient between the number of cycles a gene is a member of and the objective score difference from this gene knock-out, grouped in selected for robustness transsys program and the total transsys program population.

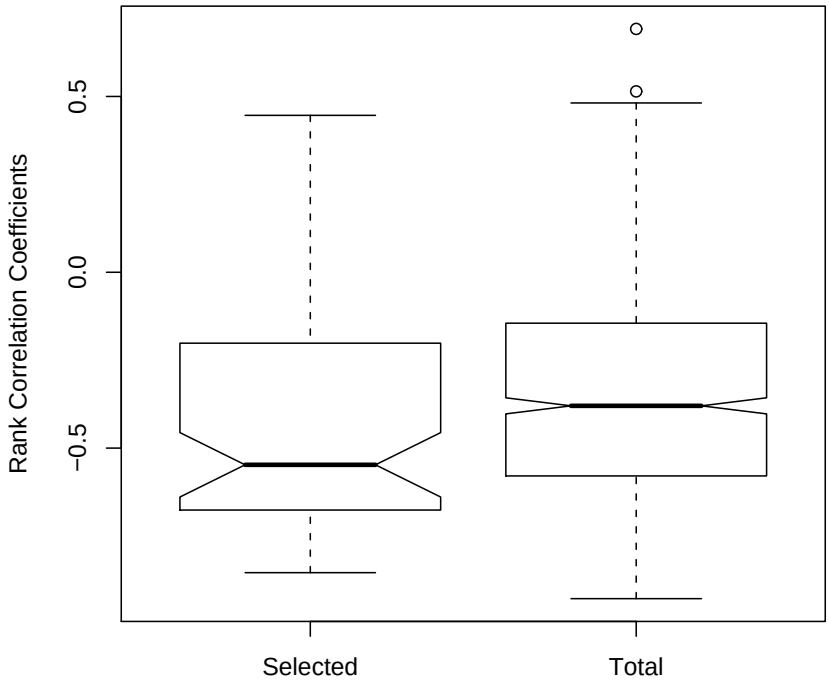


FIGURE E.2: Notched boxplot of the Spearman ρ rank correlation coefficient between gene closeness and the objective score difference from this gene knock-out, grouped in selected for robustness transsys program and the total transsys program population.

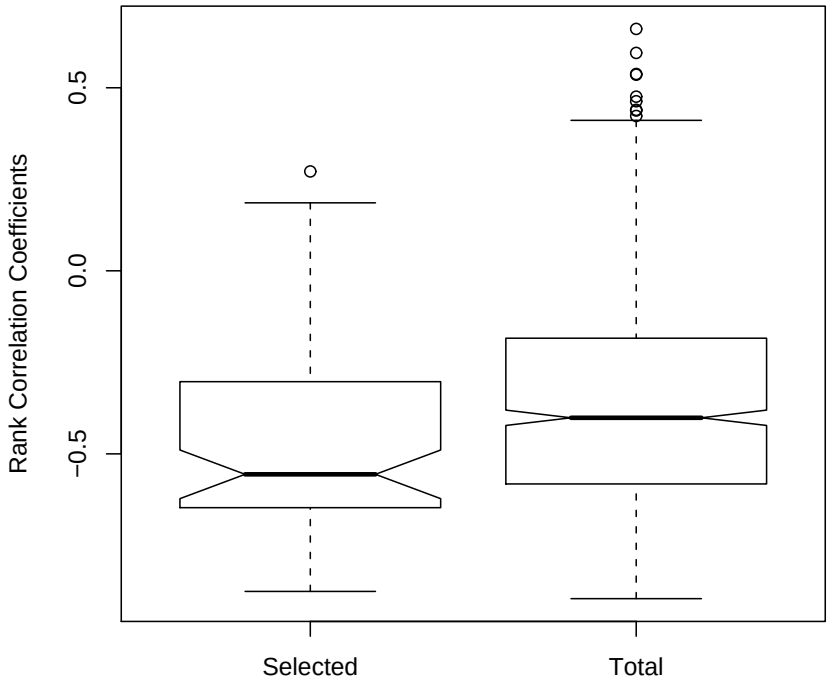


FIGURE E.3: Notched boxplot of the Spearman ρ rank correlation coefficient between gene eigenvector centrality and the objective score difference from this gene knock-out, grouped in selected for robustness transsys program and the total transsys program population.

Glossary

Bistability Is the property of a system to rest in two stable states. In electronics is realised as a Flip-Flop switch. 154

Cellular Differentiation Is the process by which cells acquire a *type*. The morphological features of cells are changing dramatically during differentiation leading to cells that do not share common characteristics and thus belong to different *types*. 154

Centrality A centrality is a function $\mathtt{C}$ that assigns to a vertex $v \in \mathcal{V}$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ a real value $\mathtt{C}(v) \in \mathbb{R}$. 41

Density A directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where loops are allowed, will have density $D_{\mathcal{G}}$ as:

$$D_{\mathcal{G}} = \frac{|\mathcal{E}|}{|\mathcal{V}|^2}$$

. 76

Dynamical Parameters The set of real valued parameters that determine the strength and the nature of the regulatory interactions between genes in GRNs, as well as parameters that determine gne product properties such as the degradation rate and the capability to difuuse. Dynamical parameters are defined in continuous modelling of GRNs and the exact set of dynamical parameters depnds on the particular modelling approach (e.g. Michaelis-Menten parameters, Hills coefficients etc.). 8

Gene Regulatory Network A set of DNA segments (genes) and the set of their interactions. Gene are interacting with each other indirectly (through their gene products, i.e. proteins, RNA), thereby governing the rates at which genes are expressed. 7

Homeostasis The property of biological systems to regulate their internal environment to a stable condition. 7

Multistationarity Multistationarity is the property of systems to exhibit a bistable behaviour (see Bistability) (when one element is ON the other is OFF and vice versa) and stably maintain initial stimuli (like a flip-flop switch in electronics), therefore serving as on bit of memory. Here we explore the connection of multistationarity with Cellular Differentiation. 7

Spatial Correlation As spatial correlation in this work we calculated the Pearson correlation coefficient between the Euclidean distance of factor concentration between all pairs of cells on a collection over the Manhattan distance.. 51

Stripy Lattice A colloquial term introduced in the context of this thesis to describe with one phrase the following phenomenon: The emergence of gene expression heterogeneity in forms of stripes of alternating factor concentration levels in transsys instances along a spatially extended system (i.e. the lattice reactor) and not on a system which lacks spatial organisation (i.e. the well stirred reactor). 69, 72

Topology Topology is a general area of mathematics studying the structure of space and describing how entities are arranged in space. Here we refer solely on Network Topology, which is the study of the arrangement of the elements (links, nodes, etc.) of a network.. 4, 8

References

Réka Albert and Albert-Lázló Barabási. The statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002. 36, 38, 39

Maximino Aldana and Philippe Cluzel. A natural class of robust networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(15):8710–8714, 2003. 37

Eivind Almaas. Biological impacts and context of network theory. *Journal of Experimental Biology*, 210(Pt 9):1548–1558, 2007. 67, 118, 129

Uri Alon. Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007. 41

Elena R Alvarez-Buylla, Mariana Benítez, Enrique Balleza Dávila, Alvaro Chaos, Carlos Espinosa-Soto, and Pablo Padilla-Longoria. Gene regulatory network models for plant development. *Current Opinion in Plant Biology*, 10(1):83–91, 2007. 67

Albert-Lásló Barabási. Taming complexity. *Nature Physics*, 1:68–70, 2005. 36

Albert-László Barabási. *Linked: How everything is connected to everything else and what it means*. Penguin Group New York, 2003. 34

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. 35, 58, 59, 67

Albert-László Barabási and Zoltán N Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. 4, 10

Smadar Ben-Tabou de Leon and Eric H Davidson. Gene regulation: Gene control network in development. *Annual Review of Biophysics and Biomolecular Structure*, 36:191, 2007. 9

Smadar Ben-Tabou de Leon and Eric H Davidson. Modeling the dynamics of transcriptional gene regulatory networks for animal development. *Developmental Biology*, 325(2):317–328, 2009. 5, 7

Mariana Benítez, Carlos Espinosa-Soto, Pablo Padilla-Longoria, and Elena R. Alvarez-Buylla. Interlinked nonlinear subnetworks underlie the formation of robust cellular patterns in arabidopsis epidermis: A dynamic spatial model. *BMC Systems Biology*, 2:98, 2008. 26

Mariana Benítez, Carlos Espinosa-Soto, Pablo Padilla-Longoria, José Díaz, and Elena R Alvarez-Buylla. Equivalent genetic regulatory networks in different contexts recover contrasting spatial cell patterns that resemble those in arabidopsis root and leaf epidermis: A dynamic model. *International Journal of Developmental Biology*, 51(2):139–155, 2007. 20, 26

C. Berding and T. Harbich. On the dynamics of a simple biochemical control circuit. *Biological Cybernetics*, 49(3):209–219, 1984. 18

Franco A. Bignone. Cells-gene interactions simulation on a coupled map lattice. *Journal of Theoretical Biology*, 161(2):231–249, 1993. 24, 26, 46, 127

S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175 – 308, 2006. ISSN 0370-1573. 38

Maatren C. Boerlijst and Pauline Hogeweg. Attractors and spatial patterns in hypercycles with negative interactions. *Journal of Theoretical Biology*, 176(2): 199–210, 1995. ISSN 0022-5193. 9, 11

B. Bollobás. *Random graphs*. Cambridge Univ Pr, 1985. 35

Costas Bouyioukos and Jan T. Kim. Gene regulatory network properties linked to gene expression dynamics in spatially extended systems. In George Kampis, editor, *Advances in Artificial Life (Proccedings of the* 10^{th} *European Conference in Artificial Life)*, number 5777/5778 in LNCS/LNAI. Springer–Verlag, 2009. 33, 76, 96

Dennis Bray. Molecular networks: The top-down view. *Science*, 301(5641):1864–1865, 2003. 4

Álvaro Chaos, Max Aldana, Carlos Espinosa-Soto, Berenice León, Adriana Arroyo, and Elena Alvarez-Buylla. From genes to flower patterns and evolution: Dynamic models of gene regulatory networks. *Journal of Plant Growth Regulation*, 25(4):278–289, 2006. 17

Arturo Chavoya and Yves Duthen. A cell pattern generation model based on an extended artificial regulatory network. *Biosystems*, 94(1-2):95–101, 2008. 27

Joshua L. Cherry and Frederick R. Adler. How to make a biological switch. *Journal of Theoretical Biology*, 203(2):117–133, 2000. 19

Enrico S. Coen and E. M. Meyerowitz. The war of the whorls: Genetic interactions controlling flower development. *Nature*, 353(6339):31–37, 1991. 26, 27

Gábor Csárdi and Tamás Népusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. 72

Luciano F. da Costa, Francisco A. Rodrigues, Gonzalo Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007. 38, 41

Maria I Davidich and Stefan Bornholdt. Boolean network model predicts cell cycle sequence of fission yeast. *PLoS One*, 3(2):e1672, 2008. 17

Eric H Davidson and Douglas H Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(5762):796–800, 2006. 25

Eric H Davidson and Michael S Levine. Properties of developmental gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of AMerica*, 105(51):20063–20066, 2008. 10

Eric H Davidson, David R McClay, and Leroy Hood. Regulatory gene networks and the properties of the developmental process. *Proceedings of the National Academy of Sciences of the United States of America*, 100(4):1475–1480, 2003. 10

Eric H. Davidson, Jonathan P. Rast, Paola Oliveri, Andrew Ransick, Cristina Calestani, Chiou-Hwa Yuh, Takuya Minokawa, Gabriele Amore, Veronica Hinman, Cesar Arenas-Mena, Ochan Otim, C. Titus Brown, Carolina B. Livi, Pei Yun Lee, Roger Revilla, Alistair G. Rust, Zheng jun Pan, Maria J. Schilstra, Peter J. C. Clarke, Maria I. Arnone, Lee Rowen, R. Andrew Cameron, David R. McClay, Leroy Hood, and Hamid Bolouri. A genomic regulatory network for development. *Science*, 295(5560):1669–1678, 2002. 9

Hidde de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9(1):67–103, 2002. 34

Hidde de Jong, Jean-Luc Gouzé, Céline Hernandez, Michel Page, Tewfik Sari, and Johannes Geiselmann. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bullentin of Mathematical Biology*, 66(2):301–340, 2004. 23

Gabriel del Rio, Dirk Koschützki, and Gerardo Coello. How to identify essential genes from molecular networks? *BMC Systems Biology*, 3:102, 2009. 43

Reinhard Diestel. *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer-Verlag, Heidelberg, 3^{rd} edition, 2005. 38, 76

Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25): 14863–14868, 1998. 18

Paul Erdös and Alfred Rényi. On random graphs. *Publicationes Mathematicae*, 6: 290–297, 1959. 35, 38, 58, 67

Carlos Espinosa-Soto, Pablo Padilla-Longoria, and Elena R Alvarez-Buylla. A gene regulatory network model for cell-fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. *Plant Cell*, 16(11):2923–2939, 2004. 11, 26

Giorgio Fagiolo. Clustering in complex directed networks. *Physical Review E Statistical Nonlinear and Soft Matter Physics*, 76(2 Pt 2):026107, 2007. 87

Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262, New York, NY, USA, 1999. ACM. ISBN 1-58113-135-6. 35, 58

Illés J. Farkas, H. Jeong, Tamás Vicsek, Albert-László Barabási, and Zoltan N. Oltvai. The topology of the transcription regulatory network in the yeast, *Saccharomyces Cerevisiae*. *Physica A*, 318(3-4):601–612, 2003. 36

Kurt Fleischer. Investigations with a multicellular developmental model. In Chris Langton and T Shimohara, editors, *Artificial Life V*, pages 229–236. MIT Press, 1996. ISBN: 0262621118. 24

Kurt Fleischer and Alan H. Barr. A simulation testbed for the study of multicellular development: The multiple mechanisms of morphogenesis. In Chris Langton, editor, *Artificial life III*, pages 389–416. Addison-Wesley, 1993. 24

Jeffrey J. Fox and Colin C. Hill. From topology to dynamics in biochemical networks. *Chaos*, 11(4):809–815, 2001. 83

Evelyn Fox Keller. Revisiting "scale-free" networks. *BioEssays*, 27:1060–1068, 2005. 37

Stefanie Fuhrman, Mary Jane Cunningham, Xiling Wen, Gary Zweiger, Seilhamer Jeffrey J., and Roland Somogyi. The application of Shannon entropy in the identification of putative drug targets. *Biosystems*, 55(1-3):5–14, 2000. 49

Nicholas Geard and Janet Wiles. A gene network model for developing cell lineages. *Artificial Life*, 11(3):249–267, 2005. 24

Alfred. Gierer and Hans Meinhardt. A theory of biological pattern formation. *Kybernetik*, 12:30–39, 1972. 11, 24, 129

M. Girvan and Mark E J Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, 2002. 43

Leon Glass and Stuart A. Kauffman. The logical analysis of continuous, non-linear biochemical control networks. *Journal of Theoretical Biology*, 39(1):103–129, 1973. 22

Jing-Dong J Han, Denis Dupuy, Nicolas Bertin, Michael E Cusick, and Marc Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23(7):839–844, 2005. 37

David Harel. A turing-like test for biological modeling. *Nature Biotechnology*, 23 (4):495–496, 2005. 131

L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–C52, 1999. 21

Jeff Hasty, D. McMillen, F. Isaacs, and J. J. Collins. Computational studies of gene regulatory networks: *In numero* molecular biology. *Nature Reviews Genetics*, 2 (4):268–279, 2001. 23

Stefan Hoops, Sven Sahle, Ralph Gauges, Christine Lee, Jrgen Pahle, Natalia Simus, Mudita Singhal, Liang Xu, Pedro Mendes, and Ursula Kummer. COPASI–a COmplex PAthway SImulator. *Bioinformatics*, 22(24):3067–3074, 2006. 21

T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000. 5

Piers J Ingram, Michael P H Stumpf, and Jaroslav Stark. Network motifs: structure does not determine function. *BMC Genomics*, 7:108, 2006. 41

Johannes Jaeger and Alfonso Martinez-Arias. Getting the measure of positional information. *PLoS Biology*, 7(3):e81, 2009. 11

H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000. 36, 58

Jieun Jeong and Piotr Berman. On cycles in the transcription network of saccharomyces cerevisiae. *BMC Systems Biology*, 2:12, 2008. 40

Henrik Jönsson, Marcus Heisler, G. Venugopala Reddy, Vikas Agrawal, Victoria Gor, Bruce E. Shapiro, Eric Mjolsness, and Elliot M. Meyerowitz. Modeling the organization of the wuschel expression domain in the shoot apical meristem. *Bioinformatics*, 21 Suppl 1:i232–i240, 2005. 26

Richard Jovelin and Patrick C Phillips. Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biology*, 10(4):R35, 2009. 92, 93

Björn H Junker, Dirk Koschützki, and Falk Schreiber. Exploration of biological network centralities with CentiBIN. *BMC Bioinformatics*, 7:219, 2006. 43

K. Kappler, R. Edwards, and Leon Glass. Dynamics in high-dimensional model gene networks. *Signal Processing*, 84(4):789–798, 2003. 22

Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Moleular Cell Biology*, 9(10):770–780, 2008. 33

Stuart A. Kauffman. Homeostasis and differentiation in random genetic control networks. *Nature*, 224(5215):177–178, 1969a. 15

Stuart A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22(3):437–467, 1969b. 15, 22, 78

Stuart A Kauffman. The large scale structure and dynamics of gene control circuits: An ensemble approach. *Journal of Theoretical Biology*, 44(1):167–190, 1974. 16

Stuart A. Kauffman. Developmental logic and its evolution. *Bioessays*, 6(2):82–87, 1987. 9, 15

Stuart A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, 1993. 15, 78

Stuart A Kauffman. *At Home in the Universe: The Search for Laws of Self-Organization and Complexity*. Oxford University Press, 1996. 15

Soile V. E. Keränen. Simulation study on effects of signalling network structure on the developmental increase in complexity. *Journal of Theoretical Biology*, 231(1):3–21, 2004. 26, 46, 127

Harold D Kim, Tal Shay, Erin K O'Shea, and Aviv Regev. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science*, 325(5939):429–432, 2009. 131

Jan T. Kim. `transsys`: A generic formalism for modeling regulatory networks in morphogenesis. In Josef Kelemen and Sosík Petr, editors, *Advances in Artificial Life (Proccedings of the 6th European Conference in Artificial Life)*, volume 2159 of *Lecture Notes in Artificial Intelligence*, pages 242–251, Berlin, Heidelberg, 2001. Springer Verlag. 27, 33

Jan T. Kim. Effects of spacial growth on gene expression dynamics and on regulatory networks reconstruction. In M. Capcarrere, A.A. Freitas, P.J. Bentley, C.G. Johnson, and J. Timmis, editors, *Advances in Artificial Life (Proccedings of the 8th European Conference in Artificial Life)*, volume 3630 of *Lecture Notes in Artificial Intelligence*, pages 825–834, Berlin, Heidelberg, 2005. Springer Verlag. ISBN 3-540-28848-1. 33

Jan T. Kim. A rule-based approach to selecting developmental processes. In Peter Dittrich and Stefan Artmann, editors, *Proceedings of the 7th German Workshop on Artificial Life*, pages 63–74, Berlin, 2006. Akademische Verlagsgesellschaft Aka. 8, 33

Jan T. Kim. The transsys home page, 2009. `http://www.transsys.net/`. 33

Junil Kim, Tae-Geon Kim, Sung Hoon Jung, Jeong-Rae Kim, Taesung Park, Pat Heslop-Harrison, and Kwang-Hyun Cho. Evolutionary design principles of modules that control cellular differentiation: consequences for hysteresis and multistationarity. *Bioinformatics*, 24(13):1516–1522, 2008. 40

Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS Computational Biology*, 5(5):e1000385, 2009. 44

Steffen Klamt, Julio Saez-Rodriguez, and Ernst D Gilles. Structural and functional analysis of cellular networks with cellnetanalyzer. *BMC Systems Biology*, 1:2, 2007. 44

Johannes F Knabe, Chrystopher L Nehaniv, and Maria J Schilstra. Do motifs reflect evolved function?–no convergent evolution of genetic regulatory network subgraph topologies. *Biosystems*, 94(1-2):68–74, 2008a. 41

Johannes F. Knabe, Chrystopher L. Nehaniv, and Maria J. Schilstra. Evolution and morphogenesis of differentiated multicellular organisms: autonomously generated diffusion gradients for positional information. In *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 321–328. MIT Press, 2008b. 27

Dirk Koschützki and Falk Schreiber. Comparison of centralities for biological networks. In R. Giegerich and J. Stoye, editors, *Proceedings of the German Conference on Bioinformatics (GCB '04)*, volume P-53 of *Lecture Notes in Informatics*, pages 199–206, 2004. 43, 92

Dirk Koschützki and Falk Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regulation and Systems Biology*, 2:193–201, 2008. 92

Dwight P. Kuo, Wolfgang Banzhaf, and André Leier. Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplicationa nd divergence. *Biosystems*, 85:177–200, 2006. 83

Fangting Li, Tao Long, Ying Lu, Qi Ouyang, and Chao Tang. The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14):4781–4786, 2004. 17

J. C W Locke, A. J. Millar, and M. S. Turner. Modelling genetic networks with noisy and varied experimental data: The circadian clock in *Arabidopsis Thaliana*. *Journal of Theoretical Biology*, 234(3):383–393, 2005. 6

Laurence Loewe. A framework for evolutionary systems biology. *BMC Systems Biology*, 3:27, 2009. 33

William Longabaugh and Hamid Bolouri. Understanding the dynamic behavior of genetic regulatory networks by functional decomposition. *Current Genomics*, 7 (6):333–341, 2006. 8

William J R Longabaugh, Eric H Davidson, and Hamid Bolouri. Computational representation of developmental genetic regulatory networks. *Developmental Biology*, 283(1):1–16, 2005. 25

Nicholas M Luscombe, M. Madan Babu, Haiyuan Yu, Michael Snyder, Sarah A Teichmann, and Mark Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312, 2004. 40

Avi Ma'ayan, Guillermo A Cecchi, John Wagner, A. Ravi Rao, Ravi Iyengar, and Gustavo Stolovitzky. Ordered cyclic motifs contribute to dynamic stability in biological and engineered networks. *Proceedings of the National Academy of Sciences of the United States of America*, 105(49):19235–19240, 2008. 40

Athanasius F Mareé, Alexander V Panfilov, and Paulien Hogeweg. Migration and thermotaxis of dictyostelium discoideum slugs, a model study. *Journal of Theorecital Biology*, 199(3):297–309, 1999. 131

Robert M. May. Network structure and the biology of populations. *Trends in Ecology and Evolution*, 21(7):394–399, 2006. 36

John Maynard Smith. The 1999 crafoord prize lectures. the idea of information in biology. *The Quarterly Review of Biology*, 74(4):395–400, 1999. 49

John Maynard Smith. The concept of information in biology. *Philosophy of Science*, 67(2):177–194, 2000. 49

Harley H. McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 94(3):814–819, 1997. 20

Hans Meinhardt. *Models of biological pattern formation.* Academic Press London, 1982. 11

Hans Meinhardt. From observations to paradigms; the importance of theories and models. an interview with hans meinhardt by richard gordon and lev beloussov. *International Journal of Developmental Biology*, 50(2-3):103–111, 2006. 24

Pedro Mendes. Biochemistry by numbers: Simulation of biochemical pathways with Gepasi 3. *Trends in Biochemical Sciences*, 22(9):361–363, 1997. 21

Pedro Mendes, Wei Sha, and Keying Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19(Suppl.2):ii122–ii129, 2003. 21

Luis Mendoza and Elena R. Alvarez-Buylla. Dynamics of the genetic regulatory network for *Arabidopsis Thaliana* flower morphogenesis. *Journal of Theoretical Biology*, 193(2):307–319, 1998. 9, 17, 45

Luis Mendoza, Denis Thieffry, and Elena R. Alvarez-Buylla. Genetic control of flower morphogenesis in *Arabidopsis Thaliana*: A logical analysis. *Bioinformatics*, 15(7-8):593–606, 1999. 26

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298 (5594):824–827, 2002. 40

Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004. 87

Patrycja Vasilyev Missiuro, Kesheng Liu, Lihua Zou, Brian C Ross, Guoyan Zhao, Jun S Liu, and Hui Ge. Information flow analysis of interactome networks. *PLoS Computational Biology*, 5(4):e1000350, 2009. 94

Eric Mjolsness, D. H. Sharp, and J. Reinitz. A connectionist model of development. *Journal of Theoretical Biology*, 152(4):429–453, 1991. 44

Adilson E. Motter, Manuel A. Matías, Jürgen Kurths, and Edward Ott. Dynamics on complex networks and applications. *Pysica D*, 224(1–2):vii–viii, 2006. 10

Andreea Munteanu and Ricard V Solé. Neutrality and robustness in evo-devo: Emergence of lateral inhibition. *PLoS Computational Biology*, 4(11):e1000226, 2008. 27

Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003. 36, 39

Mark E. J. Newman, Albert-Lászlo. Barabasi, and Duncan J. Watts. *The structure and dynamics of networks*. Princeton Univiveristy Press, 2006. 34

Mark EJ Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005. 43

Paola Oliveri, Qiang Tu, and Eric H Davidson. Global regulatory logic for specification of an embryonic cell lineage. *Proceedings of the National Academy of Sciences of the United States of America*, 105(16):5955–5962, 2008. 67

Stig W. Omholt, Erik Plahte, Leiv Øyehaug, and Kefang Xiang. Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics*, 155:969–980, 2000. 8, 10

R. Pastor-Satorras and A. Vespignani. Epidemic dynamics and endemic states in complex networks. *Physical Review E Stat Nonlin Soft Matter Phys*, 63(6 Pt 2): 066117, 2001. 36

Erik Plahte. Pattern formation in discrete cell lattices. *Journal of Mathematical Biology*, 43(5):411–445, 2001. 46

Robert J Prill, Pablo A Iglesias, and Andre Levchenko. Dynamic properties of network motifs contribute to biological network organization. *PLoS Biology*, 3 (11):e343, 2005. 8, 41

Pyhton Softaware Foundation. The Python programming language, 1996–2010. 75

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0. 75

D. A. Rand, B. V. Shulgin, J. D. Salazar, and A. J. Millar. Uncovering the design principles of circadian clocks: Mathematical analysis of flexibility and evolutionary goals. *Journal of Theoretical Biology*, 238(3):616–635, 2006. 6

J. Reinitz, E. Mjolsness, and D. H. Sharp. Model for cooperative control of positional information in drosophila by bicoid and maternal hunchback. *Journal of Experimental Zoology*, 271(1):47–56, 1995. 45

Dirk Repsilber and Jan T. Kim. Developing and testing methods for microarray data analysis using an artificial life framework. In Wolfgang Banzhaf, Thomas Christaller, Peter Dittrich, Jan T. Kim, and Jens Ziegler, editors, *Advances in Artificial Life (Proccedings of the 7^{th} european conference in Artificial Life)*, pages 686–695, Berlin, Heidelberg, 2003. Springer Verlag. 33

Marcel Salathé, Robert M May, and Sebastian Bonhoeffer. The evolution of network topology by selective removal. *Journal of the Royal Society, Interface*, 2 (5):533–536, 2005. 37

Isaac Salazar-Ciudad, Jordi Garcia-Fernandez, and Ricard V. Sole. Gene networks capable of pattern formation: From induction to reaction-diffusion. *Journal of Theoretical Biology*, 205:587–603, 2000. 26, 129

Isaac Salazar-Ciudad, Stuart A. Newman, and Ricard V. Sol. Phenotypic and dynamical transitions in model genetic networks. I. emergence of patterns and genotype-phenotype relationships. *Evolution & Development*, 3(2):84–94, 2001. 129

Maria J Schilstra and Chrystopher L Nehaniv. Bio-Logic: gene expression and the laws of combinatorial logic. *Artificial Life*, 14(1):121–133, 2008. 16

Claude E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948. 49

Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of *Escherichia Coli*. *Nature Genetics*, 31 (1):64–68, 2002. 41

P. Smolen, D. A. Baxter, and J. H. Byrne. Modeling transcriptional control in gene networks–methods, recent results, and future directions. *Bullentin of Mathematical Biology*, 62(2):247–292, 2000. 23, 33, 34

Ricard V. Solé, I Salazar-Ciudad, and Stuart A. Newman. Gene network dynamics and the evolution of development. *Trends in Ecology and Evolution*, 15(12):479–480, 2000. 24

Christophe Soulé. Graphic requirements for multistationarity. *ComPlexUs*, 1(3): 123–133, 2003. 23, 40

Christophe Soulé. Mathematical approaches to differentiation and gene regulation. *Comptes Rendus Biologies*, 329(1):13–20, 2006. 23, 40

Angelike Stathopoulos and Michael Levine. Genomic regulatory networks and animal development. *Developmental Cell*, 9(4):449–462, 2005. 9

Michael P. H. Stumpf, Carsten Wiuf, and Robert M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12): 4221–4224, 2005. 37

René Thomas. Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42(3):563–585, 1973. 16, 22, 23

René Thomas. Logical analysis of systems comprising feedback loops. *Journal of Theoretical Biology*, 73(4):631–656, 1978. 16, 39

René Thomas. Laws for the dynamics of regulatory networks. *The International Journal of Developmental Biology*, 42(3):479–485, 1998. 16, 39

René Thomas and Richard D'Ari. *Biological Feedback*. CRC, 1990. 15, 87

René Thomas and M. Kaufman. Multistationarity, the basis of cell differentiation and memory. I. structural conditions of multistationarity and other nontrivial behavior. *Chaos*, 11(1):170–179, 2001a. 23

René Thomas and M. Kaufman. Multistationarity, the basis of cell differentiation and memory. II. logical analysis of regulatory networks in terms of feedback circuits. *Chaos*, 11(1):180–195, 2001b. 23

C. Titus Brown, Alistair G Rust, Peter J C Clarke, Zhengjun Pan, Maria J Schilstra, Tristan De Buysscher, Gareth Griffin, Barbara J Wold, R. Andrew Cameron, Eric H Davidson, and Hamid Bolouri. New computational approaches for analysis of cis-regulatory networks. *Developmental Biology*, 246(1):86–102, June 2002. 21

Alan M. Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences.*, 237(641):37–72, 1952. 11, 24, 129

John J. Tyson, K. Chen, and B. Novak. Network dynamics and cell physiology. *Nature Reviews, Molecular Cell Biology*, 2(12):908–916, 2001. 7, 20

John J. Tyson and Hans G. Othmer. The dynamics of feedback control circuits in biochemical pathways. *Progress in Theoretical Biology*, 5:1–62, 1978. 18

Tim van den Bulcke, Koenraad van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart de Moor, and Kathleen Marchal. SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7:43, 2006. 21

George von Dassow, Eli. Meir, Edwin. M. Munro, and Garrett M. Odell. The segment polarity network is a robust developmental module. *Nature*, 406(6792): 188–192, 2000. 25, 27, 130

Andreas Wagner. *Robustness and evolvability in living systems*. Princeton Studies in Complexity. Princeton University Press Princeton, NJ, 2005. ISBN: 978-0-691-13404-8. 44

CC Walker and WR Ashby. On temporal characteristics of behavior in certain complex systems. *Biological Cybernetics*, 3(2):100–108, 1966. 15

Duncan J. Watts. The "new" science of networks. *Annual Review of Sociology*, 30:243–270, 2004a. 34

Duncan J. Watts. *Six degrees: The science of a connected age*. WW Norton & Company, 2004b. 34

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393(6684):440–442, 1998. 35, 36, 39, 86, 118

Sebastian Wernicke and Florian Rasche. FANMOD: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006. 74

Walter Willinger, David Alderson, John C. Doyle, and Lun Li. More "normal" than normal: Scaling distributions and complex systems. In R.G. Ingalls, M.D. Rossetti, J.S. Smith, and Peters B.A., editors, *Proceedings of the 2004 Winter Simulation Conference.*, pages 130–141, 2004. 37

Denise M. Wolf and Frank H. Eeckman. On the relationship between genomic regulatory element organization and gene regulatory dynamics. *Journal of Theoretical Biology*, 195(2):167–186, 1998. 19

Lewis Wolpert. Positional information and the spatial pattern of cellular differentiation. *Journal of Theoretical Biology*, 25(1):1–47, 1969. 27

Jeongah Yoon, Anselm Blumer, and Kyongbum Lee. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics*, 22(24):3106–3108, 2006. 43

Haiyuan Yu, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Computational Biology*, 3(4):e59, 2007. 42

Etay Ziv, Ilya Nemenman, and Chris H Wiggins. Optimal signal processing in small stochastic biochemical networks. *PLoS One*, 2(10):e1077, 2007. 40

Elena Zotenko, Julian Mestre, Dianne P O'Leary, and Teresa M Przytycka. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol*, 4(8):e1000140, 2008. 42