Freie Universität zu Berlin

Masterthesis

Inference of Boolean Networks considering real-life time course Data

Nina Valery Kersten

Supervisors

Prof. Dr. Heike Siebert
Prof. Dr. Alexander Bockmayr

Advisor

Phd. Robert Schwieger

November 20, 2018

Abstract

Contents

List of Figures

List of Tables

1 Introduction

The development and functioning of a cell and an organism in general is a product of a complex cellular machinery. This machinery is compound by the interaction of genes, proteins, mRNA and many other substances to induce a cascade of extracellular signals transducted by mechanisms of the cell membrane,reaching the nucleus of the cell, initiating a transcription process that controls the production and abundance of proteins. Proper functioning of these networks is essential to the survival and adapation of all living organisms, while malfunctioning of these networks has been identified as the cause of various diseases (**?**). To understand the behaviour of a biological system it is necessary to find and analyze the main important processes in a system in a dynamical manner. Therefore high-throughput techniques provide a big abundance of information about various biological interactions measured over a series of time. Biological information can be considered as different systems such as signal transduction, gene regulation, protein-protein-interaction or metabolism. These information are put in a network which can be yielded by several strategies like Baysian networks, Boolean regulatory networks, Ordinary differential equation models and Neural networks(Saadatpour & Albert 2013).Once a network is constructed further anaylsis of the network by validating the network (e.g. pertubations like gene manipulation and external treatments, reductions etc.)can be done to figure out the main interactions whose disfunctionalities cause diseases. This masterthesis focuses on constructing a pipeline by creating a boolean network from a real-life time course data set. In the following section the biological background of data, it's preprocessing by normalization and descretization,the graphtheoretical background and scoring methods are described. In the second section it is getting more detailed by giving an overview of different inference algorithms, describing a tool called PyBoolNet and showing with an example data set and a real-life time course data set of breast cancer celllines how the pipeline is applied. In the last section the scoring results for the real-life time course data set are analyzed reflecting the peformance of the constructed network.

2 Background

2.1 Biological Background

Depending on the aim of a network inference the biological input data can be depicted by the interaction of proteins, genes and metabolic substances. In this section the intention of using different types of input data is explained and what potentially will be the occuring problems. Different types of biological input data provide different structures of the input data for further network inference algorithms.

Transcriptional Gene Regulatory Networks

In a Transcriptional Gene regulatory network (GRN) the nodes are depicted by the genes and the arcs are directed and show whether a gene produces RNA (transcript of the source gene,resp. regulator) which inhibits oder activtes the target gene (regulatee). For network inference computational algorithms take the mRNA expression levels of genes as the input data.

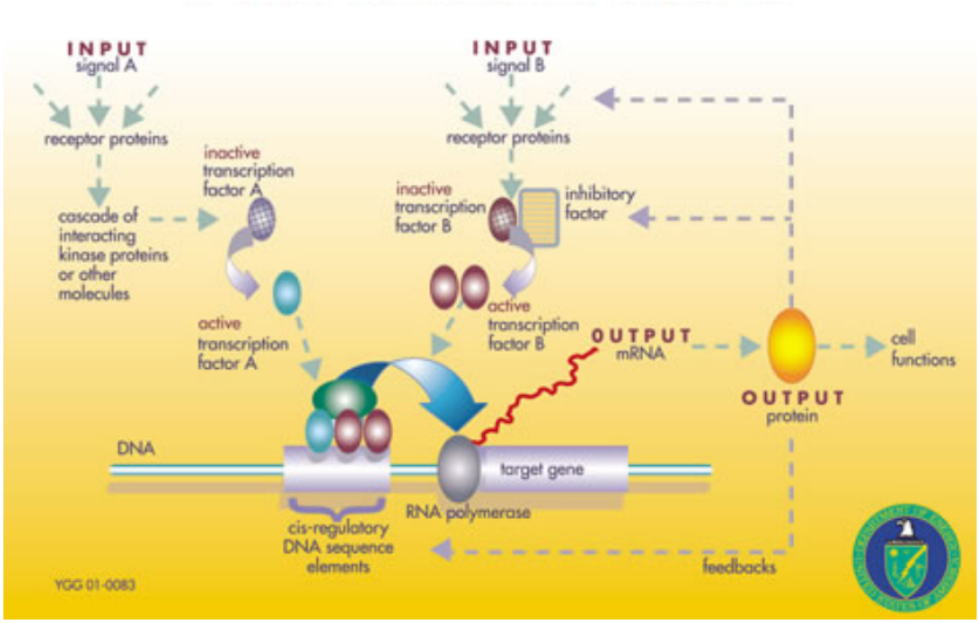

Figure 2.1: Transcriptional Gene Regulatory Network (GRN).
In this example two different signals have an impact of a single target gene. Signal molecule A triggers the conversion of inactive transcription factor A (green oval) into an active form that binds directly to the target gene's cis-regulatory sequence. The process for signal B is more complex. Signal B triggers the separation of inactive B (red oval) from an inhibitory factor (yellow rectangle). B is then free to form an active complex that binds to the active A transcription factor on the cis-regulatory sequence. The net output is expression of the target gene, leveled by A nd B. Thus cis-regulatory DNA sequences with the proteins that assemble on them, integrate information from multiple signaling inputs to produce mRNA-Output . (**?**)

Metabolic Interaction

Metabolic interactions represent the most complex cellular processes. Connections between biochemical reaction via substrate and product metabolites create complex metabolic networks. The focus is set on the different aspects of enzyme chemistry, enzymestructure and metabolite structure. Thus an individual's metabolism is determined by one's genetics, environment, and nutrition. By investigating the chemical structure of metabolites and systematically classify the functions of the enzymes the understanding of a metabolism and the prediction of enzyme function and novel metabolic pathways is improved.(**?**)(**?**)

Signal Transduction

In the developement of multicellular organisms the action of extracellular growth factors activate a cascade of intracellular signaling pathways. These pathways regulate major aspects of cell regulation like cell profileration, cell migration, cell differentiation,cell survival and cell death. To understand the developement of diseases (e.g. cancer) major prozesses (e.g. phosphorylation, ubiquitylation, methylation, etc.) of signal transduction pathways can be delightet by the prediction of a network . In signal transduction proteins are the nodes and directed edges represent interaction, where the biochemical modification of the regulatee is changed by the impact of the regulator. The concentration of signalling pathway underlies high fluctuation over time due to transcriptional and translational regulation, such that the inference of a network is a challenging task (Kestler et al. 2008)

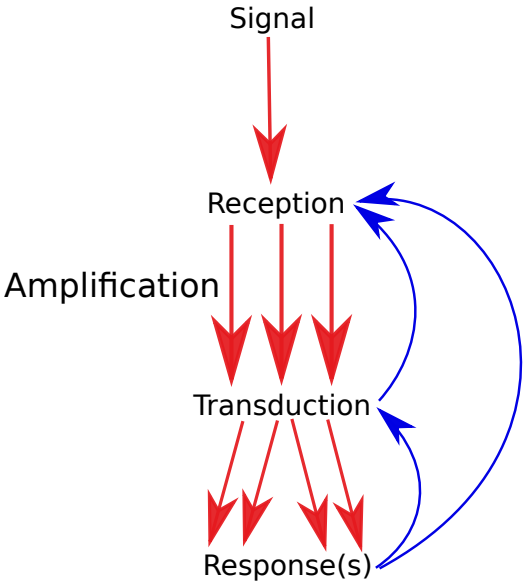

Figure 2.2: Signal Transduction. An environmental signal (e.g. hormone) interacts with a cellular component, most often a cell-surface receptor. The information that the signal has arrived is then converted into other chemical forms, or transduced. The signal is often amplified before introducing a response. Feedback pathways regulate the entire signaling process.(**????**)

Protein-Protein-Interaction

In contrast to the gene regulatory interaction network the protein-protein interactions (PPIs) act directly among themselve. Thus the nodes in a network are the interacting proteins. Proteins interact by physical contacts(e.g. electrostatic forces) of high specificity. PPIs play a big role in electron transfer,signal transduction, transport across membranes and cell metabolism. A variety of techniques are known to detect PPIs. The most applicated ones are immuno-precipitations and the yeast two-hybrid approach. The two-hybrid assay is not a relieable indication that two proteins interact *in vivo*,because the two interacting proteins are overexpressed.Thus the interaction may not be present in the wild type cells where the concentrations may be significantly lower. For this reason additional information are included to figure out the occurence of true interaction, such as cellular localization and mRNA expression level. The underlying assumption is that true interactions are likelyto occur between proteinsinvolved in the same biological process, proteins found in the same cell compartment, and proteins whose mRNA are coexpressed. For the identification of protein complexes affinity purification technique is used followed by mass spectrometry (MS) to sequence the proteins in the complexes. The detection of interactions of protein with DNA is done by chromatin immunoprecipitation (ChIP) in addition with expression microarrays, so-called ChIP on Chip approach. This method provide information about the interaction of transcription factors with DNA and the binding sites of transcription factors. Furthermore computational methods are included to predict the protein interactions by protein fusion and using phylogenetic analysis. Interaction networks of PPIs may depict how drug- protein interactions lead to toxic side effects. (Pellegrini et al. 2004)

2.2 Preprocessing

2.3 Boolean Network

Definition: Undirected Graph
In general, an undirected graph $G = (V, E)$ is defined as a set of vertices V describing the nodes of the system and a set of undirected edges $E = \{(i, j)|i, j \in V\}$ that define a relationship between node i and j.

This definition can be extended to obtain a notion of a directed graph:

Definiton: Directed Graph
A directed graph is an ordered pair $G = (V, A)$, defined as a set of vertices V (nodes) and a set of directed edges A (arcs). A set of directed edges $A = \{(i, j)| \in V\}$ describes the flow of information in the system, where (i, j) describes the flow from i (tail)to j (head).

2.4 Interaction Graph and State Transition Graph

Definition: Interaction Graph

Definition: State transition Graph
Let X be an n-dimensional binary vector that represents the current state of the system. Each element $X_i \in X$ corresponds to the state (0 or 1) of species i. A Boolean network defined by a set F of n Boolean functions. For every $f_i \in F$, such that $1 \leq i \leq n$, $f_i(X(t)) = X_i(t+1)$.

In other words,given a current state of thesystem $X(t)$, f_i determines the (binary) value of species X_i at time $t + 1$. Given a Boolean network N on n variables and an initial state $X(0) \in \{0, 1\}^n$, the dynamics of the system can be simulated by repeatedly applying the Boolean functions and updating the "current" state.
Definition: Boolean Regulatory Network

(Berestovsky & Nakhleh 2013)A boolean regulatory network consists of nodes (vertices) representing the components of a system and the edges (links) representing the interaction between the nodes. Each node can take two possible values of 1 (ON) or 0 (OFF). Depending on the input data this could mean, a gene is expressed or not, a transcription factor is active oder inactive, a molecular's concentration is above or below a certain threshold. The edges can be directed or undirected and show the orientation of masstransfer or information

respectively. The future state of a node is determined by Boolean rule (function) shown in Figure 1.3. For instance the regulator $v1$ should be inactive such that the regulatee $v2*$ can be active. Here $v2*$ describes the future state of $v2$ (Saadatpour & Albert 2013).

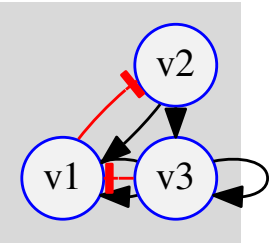

Figure 2.3: Interaction Graph. Shown is a selfinvented Interaction Graph constructed in PyBoolNet. The nodes $v1, v2, v3$ are denoted by blue circles, inhibitory arcs are red and activating arc are black.

$$f(v_{1,2,3}) = \begin{pmatrix} v_1 & \wedge v_2 & \wedge \neg v_2 \\ \neg v_1 & & \\ & v_2 & \wedge v_3 \end{pmatrix} \tag{2.1}$$

Beside the mathematical annotation, the boolean rule can be constructed with the operator AND,OR, NOT or they can be written as & , ||, ! . The graph in Fig.1. is called the Interaction Graph. Updating the state of a node yield in a network several constellation of states for each node. Regarding the example in Fig.1 for each boolean rule in $f(v)$ all the possible combination of states every node can have is inserted. With $f(v)$ it is possible to determine the next possible state of each node to build an *State Transition Graph*. The *State Transition Graph (STG)* is a directed graph representing the dynamical behaviour of a Regulatory Graph. Nodes of this graph represent possible states of the model, assigning a value to each component. Arcs of the STG represent transitions from one state to another. It is an important network to analyze how data behaves over time, thus the most possible state of each component can be computed. But not every state of a component of a node is updated at the same time. Therefore the *State Transition Graph* is either synchronous, updating all the node's states simultaneously, or asynchrounus, the node's states are updated based on their individual timesales(Lee et al. 2002).

Out of the *Interaction Graph* of Fig.1. the synchronous and anysnchronous STG is computed and shown in Fig.2.

$$f(v) = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \text{ where } v_{1,2,3} \in \{0,1\} \tag{2.2}$$

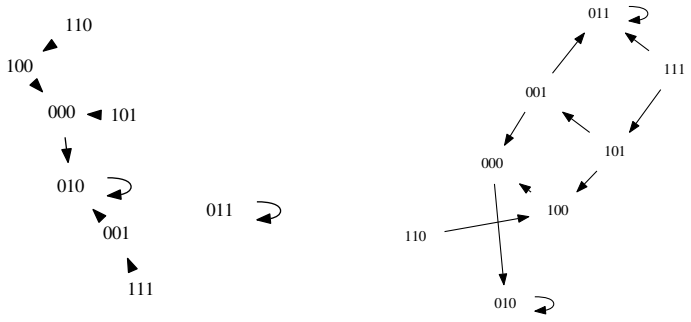

Figure 2.4: Synchronous and Asynchronous State Transition Graph. Left: Synchronous State Transition Graph; Right: Asynchronous State Transition Graph

(ÃĽlisabeth Remy et al. 2008)

2.5 Network Evaluation

3 Materials and Methods

3.1 Inferencealgorithms

regression (Chai et al. 2014)

Boolean Approach

Baysian networks

Ordinary differential equation models

Neural networks regression

BIBN (Bayesian inference approach for a Boolean network) REVEAL (Reverse Engineering algorithm) PCA-CMI (Path consistency algorithm- Conditonal mutual information) ARCANE (Time-delay algorithm for reconstruction of accurate cellular networks) MIDER (Mutual information distance and entropy reduction)(**?**) defines a mutual information based distance between genes to specify the directionality BESTFIT ()
These mutual information-based methods are computationally expensive, because they are implemented to compute exact mutual information values over all possible combinations of genes.
RelNet (Revelance network algorithm)
CLR (context likelihood of relatedness method) CST (chi-square test)

3.2 PyBoolNet

3.3 Data Selection

3.3.1 Example data set

3.3.2 HPN-DREAM breast cancer data set

Now it is shown how the Pipeline can be applied to a real-life time course data set.

What is the Dream Challenge?

For a Boolean network inference the data of a platform so-called Dialogue on Reverse Engineering Assessment and Methods (DREAM) - Challenge is used. The DREAM-Challenge is a non-profit, collaborative community effort consisting of contributors from across the research spectrum of questions in biology and medicine. This organization was built in 2006 and publishes crowdsourcing challenges with transparent sharing of data, thus everyone can participate the challenge. The DREAM-Challenge has partnered with Sage Bionetworks, which provide the infrastructure by Sage Bionetworks Synapse platform to get access to the open collaborative data analysis. Overall the DREAM-Challenge is a helpful instrument to get real-life data, comparing results and interact with other researchers all over the world, while contribute solutions to biological and medical questions.(**?**). The challenging question was to decide which Dream Challenge data set could be useful for this masterthesis. For inferring a Boolean network and further analysis of the state tarnsition graph the desired data set should contain measurements of experiments with less pertubational information and a in a time course context. The Dream Challenge 5 dealing with gene-gene interaction,providing test and training data sets of gene expressions seemed to be an appropriate candidate. But there was less time course information and a high abundance of pertubation. Thus the Dream8 Challenge is was chosen. This challenge describes protein-protein interaction and measurements for multiple timepoints.

DREAM8

Data Collection

The collection of the HPN8 breast cancer PPI data is done by a technique so called Reverse phase protein array(RPPA). This technique is divided up into 6 parts:

Sample collection An inhibitor or stimulus in form of drugs is added to a set of celllines at the same time and the celllines are then processed at different time points.

Cell Lysis Cell fragments are lysed with a celllysis buffer to obtain high protein concentration.The choice of a buffer decides the quantity of proteins can be lysed out of the cell.

Dilution Dilution of the celllysed probes.

Antibody screening The lysates are pooled and resolved by SDS-PAGE followed by western blotting on a nitrocellulose membrane. The membrane is cut into 4mmm strips. Each slide is probed with a different antibody, primary with a secondary antibody.

Fluorometric detection Primary and secondary antibody are diluted.Detection reagentisput on each slide. Signal amplification and detection is done by an optical flatbed scanner if colormetric technique is used orby laser scanning.

Data set structure Missing data points and outliers are detected and deleted from the data set. The data set is normalized

4 Appilcation of the Pipeline

4.1 Example data set

4.2 HPN-DREAM breast cancer data set

5 Results

5.1

5.2

5.3

6 Conclusion

References

Berestovsky, N., & Nakhleh, L. (2013, 06). An evaluation of methods for inferring boolean networks from time-series data. *PLOS ONE*, *8*(6), 1-9. Retrieved from https://doi.org/10.1371/journal.pone.0066031 doi: 10.1371/journal.pone.0066031

Chai, L. E., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S., & Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine*, *48*, 55 - 65. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010482514000420 doi: https://doi.org/10.1016/j.compbiomed.2014.02.011

De Las Rivas, J., & Fontanillo, C. (2010, 06). ProteinâĂŞprotein interactions essentials: Key concepts to building and analyzing interactome networks. *PLOS Computational Biology*, *6*(6), 1-8. Retrieved from https://doi.org/10.1371/journal.pcbi.1000807 doi: 10.1371/journal.pcbi.1000807

Elena S. Dimitrova, J., M. Paola Vera Licona. (2010). Discretization of time series data. *Journal of Computational Biology*(1), 853 - 868. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3203514/ doi: http://doi.org/10.1089/cmb.2008.0023

Kestler, H. A., Wawra, C., Kracher, B., & Kuehl, M. (2008). Network modeling of signal transduction: establishing the global view. *BioEssays*, *30*(11-12), 1110–1125. Retrieved from http://dx.doi.org/10.1002/bies.20834 doi: 10.1002/bies.20834

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., ... Young, R. A. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, *298*(5594), 799–804. Retrieved from http://science.sciencemag.org/content/298/5594/799 doi: 10.1126/science.1075090

Pellegrini, M., Haynor, D., & Johnson, J. M. (2004). Protein interaction networks. *Expert Review of Proteomics*, *1*(2), 239-249. Retrieved from https://doi.org/10.1586/14789450.1.2.239 (PMID: 15966818) doi: 10.1586/14789450.1.2.239

Saadatpour, A., & Albert, R. (2013). Boolean modeling of biological regulatory networks: A methodology tutorial. *Methods*, *62*(1), 3 - 12. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1046202312002770` (Modeling Gene Expression) doi: https://doi.org/10.1016/j.ymeth.2012.10.012

ÃLlisabeth Remy, Ruet, P., & Thieffry, D. (2008). Graphic requirements for multistability and attractive cycles in a boolean dynamical framework. *Advances in Applied Mathematics*, *41*(3), 335 - 350. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0196885808000146` doi: https://doi.org/10.1016/j.aam.2007.11.003

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbständig und ohne unerlaubte fremde Hilfe angefertigt, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Quellen und Hilfsmitteln wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Berlin, den 07. September 2018

.....

(Unterschrift .....)