Aalto University
School of Science
Degree Programme of Computer Science and Engineering

Erkka Valo

Prediction of drug effects in gene regulatory networks:

Boolean modeling approach

Master's Thesis
Espoo, September 30, 2013

Supervisor: Professor Harri Lähdesmäki
Instructor: Professor Sampsa Hautaniemi

Aalto University
School of Science
Degree Programme of Computer Science and Engineering

ABSTRACT OF
MASTER'S THESIS

Author: Erkka Valo

Title:
Prediction of drug effects in gene regulatory networks: Boolean modeling approach

Date: September 30, 2013 **Pages:** 56

Professorship: Bioinformatics **Code:** T-61

Supervisor: Professor Harri Lähdesmäki

Instructor: Professor Sampsa Hautaniemi

Gene regulatory networks (GRNs) control the amount and the temporal patterns of gene products, both of which are crucial for the correct functioning of the living cells of an organism. In many diseases, such as cancer, biological processes controlled by GRNs are perturbed. Understanding the functioning of GRNs may lead to a better understanding of the mechanisms behind disease and ultimately to the identification of putative drug targets.

The amount of information on the components of the GRNs and the interactions between them is increasing rapidly. Many modeling approaches have been applied to simulate the behavior of GRNs. Boolean networks give qualitative predictions of the dynamic behavior of the GRNs. They are applicable especially for large GRNs where all the mechanistic details of different reactions are not known.

In this thesis, an analysis framework to predict the effects of drugs in the context of GRNs was developed. A network consisting of genes, drugs and biological processes was constructed based on knowledge in biological databases. The behavior of the network was simulated with Boolean networks. To predict the effect of perturbing the network with a drug, an activation score was developed to estimate the activity of different components of the network before and after the perturbation. The method was applied to triple-negative breast cancer data to search for putative drug targets.

Keywords: gene regulatory networks, cancer drugs, breast cancer, Boolean modelling

Language: English

Aalto-yliopisto
Perustieteiden korkeakoulu

Aalto-yliopisto
Perustieteiden korkeakoulu

Aalto-yliopisto
Perustieteiden korkeakoulu

DIPLOMITYÖN TIIVISTELMÄ

Tietotekniikan tutkinto-ohjelma

Tekijä:	Erkka Valo		
Työn nimi:	Lääkeainevaikututusten arviointi geenisäätelyverkoissa: Boolean mallinnus		
Päiväys:	30. syyskuuta 2013	**Sivumäärä:**	56
Professuuri:	Bioinformatiikka	**Koodi:**	T-61
Valvoja:	Professori Harri Lähdesmäki		
Ohjaaja:	Professori Sampsa Hautaniemi		

Geenisäätelyverkot (GSV) kontrolloivat geenituotteiden määrää ja ajallista ilmentymistä, jotka ovat ratkaisevassa roolissa eliön solujen virheettömässä toiminnassa. Monissa sairauksissa, kuten syövässä, GSV:n kontrolloimat biologiset prosessit ovat häiriintyneet. GSV:n toiminnan ymmärtäminen voi johtaa sairauksien takana piilevien mekanismien parempaan ymmärtämiseen ja lopulta mahdollisten lääkeainekohteiden tunnistamiseen.

Tieto GSV:n osista ja niiden välisistä vuorovaikutuksista kasvaa nopeasti. Monia eri mallinnusmenetelmiä on sovellettu GSV:hin. Boolean verkot antavat laadullisia ennustuksia GSV:n dynaamisesta käyttäytymisestä. Ne soveltuvat erityisesti isojen GSV:n mallintamiseen ja tilanteisiin, joissa kaikkia yksityiskohtia eri vuorovaikutuksista ei tunneta.

Tässä työssä toteutettiin analyysikehikko, jolla voidaan ennustaa lääkeaineiden vaikusta GSV:n kontekstissa. Geeneistä, lääkeaineista ja biologisista prosesseista koostuva verkko luotiin biologisten tietokantojen sisältämän tiedon perusteella. Verkon käyttäytymistä simuloitiin Boolean verkoilla. Lääkeaineen vaikutuksen määrittämiseksi kehitettiin aktivaatiomitta, jolla arvioidaan eri osien aktiivisuutta ennen ja jälkeen lääkeaineen lisäämistä verkkoon. Menetelmää sovellettiin triplanegatiiviseen rintasyöpädataan mahdollisten lääkeainekohteiden selvittämiseksi.

Asiasanat:	geenisäätelyverkot, syöpälääkkeet, rintasyöpä, Boolean mallinnus
Kieli:	Englanti

Acknowledgements

I wish to thank the teachers and the personnel of the Master's Programme in Bioinformatics in the Aalto University as well as in the University of Helsinki. The interdisciplinary study environment of the master's programme was as exciting as rewarding. Päivi Koivunen and Satu Kähäri I want to thank for guiding me through the intricacies of bureaucracy at times when I was not following the standard path of studies. Further, I extend my gratitude to Harri Lähdesmäki for acting as my thesis supervisor.

The work was done at the Faculty of Medicine of the University of Helsinki in the lab of Sampsa Hautaniemi who I want to thank for providing an exciting and supporting working environment and also for the guidance and the ideas he provided as my supervisor. I want to thank my lab colleges for the enjoyable working environment they have all helped to create. Riku Louhimo I thank for providing the TCGA expression data and being very helpful regarding any questions I had on the TCGA data or the use of Moksiskaan. I want to thank Marko Laakso for the invaluable help with Moksiskaan, Javier Núñez-Fontarnau for helping with mathematical formulation of graph theoretical concepts and Chengyu Liu for getting me started in using Moksiskaan. I want to thank the afternoon lunch gang, Alejandra Cervera Taboada, Tiia Pelkonen, Antonio Neme Castillo, Chiara Facciotto and Amjad Alkodsi for the invigorating discussions not forgetting the excellent coffee. Special thanks to Tiia Pelkonen for proofreading the text. I also want to thank Ville Rantanen and Vladimir Rogojin for their willingness to help in any issues.

I want to thank Aaltonen lab and especially Lauri Aaltonen for providing

me a opportunity to play floorball with them, which I enjoyed tremendously. I want to thank Tatiana Cajuso Pons and Magdalena Schlager for the refreshing discussions we had on the balcony.

I want to thank my family and friends for their love and support.

Helsinki, September 30, 2013

Erkka Valo

Abbreviations and Acronyms

ATO	Arsenic trioxide
DNA	Deoxyribonucleic acid
DEG	Differentially expressed gene
ER	Estrogen receptor
FDR	False discovery rate
FDA	Food and Drug Administration
GO	Gene Ontology
GRN	Gene regulatory network
KEGG	Kyoto Encyclopedia of Genes and Genomes
LSS	Logical steady state
mRNA	messenger RNA
PR	Progesterone receptor
RNA	Ribonucleic acid
TNBC	Triple-negative breast cancer
TCGA	The Cancer Genome Atlas
TF	Transcription factor
TFBS	Transcription factor binding site

Contents

Chapter 1

Introduction

A Gene regulatory network (GRN) consists of a set of molecular species and their interactions which together control the amount and temporal patterns of gene products in a cell [1]. The correct functioning of GRNs is essential to carry out the relevant processes of living cells, including control of cell cycle, cell metabolism and signal transduction [1]. Understanding the functioning of GRNs will assist in understanding the mechanism of diseases where these cellular processes are affected [1].

The vast amount of experimental data and complex structure of the networks make computational tools essential for the analysis of GRNs [1]. The models applied to simulate GRNs can be divided into three categories [1]. Logical models give a dynamic qualitative description of the behavior of the system under study [1]. Continuous models on the other hand give a more detailed view of the concentrations of different species and their temporal development [1]. Lastly, single-molecule level models take into account the stochastic nature of the molecular interactions on the single-molecule level[1].

Boolean networks belong to the class of logical models [1]. Genes are modeled as active or inactive, and the interactions between the genes are represented with Boolean functions [2]. The qualitative dynamic behavior of a GRN of hundreds of components can be simulated using Boolean networks [2]. Boolean networks do not require information on the kinetic parameters

or mechanistic details of different reactions and are well suited to model systems where this information is insufficient[2].

The aim of this thesis is to develop an analysis framework to predict drug effects in the context of GRNs. The GRN is constructed utilizing information in biological databases. The network is expanded to include Gene Ontology (GO) terms representing different biological processes affected by the genes in the network and drugs targeting the genes in the network. The vertices of the network represent the different biological entities, namely, genes, biological processes and drugs. The edges between the vertices describe the regulatory interactions between the different biological entities.

Experimental transcriptomics data, such as gene expression or RNA-seq data, is used to preserve only those interactions fetched from the databases that are relevant for the specific biological system in question. The network is transformed into a Boolean network, and the resulting Boolean network is simulated using individual drugs to perturb the state of the network one by one. An activation score is developed to estimate the effect of a drug on different biological processes included in the network. The activation score is calculated for each gene and GO term based on the states of the genes in the network before and after perturbation with a drug. The difference of the activation scores is used to predict the effect of a drug on genes and different biological processes.

The analysis framework is applied to suggest potential drug targets in triple negative breast cancer (TNBC) patients. Cancer is a disease that affects millions of people world wide every year [3]. In cancer the normal cells become malignant and proliferate without control [4]. The capabilities required from cancer cells are orchestrated through changes in gene expression [4]. The identification of putative cancer drug targets is an important task to facilitate medical research and may eventually lead to the improvement of patient survival and quality of life.

The following Chapter 2 introduces the basic biological background of GRNs and gives some insight into cancer biology. Also, the principles of

Boolean networks and modeling GRNs using them are presented. Finally, Anduril and BoolNet are described. Anduril is a workflow engine that facilitates development of complex analysis pipelines and BoolNet is a software package dedicated to the simulation and analysis of Boolean networks. Chapter 3 gives a general view of constructing a drug effect analysis pipeline using a simple example network. Chapter 4 presents a case study of predicting drug effects in the context of TNBC patients. The network is constructed using biological databases, and transcriptomics data from TNBC patients is employed to fit the network to the biological context in question. Chapter 5 discusses the reached conclusions and the limitations as well as the possible improvements of the analysis framework.

Chapter 2

Background

This chapter gives an introduction to the biological background of gene regulatory networks and modeling them with a Boolean networks. The basics of cancer and specifically TNBC are presented. Additionally, the main tools and databases utilized to create the drug effect prediction analysis framework are introduced.

2.1 Basics of molecular biology

The majority of genetic information of an organism is stored inside the nucleus of the cells of the organism [5]. The information is stored in chromosomes which are structures composed of proteins and a deoxyribonucleic acid (DNA) molecules [5]. DNA is composed of a sugar-phosphate backbone and four bases, namely adenine (A), cytosine (C), guanine (G) and thymine (T) [5]. The specific sequence of the bases encodes information [5]. Genes are stretches of DNA along the chromosome encoding for functional ribonucleic acid (RNA) molecules [5].

The genes function as templates for creating a functional RNA molecule [5]. The RNA molecules can be divided into two classes [5]. The coding RNAs contain sequences which encode the polypeptide sequence of a protein [5]. Most of the genes encode coding RNA molecules, which are subsequently used

as templates to construct a protein [5]. The other class of RNA molecules is non-coding RNA [5]. The non-coding RNA molecules are often involved in the regulation of the expression of other genes [5].

The process of creating an RNA molecule from the DNA sequence of a gene is called transcription [5]. For the class of coding RNA molecules, the RNA is called a messenger RNA (mRNA), since it carries information from DNA to the protein synthesis machinery [5]. The protein synthesis machinery is located outside the nucleus in the cytoplasm [5]. The mRNA is transported to the cytoplasm where it is used as a template to produce a protein in a process named translation [5].

2.1.1 Gene regulatory networks

A subset of all the genes in an organism encode for proteins known as transcription factors (TFs) [6]. TFs bind DNA to the regulatory regions of genes on specific patterns of DNA called transcription factor binding sites (TFBS) and regulate the rate of transcription either positively (activators) or negatively (repressors) [6]. The regulation of the transcription rate of a single gene is generally achieved by the interplay of multiple activators and repressors [7]. This is a simplified view of transcriptional regulation of genes, which in addition to TFs also involves many other types of proteins, such as chromatin-modifying factors [7]. Also, non-coding RNAs are known to regulate the expression of other genes [5].

All the transcriptional regulation interactions between genes form a complex transcriptional regulation network [8], which is referred to as a gene regulatory network in this thesis. A toy example of a gene regulatory network is shown in figure 2.1. Here the gene regulatory network consists of three genes of which all are TFs regulating the transcriptional rate of other genes. For example, gene A regulates the transcription of itself and gene C.

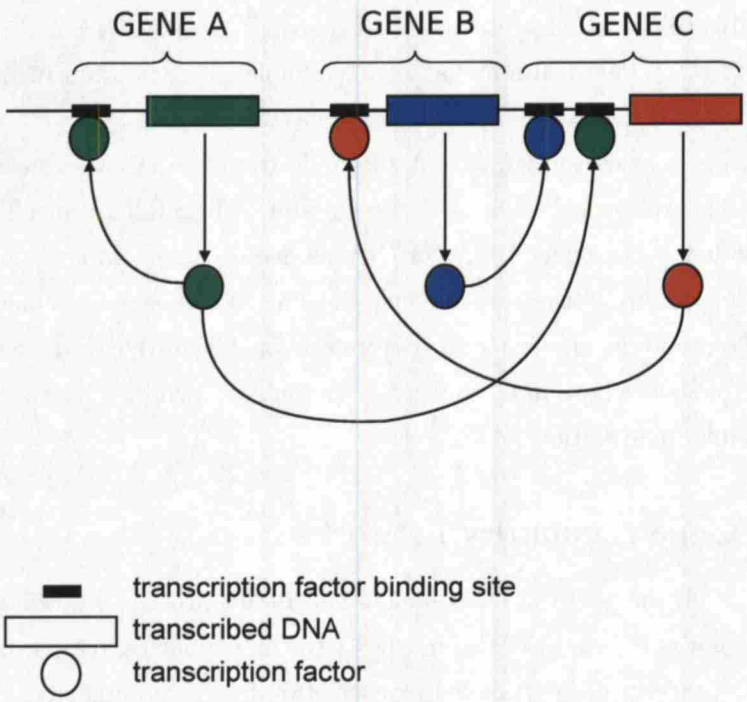


Figure 2.1: An example of a toy gene regulatory network. A gene regulates another gene, if the transcription factor binds to the left side of the transcribed DNA of the gene.

2.2 Cancer

In 2008, there were an estimated 12.4 million new cancer cases and 7.6 million cancer deaths [3]. The number of cancer cases and deaths in 2030 are projected to be 20.0 million and 12.9, respectively [3]. The most common cancers in the world in 2008 in terms of the number of new cases were lung, breast and colorectal cancer [3].

In cancer, the normal cells of an organism proliferate without normal control giving rise to tumors [9]. Almost all deaths caused by cancer follow from a malignant tumor that invades the surrounding tissues and gives rise to distant tumors in other parts of the organism [9].

The development of normal cells into malignant tumor cells proceeds in steps where the cells acquire capabilities required for the malignant phenotype [4]. The capabilities needed, the hallmarks of cancer, are sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis and activating invasion and metastasis [10]. Recently, reprogramming energy metabolism and evading immune destruction have also appeared as potential hallmarks of cancer [4]. The main driver behind the acquisition of these hallmark capabilities is genomic instability in cancer cells [4]. Genomic instability leads to accumulation of changes in the genome of the cancer cells, which enable the emergence of hallmark capabilities [4].

2.2.1 Triple-negative breast cancer

Invasive breast carcinomas are cancers arising from the mammary epithelium [11]. They are characterized by invasiveness to surrounding tissue and tendency to form metastases in distant sites [11]. Breast cancer is the most common type of cancer in women accounting for approximately one quarter of all the cancers in women [11]. In the high risk areas of Europe, North America and Australia, six percent of women develop invasive breast carcinoma before the age of 75 [11]. However, in the low risk areas, only two percent of women develop invasive breast carcinoma before the age of 75 [11].

Triple negative breast cancer (TNBC) is a type of breast cancer defined as lacking expression of estrogen receptor (ER), progesterone receptor (PR) and HER2 in the tumor [12]. Approximately 12% to 17% of women with breast cancer have TNBC [12]. The patients with TNBC have as a group relatively poor outcome and are mainly treated with chemotherapy [12].

2.3 Modeling gene regulatory networks

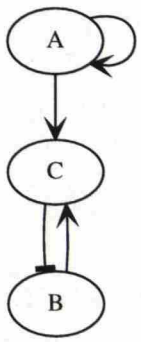
Gene regulatory networks are essential for the correct functioning of living cells [1]. By understanding the functioning of these networks, the mechanisms

behind diseases caused by dysfunctioning gene regulatory networks can be elucidated [1]. Gene regulatory network is the collection of molecular species and the interactions between the species that control the amount of gene products in a cell [1].

Different approaches to modeling gene networks can be divided into logical, continuous and single-molecule level models [1]. Logical models represent the levels of different entities of the system as discrete values. The levels of the entities are updated at each time step according to regulation functions. Logical models include Boolean networks, probabilistic Boolean networks and Petri nets [1]. In continuous models, the levels of the entities are real-valued and are modeled over a continuous timescale [1]. Continuous models include continuous linear models and ordinary differential equations [1]. Single-molecule level models take into account the fluctuations that occur on the molecular level. Single-molecule level models have been implemented using Gillespie's stochastic simulation algorithm and approximation to it [1].

2.3.1 Modeling gene regulatory networks with Boolean networks

Boolean networks consist of vertices with binary states [2]. The states of the nodes are determined by the states of the other nodes through a Boolean function [2]. Boolean networks were first introduced for the analysis of genetic regulatory systems by Kauffman [13]. Boolean networks can be used to simulate the qualitative behavior of the system over time and to predict the effect of perturbations on the system [2]. Unlike continuous models, Boolean networks do not require the kinetic parameters of the interactions of different species [2]. Boolean networks are able to provide qualitative description of the dynamic behavior of the system without the kinetic parameters even for systems with hundreds of species [2].



(a) Network structure

$$f_A = x_A$$
$$f_B = NOT\ x_C$$
$$f_C = x_A\ AND\ x_B$$

(b) Boolean functions

Figure 2.2: (a) The directed graph representation of the example GRN in figure 2.1. The activating interactions are shown with arrowhead edges and the inhibitory interaction with a tee-head edge. (b) The Boolean functions associated to the Boolean network model of the example GRN in figure 2.1.

2.3.1.1 Boolean networks

A Boolean network consists of a set of Boolean variables $\{x_1, x_2, ..., x_n\}$ and a set of Boolean functions $f = \{f_1, f_2, ..., f_n\}$ [14]. The Boolean functions determine the state of the corresponding Boolean variable at time t as a function of the Boolean variables in the network [14].

The Boolean variables corresponding to different biological entities are binary valued. They take values 0 or 1 corresponding to the logical values of $FALSE$ and $TRUE$. The value 0 is referred to as inactive and the value 1 as active [15]. For example, a gene is active when it is expressed and inactive when it is not expressed.

The Boolean functions are mappings $f : \{0,1\}^k \rightarrow \{0,1\}$, where k is the number of input variables for the Boolean function [2]. The Boolean function determines how to calculate the output value from the input values using logical operators, which include AND, OR and NOT [2]. For example, if species i activates species j, x_i is present in the Boolean function f_j. Inhibitory interaction is modeled with the NOT operation. The Boolean

functions can also be defined by truth tables, which list the output values for all the combinations of the input values [2].

A Boolean network can be projected into a directed graph [2]. Directed graph $G = (V, E)$ consists of a set of vertices $V = \{v_1, v_2, ..., v_n\}$ and a set of directed edges E. The vertices correspond to the Boolean variables, and the edges are defined by the Boolean functions [2]. The vertices corresponding to the input variables of Boolean function f_i have an edge incident on v_i. The edges have a sign implying whether the input vertex has a positive or negative effect on the vertex v_i [2]. The directed graph does not define the Boolean network completely since the Boolean functions are only partially determined by the directed graph [2].

Figure 2.2a shows a graph representation of the example GRN in figure 2.1. All the regulatory interactions between the genes are assumed to be activating except C is taken to inhibit B. The activating interactions are shown with arrowhead edges and the inhibitory interaction with a tee-head edge. The Boolean functions of the network are shown in figure 2.2b, and the corresponding truth tables are shown in table 2.1. Here, it is assumed that both A and B have to be active in order C to be active.

B_A	x_A
0	0
1	1

(a)

B_B	x_C
1	0
0	1

(b)

B_C	x_A	x_B
0	0	0
0	0	1
0	1	0
1	1	1

(c)

Table 2.1: The truth tables of the Boolean functions of the example GRN. The left most column gives the output value of the Boolean function with the input value combination on the same row.

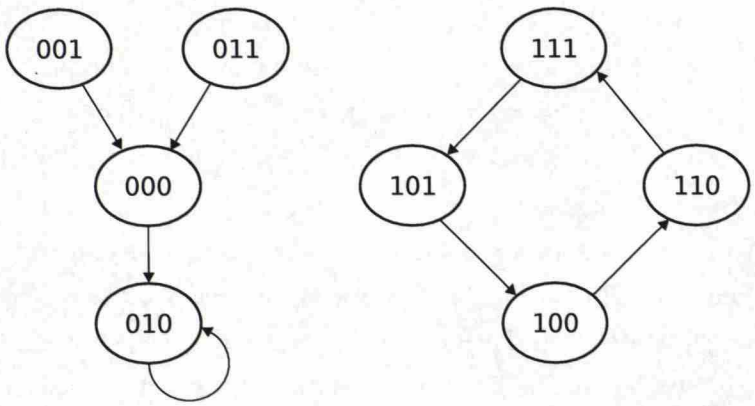


Figure 2.3: The state transition graph of the example gene regulatory network. The vertices represent states of the system and the directed edges allowed transitions. The state of the system is given in the order gene A, gene B and gene C. The allowed transitions are calculated using synchronous updating. State 010 is a point attractor, and the states $100, 101, 110, 111$ form a complex attractor.

2.3.1.2 Simulating Boolean networks

The state of the system at time point t is $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_n(t)]$, and the state space of the system consists of 2^n states [15]. The dynamics of the Boolean network is given by the equation $x_i(t+1) = f_i(t)$ [15]. Here, all the states are updated simultaneously based on the state of the system at time t. In this case, the Boolean network is called synchronous and the behavior of the network is deterministic [15]. The synchronous update scheme implies that all the regulatory interactions in the system have the same timescale [2]. In asynchronous updating scheme, the states are updated based on the previous or current states of the input variables [2]. The asynchronous scheme can be deterministic, where individual states are updated according to the relevant timescales of the biological interactions, or it can be stochastic, where the states are updated in a random order [2].

The sequence of states the system travels through in the state space is

called a trajectory [15]. The state of a synchronous Boolean network will follow a deterministic trajectory in the state space [15]. The possible trajectories can be visualized with a state transition graph, where the vertices represent the states of the system and edges allowed transitions between the different states [2]. The system will eventually end up in a recurring state, called a point attractor or a steady state, or in a recurring state cycle [15], called a complex attractor [2, 15]. The states that lead to the attractor are called transient states [15]. The attractor itself and the transient states that lead to it are called the basin of attraction for the attractor [15]. For the asynchronous Boolean network the point attractors are the same as for the synchronous Boolean network [2]. The complex attractors can be different for deterministic and stochastic Boolean networks [2]. For stochastic asynchronous Boolean networks the system can oscillate randomly in a set of states forming a complex attractor also referred to as loose attractor [2].

The state transition graph of the example gene regulatory network using synchronous updating is shown in figure 2.3. The states are represented as vertices and the allowed transitions as directed edges. For example, if the system is in state 000 at time point t, at time point $t+1$ gene B will be activated and genes A and C will remain inactive and the system will be in state 010. The system has two attractors: 010 is a point attractor, and the states $100, 101, 110$ and 111 form a complex attractor. The size of the basin of attraction is four states for both of the attractors.

For small Boolean networks the point attractors can be solved analytically [2]. For example, all the possible point attractors satisfy $f_i(x_1, ..., x_n) = x_i$, $i = 1, ..., n$, where f_i are the Boolean functions of the network and n is the number of vertices in the network. For example, the point attractors of the example GRN can be obtained by solving the equation system

$$\begin{aligned} x_A &= x_A \\ x_B &= NOT\ x_C \\ x_C &= x_A\ AND\ x_B \end{aligned}$$

The logical steady state (LSS) analysis can find partial steady states of a Boolean network [16]. The fixed initial values of a set of Boolean variables of the network are propagated through the network to identify the variables with fixed states [16]. The resulting set of variables with fixed states is called a partial LSS [16]. Complex attractors can be identified for small synchronous Boolean networks by analytical methods but for asynchronous Boolean networks the identification of loose attractors analytically is a difficult task [2].

The state transition graph of the Boolean network can also be utilized to identify attractors [2]. Point attractors are vertices without outgoing edges except to itself [2]. In figure 2.3, the point attractor 010 has only an outgoing edge to itself. In synchronous or deterministic asynchronous Boolean networks, complex attractors are sets of vertices forming a cycle without outgoing edges [2]. The compex attractor in figure 2.3 consists of a cycle of states $100, 101, 110$ and 111. In stochastic asynchronous Boolean networks, complex attractors are a set of vertices forming a strongly connected component without outgoing edges [2].

For large Boolean networks identification of attractors for both synchronous and asynchronous Boolean networks is a computationally challenging task [2]. The task can be made easier by simplifying the Boolean network prior to the search of attractors [2]. Several network reduction methods have been suggested to this end [2].

2.4 Biological networks

Cells contain a myriad of complex molecules that together carry out the functions necessary for the correct functioning of the cell. The cellular components and the interactions between them can be at the most basic level of abstraction represented as networks [17]. The vertices of the graphs represent the different biological molecules and the edges between the vertices different interactions between the molecules [17]. The network model can be used as a basis for mathematical modeling of the functioning of the network

[18]. The simulation of the network behavior can be used to understand the mechanisms of complex diseases or predict potential drug targets [18].

2.4.1 Moksiskaan

Moksiskaan is a tool to translate gene or protein lists into hypothetical pathways. The pathways are constructed by integrating data from various online databases [19]. In Moksiskaan, biological concepts such as genes, proteins, drugs, biological processes, molecular function and cellular components are called bioentities. Each bioentity has its own bioentity type. The relationships between bioentities are represented as directed edges between the bioentities. Each edge has a link type specifying the type of relationship between the bioentities [20]. The relationships include for example *gene expression* and *gene repression* between a gene and a gene, *drug inhibits* and *drug promotes* between a drug and a gene, and *positive regulation* and *negative regulation* between a gene and a biological process [21].

Moksiskaan can be used to produce hypothetical pathways consisting of defined bioentity types. From a given list of bioentities the pathway is expanded following defined link types between the bioentities. There are four modes for the pathway construction: 'up' bioentities upstream are searched, 'down' bioentities downstream are searched, 'both' bioentities both upstream and downstream are searched and 'connected' bioentities between the given list of bioentities are searched. All these modes are parametrized by the number of steps to search upstream, downstream or between the given list of bioentities[19].

Moksiskaan can also be used to prune the network using experimental data. The bioentities, such as genes, are assigned a state based on the experimental data. The edges conflicting with the experimental data are removed. Also, the state of genes for which there is no experimental data can be predicted using Moksiskaan[21].

Figure 2.4 shows an example of a Moksiskaan generated network. The network was expanded from a gene list: $SOS1$ and $MAP2K1$ one step up-

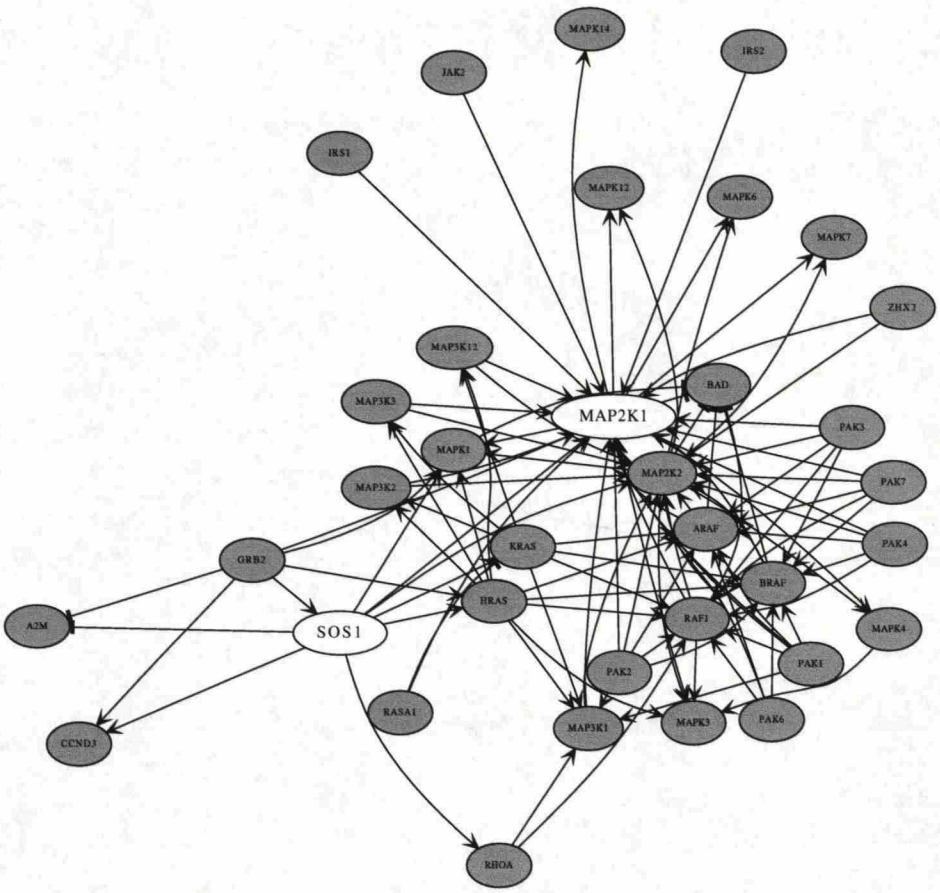


Figure 2.4: An example of a Moksiskaan generated gene regulatory network. The network was expanded one step upstream and one step downstream from genes $SOS1$ and $MAP2K1$, marked with white background color, following edges representing gene expression and gene repression. The arrowhead edges represent regulatory interactions where the source gene activates the gene expression of the target gene, and the tee-head edges represent regulatory interactions where the source gene inhibits the gene expression of the target gene.

stream and one step downstream following edges representing gene expression and repression.

2.5 Biological knowledge bases

Moksiskaan integrates information from various biological knowledge bases [19]. This section summarized the knowledge bases and their relevant content utilized in this work.

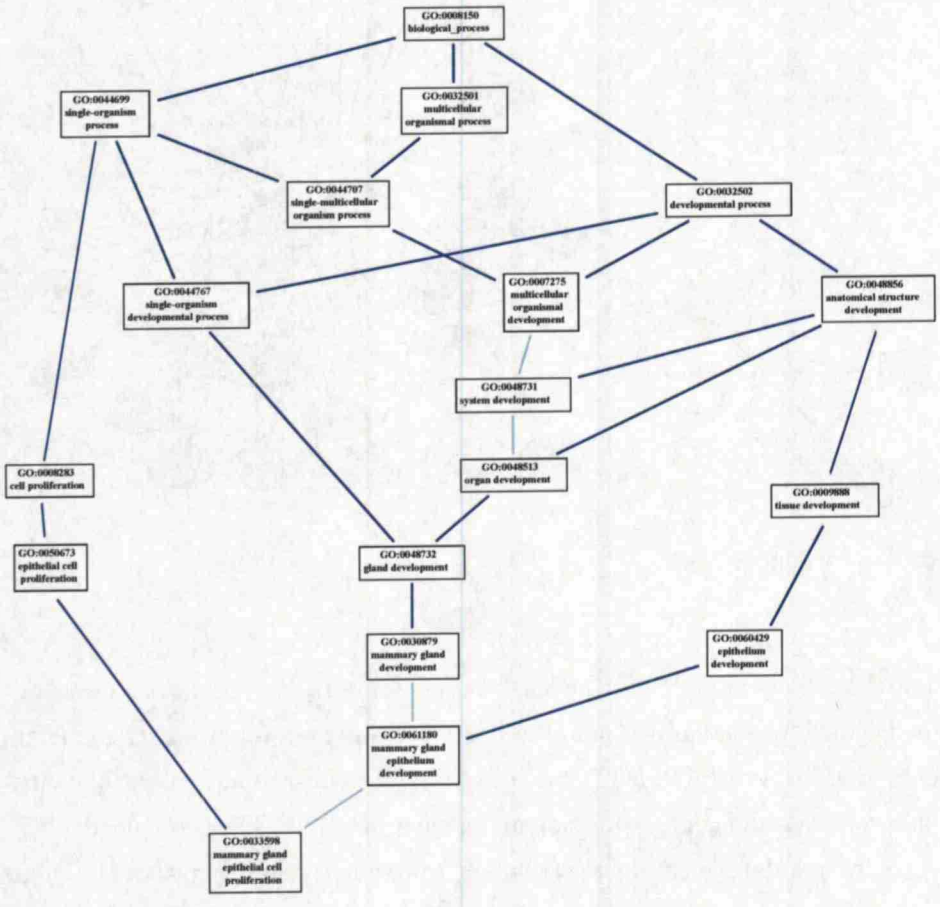


Figure 2.5: The ancestors of the GO term *GO:0033598: mammary gland epithelial cell proliferation* from the biological process ontology. The dark blue lines depict an *is a* relation and the light blue lines a *part of* relation.

2.5.1 Gene Ontology

Gene Ontology is a controlled vocabulary to describe genes and gene products [22]. It is divided into three separate ontologies: biological process, molecular function and cellular component [22]. Biological process refers to a biological objective the gene or the gene products are involved in achieving [22]. Molecular function is a biochemical activity of a gene product. Cellular component is the location where a gene product is active. The terms are linked to their parent and child terms with different relationships forming a hierarchical structure [22]. In figure 2.5 the ancestors of GO term *GO:0033598 mammary gland epithelial cell proliferation* are shown. The dark blue lines and the light blue lines mark *is a* and *part of* relations between the terms, respectively.

The structure of the GO is not static and it is updated frequently as new information is gathered [22]. As of now, there are 25786, 10482 and 3348 GO terms for the biological process, molecular function and cellular component ontologies, respectively [23]. For the same ontologies there are 188386, 103887 and 115817 associations for the human gene products [23].

2.5.2 DrugBank

DrugBank is an open access drug database containing information on drugs and drug targets [24]. The information includes nomenclature, function, action and the targets on which these drugs act upon. There are 6811 drug entries in the database including 1528 FDA-approved small molecule drugs, 150 FDA-approved protein/peptide drugs, 87 nutraceuticals and 5080 experimental drugs [25].

2.5.3 KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is an integrated database resource of 15 main databases, including a pathway database KEGG PATHWAY [26]. The KEGG pathway maps represent molecular interaction

and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development [26]. The pathways are based on published literature and curated in-house [26].

2.5.4 WikiPathways

WikiPathways is a community curated resource for biological pathways [27]. Pathways are represented as pathway diagrams and contain entities such as genes, proteins and metabolites. Currently there are 1729 pathways in the database [28]. WikiPathways offers a web service for programmatic access to the pathway information [29].

2.5.5 PathwayCommons

PathwayCommons is a freely available pathway database which integrates pathway information from various sources [30]. The data is stored in BioPax format and includes proteins, DNA, RNA, complexes, their cellular locations and different physical interactions [30]. The data can be browsed through a web-based interface, downloaded or accessed programmatically through a web service [30]. Currently PathwayCommons contains 1668 pathways comprising 86282 physical entities and 442182 interactions [31].

2.6 Anduril

Anduril [32] is a workflow engine designed for large-scale integrative data analysis. A workflow consists of interconnected processing steps. Each of the processing steps implements a well defined part of the analysis such as data import, data preprocessing or result visualization [32].

The processing steps are implemented as components. Components are reusable executable code, which can be written in any programming language such as Java, R, Python or Matlab. The components have well defined input

and output ports, which are represented as files or directory structures. The input and output ports of the components are connected to each other to form a workflow. The workflow is created using a script language called AndurilScript [33].

2.7 BoolNet

BoolNet is an R package that provides tools to analyze synchronous, asynchronous and probabilistic Boolean networks [34]. The Boolean networks can be defined using a collection of Boolean functions read from a file, reconstructing the network from time series gene expression data or importing from BioTapestry [34]. The Boolean network of the example GRN defined in figure 2.2b would be written in BoolNet format as:

```
targets, factors
A, A
B, ! C
C, B & A
```

The targets column lists the variables in the Boolean network, and the factors column determines the Boolean function of the corresponding variable [35]. Logical operators AND, OR and NOT are coded with &, | and !, respectively [35].

The Boolean networks can be analyzed with various methods, including identification of attractors [34]. Attractors can be searcher for synchronous or asynchronous Boolean networks [35]. For synchronous Boolean networks the steady state and complex attractors can be searched for using an exhaustive search or a heuristic search [35]. The exhaustive search starts from every possible initial state of the network and calculates the transitions in the state space until an attractor is reached [35]. This is plausible only with small networks consisting of maximum 29 genes [35]. The heuristic search starts from a limited number of initial states of the network and identifies

attractors to which the initial states lead [35]. For asynchronous networks the attractors are searched for with a heuristic method which starts from a subset of possible initial states and performs a number of random transitions to reach the attractors [35].

The state of a gene can be fixed to active or inactive without modifying the Boolean functions to simulate overexpression or knock-out experiments, respectively [34]. Computationally expensive algorithms have been implemented in ANSI C to ensure high performance [34].

Chapter 3

Drug effect analysis pipeline

This chapter describes a general approach to assess the effects of individual drugs on a biological system. To this end, a network consisting of different biological entities (bioentities), specifically genes, drugs and GO terms representing biological processes and molecular functions, is constructed. The regulatory interactions between the bioentities are also included in the network. Experimental data is used to prune the network, that is to say, regulatory interactions between the bioentities not supported by the data are removed. Also, the state of the genes in the network are predicted using experimental data. The effect of perturbing the network with a drug is computed and the biological effectes of the drug predicted. The analysis workflow is constructed using Anduril workflow engine. The workflow can be divided into four main stages:

1. Construction of a candidate network using biological databases.
2. Constructing a pruned network with initial states for the genes from the candidate network using experimental data.
3. Simulating the effects of a drug on the state of the network.
4. Estimating the biological effects of the drug based on the perturbed and the initial state of the network.

First, Moksiskaan is employed to construct a candidate network based on existing databases of biological knowledge. The network consists of vertices representing the different bioentities and edges between the vertices representing the regulatory interactions between the bioentities. Second, the network is pruned using transcriptomics data from TNBC patients [36]. Genes and edges between them are removed, if they are not supported by the experimental data. Also, the state of the genes with unknown state is predicted using the data and the network structure. Third, the effect of perturbing the network with a drug is simulated with a Boolean modeling framework implemented as a BoolNet R package. The attractor of the network is computed after perturbation and the steady states of the genes are determined. Finally, the information of the states of the genes before and after perturbation are used to calculate activity scores to estimate the change of activity for the genes and the GO terms following the perturbation. The difference between the scores is used to estimate the biological effects of the drugs on the biological system.

3.1 Construction of the candidate network

In order to construct a candidate network, Moksiskaan is used to retrieve information from the following databases: GO, KEGG, WikiPathways, PathwayCommons and DrugBank. The gene regulatory network is expanded from a set of candidate genes utilizing gene interaction data from KEGG, WikiPathways and PathwayCommons. GO terms are added to the network by employing data from the GO database, which provides information on how different genes regulate different biological processes and molecular functions represented by GO terms. DrugBank is used to include drugs regulating positively or negatively genes in the network.

The network is represented as a graph where the vertices are genes, drugs and GO terms from the molecular function and biological process ontologies. The bioentity types queried from Moksiskaan are: *gene*, *drug*, *biolog-*

ical_process and *molecular_function*. The vertices are connected by directed edges, which represent the following relationships:

- Gene expression and repression between a gene and a gene.
  Moksiskaan link types: *gene_expression* and *gene_repression*.
- Drug promotion and inhibition between a drug and a gene.
  Moksiskaan link types: *drug_promotes* and *drug_inhibits*.
- Positive or negative regulation between a gene and a GO term.
  Moksiskaan link types: *positive_regulation* and *negative_regulation*.

Note that all the edges are independent. There is no information on the combined regulatory effects of multiple edges incident on the same vertice. For example, two genes might be known to regulate the expression of a third gene, but the combined regulatory effect of the genes is unknown.

3.2 Pruning the candidate network with experimental data

The candidate network contains genes and their regulatory interactions from various biological contexts. All of the genes and the interactions are not necessarily present in a specific condition [37]. For this reason, the candidate network is pruned with experimental data to fit the candidate network better to the biological context under investigation.

Experimental data are used to assign the genes as either upregulated or downregulated. Employing Moksiskaan, the sets of upregulated and downregulated genes are used to prune the candidate network. Only the genes in the candidate network are submitted to the pruning operations, while drugs and GO terms are not. The sets of upregulated and downregulated genes are mapped to the states of the genes in the candidate network. The upregulated genes are considered to be active (1) and the downregulated genes inactive

(0). The state information of the genes in the candidate network is then used to prune the network and to predict the state of genes with unknown state.

Vertices and edges incompatible with the experimental data are discarded as follows. The state information is propagated to the genes lacking state information, given there are no ambiguities in their upstream regulators. Genes still lacking state information are removed from the network. Edges which are in contradiction with the known states of the genes on their both ends are removed. Finally, orphan genes that have become disconnected from the network are removed to produce the pruned network.

3.3 Calculation of the logical steady state of the network after perturbation

The effect of perturbing the network with a drug is computed separately for each drug in the pruned network. The pruned network is transformed into a Boolean network to simulate the effect of a drug on the state of the network with BoolNet R package.

To transform the network in to a Boolean network, the network was first represented as a directed graph $G = (V, E)$ with vertices V and directed edges E between the vertices. Every vertex is associated with a state, namely active or inactive: $\sigma : V \rightarrow \{0, 1\}$. Every edge has a link type T assigned to them. Function τ maps the link types to activating or inactivating link types $\tau : E \rightarrow \{-1, 1\}$. The link types are assigned to activating or inactivating as

- **Activating**: *gene expression* and *drug promotes*
- **Inactivating**: *gene repression* and *drug inhibits*

The GO terms and the drugs, except the one used as the perturbation, are removed from the graph. The graph is transformed into a Boolean network $B = (V, F)$, where $V = \{v_1, ..., v_n\}$ are the vertices of the network and $F = \{f_1, ..., f_n\}$ are the transition functions of the vertices. The transition

functions represent the regulatory interactions between the vertices. The vertices are modeled as Boolean variables x_i with two states, namely active (1) or inactive (0).

For vertex $i \in \{1, ..., n\}$, the vertices with an edge to v_i are combined into a transition function f_i represented as a Boolean expression. Let $J_i^+ = \{j_i^{+,1}, ..., j_i^{+,n_i^+}\} = \{j_i^{+,k} \in \{1, ..., n\} \mid \exists(e = (u_{j_i^{+,k}}, v_i) \in E \wedge \tau(e) = 1)\}$ be the indices of vertices with an edge to v_i with activating link types, and let $J_i^- = \{j_i^{-,1}, ..., j_i^{-,n_i^-}\} = \{j_i^{-,k} \in \{1, ..., n\} \mid \exists(e = (u_{j_i^{-,k}}, v_i) \in E \wedge \tau(e) = -1)\}$ be the indices of vertices with and edge to v_i with inactivating link type. Since there is no information on the regulatory interaction of vertices having edges to v_i, the terms are combined with the OR operator

$$f_i = x_{j_i^{+,1}} \; OR \; ... \; OR \; x_{j_i^{+,n_i^+}} \; OR \; !x^n_{j_i^{-,1}} \; OR \; ... \; OR \; !x_{j_i^{-,n_i^-}} \; , \tag{3.1}$$

or alternatively with the AND operator

$$f_i = x_{j_i^{+,1}} \; AND \; ... \; AND \; x_{j_i^{+,n_i^+}} \; AND \; !x^n_{j_i^{-,1}} \; AND \; ... \; AND \; !x_{j_i^{-,n_i^-}} \; . \tag{3.2}$$

When using the OR operator no interaction between the regulatory effects of vertices having edges to vi_i is assumed. One active vertice having an activating edge to v_i or one inactive vertice having an inactivating edge to v_i is enough to activate v_i. Conversely, when using AND operator it is assumed that for v_i to become active all vertices with activating edge to v_i must be active and all vertices with an inactivating edge to v_i must be inactive.

If vertex v_i does not have any edges incident to it, an activating edge from v_i to v_i is created. Thus, the state of vertex v_i during the simulation will be the given initial state, as the transition function is $f_i = x_i$ and thus $x_i(t+1) = f_i(t) = x_i(t)$.

The states of the genes in the pruned network are used as the initial states of the vertices. The drug used to perturb the system is fixed to active state and the partial steady state of the network is calculated. The logical steady state (LSS) analysis introduced in [16] was first applied to solve the

partial LSS of the network. However, it proved unsatisfactory as the initial states of the genes cannot be given without fixing them to constant values for the whole duration of the simulation. Instead, the attractor of the system is computed using synchronous update of the states. Since the initial states of the vertices are known and synchronous update is used, the system will at some point re-enter a state previously visited [13]. After this, the system will follow the same path trough the states as before starting from the first re-entered state [13]. Therefore, there can only exist one attractor for a given initial state of the system.

If the network is perturbed with the drug at time point $t = 0$, the vertices for which there exists t_{ss}, such that $x(t+1) = x(t)$, when $t > t_{ss}$, are taken to be in steady state after the perturbation. In other words, if the state of the vertex is constant in the attractor after some time period, the vertex is said to have an LSS in the attractor. If the state of the vertex fluctuates in the attractor, the vertex is said to have an unknown LSS in the attractor. Thus an LSS for the network can be determined after the perturbation.

3.3.1 Simple example network

Next, the computation of the LSS of a network is illustrated with a simple example network. Also, the effect of assuming AND or OR operators for combining the incoming edges of the vertices is explored.

Figure 3.1 depicts the topology of the example network and the initial states of the vertices. The vertices could represent genes, drugs or biological processes. An edge with an arrowhead represents an activating interaction and an edge with a tee-head represent an inactivating interaction. The states of the vertices are illustrated by the boundary colors of the vertices. Green codes for active state and blue for inactive state.

A perturbation is introduced to the network by fixing the state of the vertice $I1$ to active. The attractor for the network is calculated and the LSSs of the vertices in the attractor are determined. This is performed with two distinct assumptions: combining the incoming edges with the AND

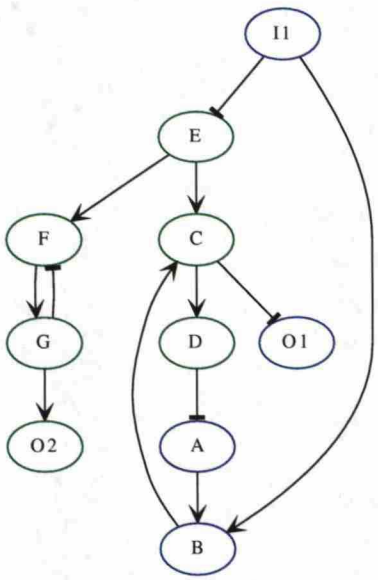


Figure 3.1: The initial state of the example network. Active vertices are marked with green boundaries and inactive vertices with blue boundaries. An arrowhead edge stands for an activating interaction and a tee-head edge stands for an inactivating interaction.

operator or combining the incoming edges with the OR operator to illustrate the differences between these two assumptions. The resulting LSSs for the vertices are shown for the case of the AND operator in figure 3.2 and for the case of the OR-operator in figure 3.3.

The states of all the vertices are determined in the LSS for the case of the AND operator. In contrast, the states of the vertices F, G and $O2$ are undetermined in the LSS for the case of the OR operator. This follows from the fact that the attractor for the case of the AND operator is a point attractor, which comprises only one state, whereas for the case of the OR operator, the attractor is a complex attractor, which comprises four states listed in table 3.1. In the complex attractor, the states of F, G and $O2$ vertices oscillate and are left thus undetermined. After the first round of

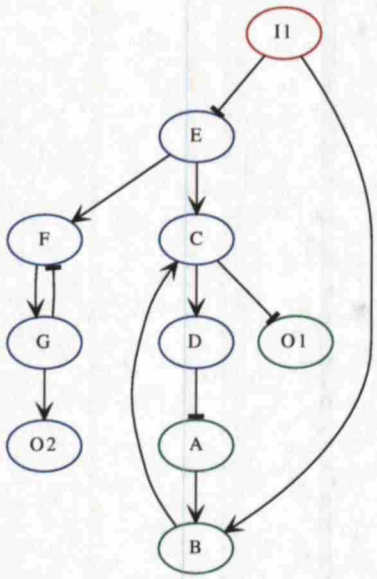


Figure 3.2: LSSs of the vertices in the example network assuming incoming edges are combined using the AND operator. Vertices with red boundaries are fixed to active state. Vertices in active state are marked with green boundaries and vertices in inactive state with blue boundaries. An arrowhead edge stands for an activating interaction and a tee-head edge stands for an inactivating interaction.

simulation $F = 1$ as it is activated by E, while $G = O2 = 1$. On the following round E has become inactive and thus $F = 0$, while $G = O2 = 1$. Next, F stays inactive and G becomes inactive, while $O2$ is active. Following this, F becomes active and $G = O2 = 0$. This leads to activation of G, while $F = 1$ and $O2 = 0$. Next, $F = 0$ and $G = O2 = 1$ which then starts the cycle again. The states of the vertices follow this cycle in the attractor, and thus the LSS cannot be determined for these vertices.

The difference between combining the incoming edges with the AND or the OR operator is also observed for the LSSs of vertices A, C, D and $O1$. For example, vertice C is inactive when assuming the AND operator since

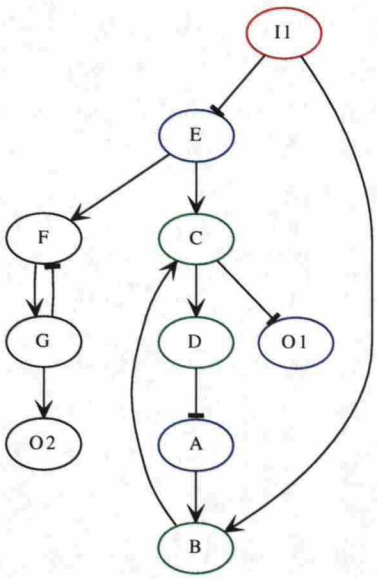


Figure 3.3: LSSs of the vertices in the example network assuming incoming edges are combined using the *OR* operator. Vertices with red boundaries are fixed to active state. Vertices in active state are marked with green boundaries and vertices in inactive state with blue boundaries. Vertices with black boundaries have undetermined state. An arrowhead edge stands for an activating interaction and a tee-head edge stands for an inactivating interaction.

E is inactive and the activation of C would require both E and B to be in the active state. For the case of the *OR* operator, the active state of B is enough to activate C.

3.4 Evaluating the biological effects of a drug

The evaluation of the biological effects of perturbing the network with a drug is done by introducing an activation score for the vertices for the initial state of the network and the LSS of the perturbed network, and then comparing

Table 3.1: The set of states forming the attractor of the example network assuming the incoming edges are combined using the OR operator. Each row represents one state and the columns the vertices. Active state is marked by 1 and inactive state by 0.

I1	E	F	G	O2	A	B	C	D	O1
1	0	1	0	0	0	1	1	1	0
1	0	1	1	0	0	1	1	1	0
1	0	0	1	1	0	1	1	1	0
1	0	0	0	1	0	1	1	1	0

the activation scores for the two different states of the network.

The activation score for a vertex in a network is calculated as follows. The network is represented as a directed graph $G = (V, E)$ with vertices V and directed edges E between the vertices. Every edge has a link type T assigned to them. Function τ maps the link types to activating or inactivating link types $\tau : E \to \{-1, 1\}$. Every vertex is associated with a state, namely active, inactive or unknown: $\sigma : V \to \{0, 1, \perp\}$. The activation score for vertex v is simply the sum of incoming activating signals and inactivating signals divided by the number of incoming edges.

More formally, for $v \in V$, let n be the cardinality of the set $\{(u, v) \in E\}$ of edges incident to v, and let n' be the cardinality of the set $E' = \{(u, v) \in E \mid \sigma(u) \neq \perp\}$ of edges incident to v with a source vertex with a known state. Activation score $S : V \to \mathbb{R}$ is defined if and only if $n \neq 0$ and $n' \neq 0$ as

$$S(v) = \frac{1}{n} \sum_{e \in E'} \tau(e)\sigma(u) \tag{3.3}$$

The difference between the activation scores for the initial state of the network and the LSS of the network after the perturbation is taken to describe the effect of the drug on each vertex $v \in V$:

$$\Delta S(v) = S_p(v) - S_0(v), \tag{3.4}$$

where $S_p(v)$ and $S_0(v)$ are the activation scores for vertex v after the perturbation and before, respectively.

Chapter 4

Prediction of drug effects for TNBC

This chapter describes how the drug effect prediction method is applied to a network constructed for TNBC patients. The effect of potential drugs for TNBC patients is evaluated. A candidate network containing genes, drugs and GO terms is constructed based on the knowledge in the biological databases. The network is pruned with TNBC gene expression data and the initial states of the vertices are derived from the data. The LSSs of the vertices are determined before and after perturbing the system with a drug. The biological effects of administrating the drug to the system are predicted by comparing the activation scores of the vertices before and after perturbation.

4.1 Candidate network construction

To predict the drug effects for TNBC patients, a network relevant in the context of TNBC patients is needed. Breast carcinomas arise from the mammary epithelium [11], and excessive proliferation is characteristic for malignant cancer cells [4]. A gene regulatory network regulating cell proliferation in TBNC is searched for. First, GO term *GO:0033598 mammary gland epithelial cell proliferation* is used as a seed to search for genes directly positively or negatively regulating cell proliferation in mammary epithelium. The set of can-

didate genes is used as the core set from which the network is expanded one step upstream and one step downstream. The information is retrieved from the following databases: GO, KEGG, WikiPathways, PathwayCommons and DrugBank.

Moksiskaan database (version 2.01) was queried for genes regulating positively or negatively biological process *GO:0033598 mammary gland epithelial cell proliferation.* The query resulted in a core set of 10 candidate genes, which are listed in table 4.1. The network expanded from the set of candidate genes one step upstream and one step downstream resulted in a candidate network of 191 vertices and 728 edges. The network consisted of genes, drugs and GO terms and the regulatory interactions between the bioentities. The vertice and edge types for the candidate network are listed in table 4.2.

Table 4.1: The set of candidate genes regulating biological process *GO:0033598 mammary gland epithelial cell proliferation.* Regulation Type indicates whether the gene has a positive or a negative regulatory effect on *GO:0033598 mammary gland epithelial cell proliferation.* Fold change in log2 and the related FDR value are given for the comparison between triple negative solid tumor samples and the solid normal samples used to prune the candidate network.

Ensembl Gene ID	**Gene Name**	**Regulation type**	**Fold Change**	**FDR**
ENSG00000139618	BRCA2	negative regulation	2.99	9.24E-21
ENSG00000107485	GATA3	negative regulation	-1.86	3.36E-05
ENSG00000175832	ETV4	negative regulation	1.15	1.51E-08
ENSG00000215021	PHB2	negative regulation	0.34	2.29E-04
ENSG00000183779	ZNF703	positive regulation	-0.27	6.11E-01
ENSG00000169855	ROBO1	negative regulation	0.26	6.89E-01
ENSG00000125686	MED1	positive regulation	0.22	3.92E-02
ENSG00000110092	CCND1	positive regulation	0.17	9.33E-01
ENSG00000183856	IQGAP3	positive regulation	0.13	7.59E-01
ENSG00000135439	AGAP2	positive regulation	NA	NA

Table 4.2: The total number and the number of different types of vertices and edges in the candidate network and the pruned network.

	Candidate network	**Pruned network**
Edge type: drug inhibits	3	0
Edge type: drug promotes	3	2
Edge type: gene expression	345	24
Edge type: gene repression	76	6
Edge type: negative regulation	140	58
Edge type: positive regulation	161	59
Number of edges	728	149
Number of vertices	191	81
Vertex type: biological process	53	40
Vertex type: drug	1	1
Vertex type: gene	130	34
Vertex type: molecular function	7	6

4.2 Pruning of the network

The network is pruned using experimental gene expression data from TCGA. Level one raw gene expression microarray data was downloaded from TCGA for 524 primary breast carcinoma tumors and 59 solid normal samples [36]. The gene expression measurements were performed on AgilentG4502A_07_3 microarray. All the arrays were normalized to have a mean of zero. Probes matching either multiple or no genes were removed using Ensembl database [38].

The clinical data for the breast cancer patients was downloaded from TCGA. The clinical information was used to select TNBC patients, namely those who lack the expression of HER2, progesterone receptor and estrogen receptor [12]. This resulted in 55 triple-negative primary breast carcinoma tumors.

Differentially expressed genes (DEGs) between the triple-negative solid tumor samples and the solid normal samples were determined as follows. First, the group median was calculated for both groups and then fold change was calculated by taking the ratio between the group medians. Second, a t-test was performed to asses the statistical significance of the differential expression for each gene. To reduce the number of false positives, a multiple hypothesis correction method was applied to calculate the false discovery rate (FDR) based on the p-values [39]. Last, genes with fold change at least 2 in either direction and $FDR \leq 0.05$ were considered differentially expressed.

The comparison of triple-negative solid tumor samples to normal solid samples resulted in 2164 upregulated and 2172 downregulated genes out of 15282 genes in total. The candidate network was pruned using the sets of DEGs, which resulted in a pruned network of 81 vertices and 149 edges. The vertice and edge types for the candidate network are listed in table 4.2. Only three genes, *BRCA2*, *GATA3* and *ETV4*, of the original set of 10 candidate genes were in the set of DEGs and thus included in the pruned network. The states of the genes in the pruned network were determined based on the DEG sets. The structure of the pruned network and the initials states of the genes are illustrated in figure 4.1. The state of the genes are coded with green and blue boundaries for active and inactive states, respectively. The states derived directly from the data are marked with thicker boundaries, whereas states predicted from the data are marked with thinner boundaries. The genes in the original candidate gene set, from where the network was expanded, are shown in white background. The network also includes the drugs and the GO terms.

4.3 Prediction of drug effects

The candidate network includes one drug, Arsenic trioxide (ATO), and three regulatory interactions between the drug and the genes. In the pruned network the number of regulatory interactions is reduced to two: the activation

of genes *JUN* and *MAPK1*. The effect of perturbing the pruned network with ATO was computed assuming the incoming edges were combined using *OR* operator. The *OR* operator was chosen because this does not assume any interaction between the regulations of the same gene. The resulting LSS state of the network is shown in figure 4.2. The activation score difference for genes and GO terms is shown in table 4.3 for those vertices with an activation score difference unequal to zero. Also, the number of activating and inhibiting signals incident on a gene or a GO term before and after perturbation and the total number of edges incident on a gene or GO term is reported.

ATO is used in the treatment of newly diagnosed and relapsed acute promyelotic leukemia patients, and it has been shown to have an anticancer effect on many solid tumors [40]. The exact molecular mechanism of ATO's anticancer effect are not known but it works mainly through elevating oxidative stress levels in the cells [40].

The simulation results predict that ATO would inactivate *mammary gland epithelial cell proliferation* ($\Delta S = -0.333$), as there is one activating signal less out of three regulators of the process after the perturbation. Also, *cell cycle* ($\Delta S = -0.500$) is predicted to be inhibited by addition of two inactivating signals out of four regulators and *endothelial cell apoptotic process* ($\Delta S = -1$) is predicted to be inhibited by the only regulator after treatment with ATO.

The results obtained here are in part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/. For the TCGA studies used here the study accession number in the database of Genotypes and Phenotypes (dbGaP) is phs000569.v1.p7.

Table 4.3: Activation score differences ΔS, activation scores for the initial state of the network S_0 and activation scores after the perturbation of the network S_P for genes and GO terms. The scores are listed for genes and GO terms with an difference in the activation score between the initial state and the LSS after the perturbation of the network with Arsenic trioxide. A_0 and A_P are the number of activating signals incident on a gene or a GO term before and after the perturbation. I_0 and I_P are the number of inhibiting signals incident on a gene or a GO term before and after the perturbation. N is the total number of edges incident on a gene or a GO term

VERTICE	ΔS	S_P	A_P	I_P	S_0	A_0	I_0	N
IL13	1.000	1.000	3	0	0.000	0	0	3
interleukin-13 secretion	1.000	1.000	1	0	0.000	0	0	1
JUN	1.000	1.000	2	0	0.000	0	0	2
transcription regulatory region DNA binding	1.000	1.000	1	0	0.000	0	0	1
IL10	1.000	1.000	2	0	0.000	0	0	2
GTPase activity	1.000	1.000	1	0	0.000	0	0	1
FOS	1.000	1.000	1	0	0.000	0	0	1
interleukin-5 secretion	1.000	1.000	1	0	0.000	0	0	1
FEV	1.000	1.000	1	0	0.000	0	0	1
protein kinase B signaling cascade	1.000	1.000	1	0	0.000	0	0	1
MYB	1.000	1.000	1	0	0.000	0	0	1
ETS1	1.000	1.000	3	0	0.000	0	0	3
T cell differentiation	1.000	1.000	1	0	0.000	0	0	1
cellular amine metabolic process	1.000	1.000	1	0	0.000	0	0	1
thyroid hormone generation	1.000	1.000	1	0	0.000	0	0	1
FABP4	1.000	1.000	2	0	0.000	0	0	2
interleukin-4 production	1.000	1.000	1	0	0.000	0	0	1
FOSL2	0.667	0.667	2	0	0.000	0	0	3
endothelial cell migration	0.500	0.500	1	0	0.000	0	0	2
GATA3	0.500	0.000	1	1	-0.500	0	1	2
cellular process	0.500	0.000	1	1	-0.500	0	1	2
immune response	0.500	0.500	1	0	0.000	0	0	2
KLK3	0.333	-0.333	1	2	-0.667	0	2	3
cell migration	0.250	0.750	3	0	0.500	2	0	4
transcription from RNA polymerase II promoter	0.182	0.227	11	6	0.045	5	4	22
transcription, DNA-dependent	-0.100	0.300	6	3	0.400	5	1	10
response to external stimulus	-0.250	-0.250	1	2	0.000	0	0	4
fat cell differentiation	-0.333	0.000	1	1	0.333	1	0	3
mammary gland epithelial cell proliferation	-0.333	-1.000	0	3	-0.667	0	2	3
cell cycle	-0.500	-1.000	0	4	-0.500	0	2	4
interferon-gamma production	-1.000	-1.000	0	2	0.000	0	0	2
glial cell-derived neurotrophic factor receptor signaling pathway involved in ureteric bud formation	-1.000	-1.000	0	1	0.000	0	0	1
fibroblast growth factor receptor signaling pathway involved in ureteric bud formation	-1.000	-1.000	0	1	0.000	0	0	1
protein catabolic process	-1.000	-1.000	0	1	0.000	0	0	1
protein kinase activity	-1.000	-1.000	0	1	0.000	0	0	1
endothelial cell apoptotic process	-1.000	-1.000	0	1	0.000	0	0	1
cell proliferation involved in mesonephros development	-1.000	-1.000	0	1	0.000	0	0	1
cell motility	-1.000	-1.000	0	1	0.000	0	0	1
interleukin-2 production	-1.000	-1.000	0	1	0.000	0	0	1

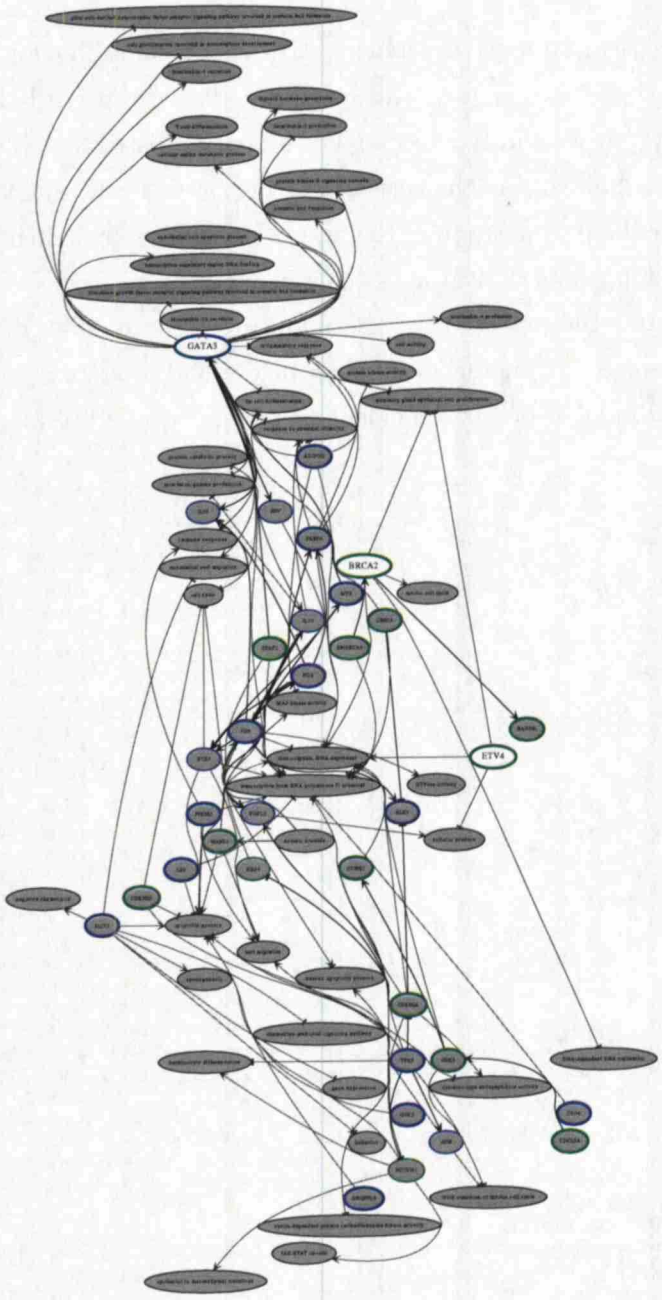


Figure 4.1: The initial state of the TNBC network. The states of the genes are coded with blue and green boundaries for inactive and active states, respectively.

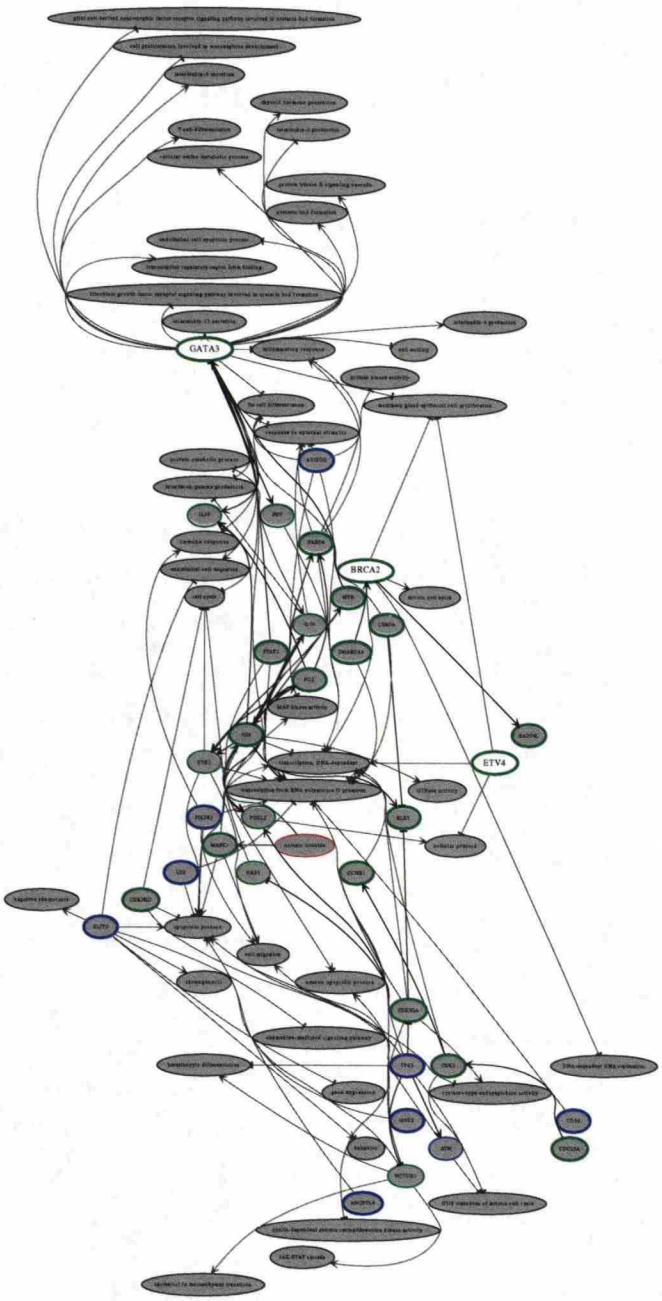


Figure 4.2: The LSS of the genes for the TNBC network after perturbation with ATO. The states of the genes are coded with blue and green boundaries for inactive and active states, respectively. ATO is shown with a red boundary.

Chapter 5

Discussion

The prediction of drug targets is an important task that can speed up biomedical research and potentially lead to improvements in patient care. This thesis implemented an analysis framework for predicting potential drug targets in the context of GRNs. The analysis framework can be applied flexibly to a number of biologically interesting systems. The network containing genes, drugs and biological processes can be generated with Moksiskaan to suit the biological question in hand, and different transcriptomics data, such as gene expression or RNA-seq data, can be used to fit the candidate network to the relevant biological context.

The effect of perturbing a biological system with a drug was estimated by simulating the system with Boolean networks. Even if information on the mechanistic details and the specific kinetic parameters of the chemical reactions are lacking, Boolean networks are able to give qualitative predictions on the dynamic behavior of the system [2]. Since Moksiskaan generated networks do not contain the required information to build continuous or single-molecular level models, Boolean networks were chosen to model the system.

The herein developed activation score was used to predict the effect of a drug on the network. The activation score measures the activation of the components of the network under a steady state. The activation score was

calculated for the genes and the GO terms before and after perturbing the network with a drug. The differences of the activation scores were used to predict the effect of a drug on the genes and the biological processes in the network.

The analysis framework was applied to predict the drug effects on TNBC patients. ATO was predicted to inactivate *mammary gland epithelial cell proliferation* and *cell cycle*, which would be in agreement with anticancer effects of ATO in other cancer types [40]. However, ATO is also predicted to inactivate *endothelial cell apoptotic process*, which could imply a tumor-promoting effect of ATO in TNBC patients. These results leave unclear the effect of ATO on malignant cancer cells in TNBC patients.

In the future, it would be interesting to take into account different timescales of different biological processes. Here, the states of the genes are updated synchronously, which implies that all the regulatory interactions are assumed to occur at similar timescales [2]. In reality, different biological interactions have different time scales [2]. Asynchronous update scheme can be utilized to take into account these differences, if timescales of different types of interactions are known [2]. For example, TF based regulation of gene expression has a different timescale compared to miRNA based regulation [41]. Also, Moksiskaan does not currently contain information on the regulatory interactions of genes regulating the same gene. For this reason, Boolean functions were assumed to contain only OR operations when simulating drug effects for TNBC patients. As more data accumulates in the biological databases, the Boolean functions can be modified to reflect this new information.

In summary, the major development in this thesis was the construction of an analysis framework that can be flexibly utilized to predict the effect of drugs on different biological systems. For this purpose, a method for calculating the partial logical steady states of the genes after perturbing the network was implemented. Furthermore, an activation score to asses the activity of the components in different conditions was created. Finally, the

method was applied to predict drug effects in TNBC patients.

Bibliography

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

[2] Rui-Sheng Wang, Assieh Saadatpour, and Réka Albert. Boolean modeling in systems biology: an overview of methodology and applications. *Physical Biology*, 9(5):055001, 2012.

[3] Peter Boyle, Bernard Levin, et al. *World cancer report 2008*. IARC Press, International Agency for Research on Cancer, 2008.

[4] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646 – 674, 2011.

[5] Tom Strachan and Andrew P. Read. *Human Molecular Genetics 4*. Garland Science/Taylor & Francis Group, 2011.

[6] David S. Latchman. Transcription factors: An overview. *The International Journal of Biochemistry & Cell Biology*, 29(12):1305 – 1312, 1997.

[7] Tong Ihn Lee and Richard A. Young. Transcription of eukaryotic protein-coding genes. *Annual Review of Genetics*, 34(1):77–137, 2000.

[8] Thomas Schlitt and Alvis Brazma. Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, 8(Suppl 6):S9, 2007.

[9] Robert A. Weinberg. *The Biology of Cancer.* Garland Science, Taylor & Francis Group, LLC, 2007.

[10] Douglas Hanahan and Robert A. Weinberg. The Hallmarks of Cancer. *Cell*, 100(1):57 – 70, 2000.

[11] Fattaneh A Tavassoli and Peter Devilee. *Pathology & genetics: tumours of the breast and female genital organs*, volume 4. IARC, 2003.

[12] William D. Foulkes, Ian E. Smith, and Jorge S. Reis-Filho. Triple-Negative Breast Cancer. *New England Journal of Medicine*, 363(20):1938–1948, 2010.

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

[14] Stuart A. Kauffman. *The origins of order: Self organization and selection in evolution.* Oxford University Press, USA, 1993.

[15] Hidde de Jong. Modeling and simulation of genetic regulatory systems : a literature review. *Journal of Computational Biology*, 9:67–103, 2002.

[16] Steffen Klamt, Julio Saez-Rodriguez, Jonathan Lindquist, Luca Simeoni, and Ernst Gilles. A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics*, 7(1):56, 2006.

[17] Gregory W. Carter. Inferring network interactions within a cell. *Briefings in Bioinformatics*, 6(4):380–389, 2005.

[18] Anna Bauer-Mehren, Laura I. Furlong, and Ferran Sanz. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular systems biology*, 5(1), 2009.

[19] Marko Laakso and Sampsa Hautaniemi. Integrative platform to translate gene sets to networks. *Bioinformatics*, 26(14):1802–1803, 2010.

[20] Marko Laakso, Riku Louhimo, Kristian Ovaska, Tiia Pelkonen, and Sampsa Hautaniemi. Moksiskaan User Guide, April 2013. `http://csbi.ltdk.helsinki.fi/moksiskaan/userguide/userguide.html`.

[21] Marko Laakso. Moksiskaan component API. Webpage. `http://csbi.ltdk.helsinki.fi/moksiskaan/anduril/index.html?q=\Moksiskaan%20project`.

[22] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[23] Seth Carbon, Amelia Ireland, Christopher J. Mungall, ShengQiang Shu, Brad Marshall, Suzanna Lewis, the AmiGO Hub, and the Web Presence Working Group. AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 2009. AmiGO 2, 2013-07-14.

[24] Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, An Chi Guo, and David S. Wishart. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research*, 39(suppl 1):D1035–D1041, 2011.

[25] DrugBank Open Data Drug & Drug Target Database. Webpage. `http://www.drugbank.ca`. Accessed on 17-07-2013.

[26] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2012.

[27] Thomas Kelder, Martijn P. van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R. Conklin, Chris T. Evelo, and Alexander R. Pico. Wikipathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(D1):D1301–D1307, 2012.

[28] WikiPathways. Webpage. `http://www.wikipathways.org`. Accessed on 17-07-2013.

[29] Thomas Kelder, Alexander R. Pico, Kristina Hanspers, Martijn P. van Iersel, Chris Evelo, and Bruce R. Conklin. Mining Biological Pathways Using WikiPathways Web Services. *PLoS ONE*, 4(7):e6447, 07 2009.

[30] Ethan G. Cerami, Benjamin E. Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D. Bader, and Chris Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(suppl 1):D685–D690, 2011.

[31] PathwayCommons. Webpage. `http://www.pathwaycommons.org`. Accessed on 17-07-2013.

[32] Kristian Ovaska, Marko Laakso, Saija Haapa-Paananen, Riku Louhimo, Ping Chen, Viljami Aittomaki, Erkka Valo, Javier Nunez-Fontarnau, Ville Rantanen, Sirkku Karinen, Kari Nousiainen, Anna-Maria Lahesmaa-Korpinen, Minna Miettinen, Lilli Saarinen, Pekka Kohonen, Jianmin Wu, Jukka Westermarck, and Sampsa Hautaniemi. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Medicine*, 2(9):65, 2010.

[33] Kristian Ovaska, Ping Chen, Marko Laakso, Ville Rantanen, Riku Louhimo, Sirkku Karinen, Javier Núñez Fontarnau, and Rogojin Vladimir. Anduril User Guide, April 2013. `http://csbi.ltdk.helsinki.fi/pub/anduril/userguide.pdf`.

[34] Christoph Müssel, Martin Hopfensitz, and Hans A. Kestler. BoolNet - an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics*, 26(10):1378–1380, 2010.

[35] Christoph Müssel, Martin Hopfensitz, and Hans A. Kestler. BoolNet package vignette. Webpage. `http://cran.r-project.org/web/packages/BoolNet/BoolNet.pdf`.

[36] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012.

[37] 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.

[38] Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Laurent Gil, Leo Gordon, Maurice Hendrix, Thibaut Hourlier, Nathan Johnson, Andreas K. Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Monika Komorowska, Gautier Koscielny, Eugene Kulesha, Pontus Larsson, Ian Longden, William McLaren, Matthieu Muffato, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Harpreet Singh Riat, Graham R. S. Ritchie, Magali Ruffier, Michael Schuster, Daniel Sobral, Y. Amy Tang, Kieron Taylor, Stephen Trevanion, Jana Vandrovcova, Simon White, Mark Wilson, Steven P. Wilder, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M. Fernández-Suarez, Jennifer Harrow, Javier Herrero, Tim J. P. Hubbard, Anne Parker, Glenn Proctor, Giulietta Spudich, Jan Vogel, Andy Yates, Amonida Zadissa, and Stephen M. J. Searle. Ensembl 2012. *Nucleic Acids Research*, 40(D1):D84–D90, 2012.

[39] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[40] Jun Lu, Eng-Hui Chew, and Arne Holmgren. Targeting thioredoxin reductase is a basis for cancer therapy by arsenic trioxide. *Proceedings of the National Academy of Sciences*, 104(30):12288–12293, 2007.

[41] Oliver Hobert. Gene Regulation by Transcription Factors and MicroRNAs. *Science*, 319(5871):1785–1786, 2008.