Freie Universität zu Berlin

Master Thesis

Inference and Analysis of Boolean Networks considering insilico and real life time-series data

Nina Valery Kersten

Supervisors

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

Advisor

Phd. Robert Schwieger

September 7, 2018

Abstract

goal: create a pipeline: Helps to inference a network of several data sets with the same algorithm (boolsch/bayes/regression....). Analyze the network with PyBoolNet and BoolFilter to figure out the most important parts of a network.

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 disfunctions cause diseases. This work is focused on the inference of a Boolean regulatory network, which is helpful to applicate tools like PyBoolNet and BoolFilter for further analysis of the inferenced network. This chapter describes the different kinds of biological data, what a Boolean Regulatory Network is as well as the algorithms to yield a network and how the performance of a algorithm can be measured.

1.1 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.

Signal Transduction

In the developement of multicellular organisms the action of extracellualr 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)

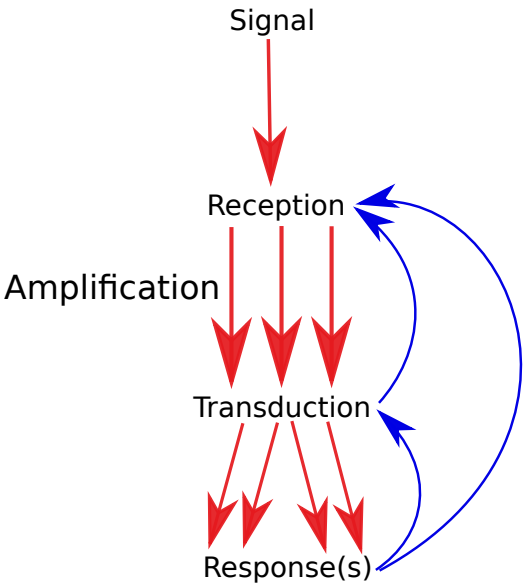

Figure 1.1: Signal Transduction. An environmental signal (e.g. hormone) interacts with a cellular component, most often a cell-surface receptor. The information that the signal has arrived is then converted into other chemical forms, or transduced. The signal is often amplified before introducing a response. Feedback pathways regulate the entire signaling process.(**????**)

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. In order to determine the appropriate nodes of the network some statistical classification of the mRNA expression level data has to be done. The modelling of a transcriptional gene regulatory network ist done by algorithms like Baysian networks, Boolean regulatory networks, Ordinary differential equation models and Neural networks regression

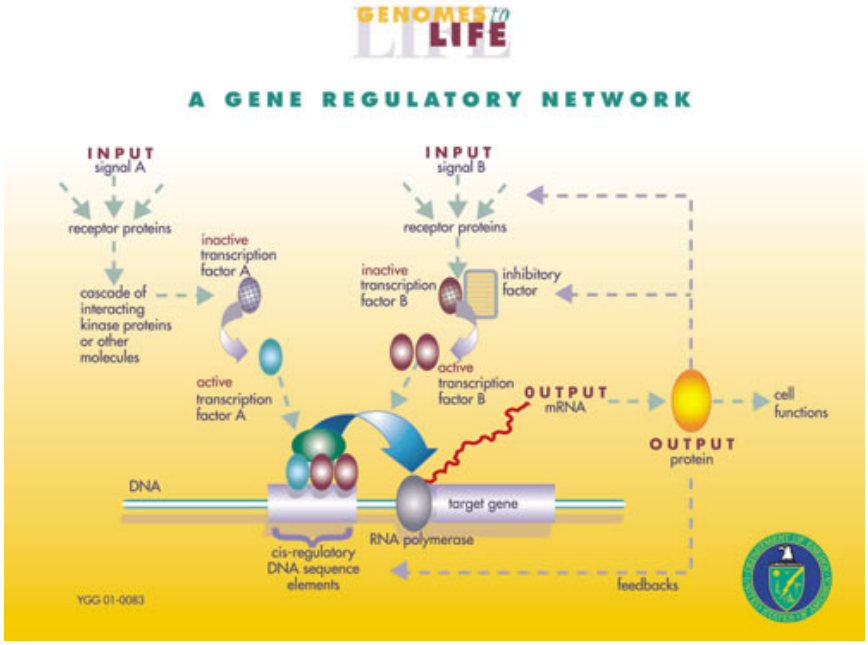

Figure 1.2: Transcriptional Gene Regulatory Network (GRN). In this example two different signals have an impact of a single target gene. Signal molecule A triggers the conversion of inactive transcription factor A (green oval) into an active form that binds directly to the target gene's cis-regulatory 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 . (**?**)

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)

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.(**?**)(**?**)

1.2 Inference Algorithms

1.2.1 Graphtheoretical Background

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).

1.2.2 Boolean network

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).

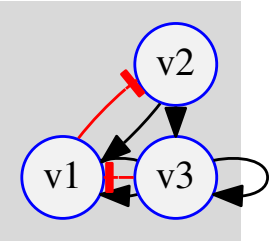

Figure 1.3: Interaction Graph. Shown is a selfinvented Interaction Graph constructed in PyBoolNet. The nodes $v1, v2, v3$ are denoted by blue circles, inhibitory arcs are red and activating arc are black.

$$f(v_{1,2,3}) = \begin{pmatrix} v_1 & \wedge v_2 & \wedge \neg v_2 \\ \neg v_1 & & \\ & v_2 & \wedge v_3 \end{pmatrix} \tag{1.1}$$

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.

$$f(v) = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \text{ where } v_{1,2,3} \in \{0,1\} \tag{1.2}$$

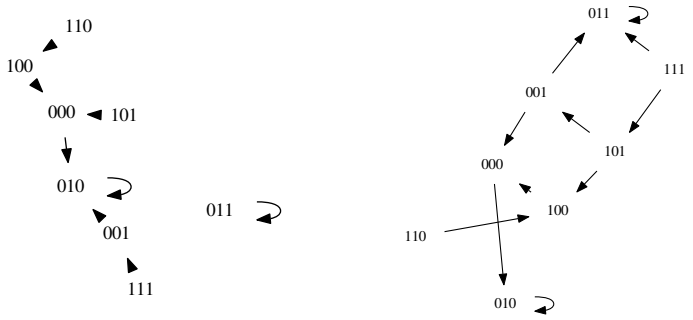

Figure 1.4: Synchronous and Asynchronous State Transition Graph. Left: Synchronous State Transition Graph; Right: Asynchronous State Transition Graph

(ÃĽlisabeth Remy et al. 2008)
(Chai et al. 2014)

1.2.3 Baysian networks

1.2.4 Ordinary differential equation models

1.2.5 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)

1.3 How is the performance of the network inference algorithms measured?

2 Materials and Methods

2.1 Data selection

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. Two DREAM-Challenges appear useful for this work. The first is the DREAM5- Network Inference Challenge (DREAM5) and the second the HPN-DREAM brest cancer network inference challenge (DREAM8)(**?**).

2.1.1 DREAM5

Living cells are the product of gene expression programs involving regulated transcription of thousands of genes. Gene expression programs depend on recognition of specific promoter sequences by transcriptional regulatory proteins. How a collection of regulatory proteins associates with genes across a genome can be described as a transcriptional regulatory network. This map of the transcriptional regulatory network describes potential pathways bacteria cells (S.aureus, E.Coli, S. cerevisiae) can use to regulate global gene expression programs (Lee et al. 2002). The DREAM5-Challenge states the challenge to infere a complete (genome-scale) transcriptional regulatory network for four organisms: in silico, S.aureus, E.Coli, S.cerevisiae.
The data is yielded by genechip experiments, where mRNA is extracted from the samples, then reversetranscriped into more stable cDNA (complementary DNA), which is fluorescent labeled. Afterwards the cDNA binds to the complementary bases on the chip. Thus the

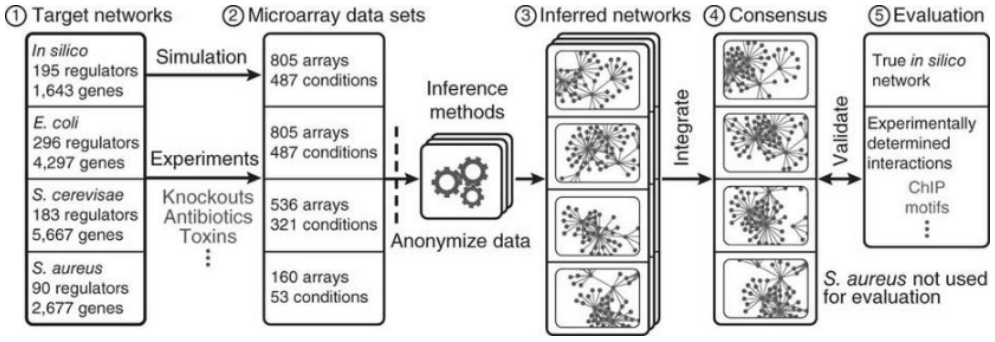

Figure 2.1: Assessment involved the following steps Assessment involved the following steps

DNA composition can be detected and it can be recognized wether a gene is active or not by the information of its transcript. Just the gene expression profiles for the in silico network were derived from another platform so-called GeneNetWeaver (**?**).

In this challenge some experiments have pertubations in terms of gene deletion, overexpressed genes and adding drugs to the system. Few experiments were done the same konfiguration several times. Some experiments are part of a time-series measurement and some are not,described in Figure 2. For each network, a list of directed, unsigned edges had to be submitted ordered according to the participants confidence scores. The participants are given four microarray sets. In each set are three tsv.-files (tab-seperated values) with the gene expression data, chip-features and transcription factors.

```
Network_i_\_expression\data.tsv
Network_i_\_chip\_features.tsv
Network_i_\_transcription\_factors.tsv
```

Chip	Experiment	Pertubations	PertubationLevels	Treatment	DeletedGenes	OverexpressedGenes	Time	Repeat
1	1	NA	NA	NA	NA	NA	NA	1
2	1	NA	NA	NA	NA	NA	NA	2
3	2	NA	NA	NA	NA	NA	NA	1
4	2	P1	0.5	NA	NA	NA	NA	1
5	2	P1	1.0	NA	NA	NA	NA	1
6	3	NA	NA	NA	NA	NA	0	1
7	3	NA	NA	NA	NA	NA	30	1
8	3	NA	NA	NA	NA	NA	60	1
9	3	NA	NA	NA	G5	NA	30	1
10	3	NA	NA	NA	G5	NA	60	1
11	4	NA	NA	NA	G5,G8	NA	NA	1
12	5	P2,P3	NA	NA	NA	G4	NA	1
13	5	P2,P3	NA	1	NA	G4	NA	1

Table 2.1: Chip information

The `Network_i_\_expression\data.tsv` is a matrix where each entry (i, j) describes the expression value of gene j (column) in chip i (row). The data has been normalized,such that the values are compareable across the experiments. For further computation it is necessary to dicretesize the data.
The `Network_i_\_chip\_features.tsv` are shown in Table 2. where every row returns an identifier for the experiment that this chip is part of, then information about added drugs (Pertubations) and the dosage of the drugs (Pertubation Level). The Treatment shows the type of treatment used to apply the pertubations. Additionally are 2 columns showing Deleted genes and Overexpressed genes. Time-scaled data is shown in the column Time and Repeat hepls to distinguish experimental replicants. Whenever an entry NA occurs, none of these pertubation properties mentioned above have been done. The file `Network_i_\_transcription\_factors.tsv` is a list of genes of the network that are potential transcription factors for the network i (where $i \in \}1, 2, 3, 4\{$). Only these transcription factors should be included as regulators in the submitted network.

The predicted network is submitted with no more than 100.000 regulatory link predictions. These predictions are ordered from the most reliable to the last reliable prediction and put in a tab-seperated column format.
Organism specific gold standards containing the known transcription factor to target gene (transcription factor-target gene) interactions (= true positives) were compiled for assessing the participating approaches. All transcription factor-target gene pairs that are not part of the gold standards are seen as negatives, although, as the gold standards are based on incomplete knowledge, they might contain yet unknown true interactions. For evaluation of the predicted networks the AUPR (Area Under Precision Recall), AUROC (Area Under Receiver Operating Characteristic) and an ove-all accuracy was used.

2.1.2 DREAM8

2.1.3 Comparision DREAM 5 vs. DREAM8

The DREAM5 challenge deals with high-throushput data of a microarray set in contratst to the DREAM8 challenge which deals with PPIs (protein-protein interactions). But in the DREAM5 challenge time-series data is mixed with data where no time was captured. Furthermore the experiments in DREAM5 contain a lot of additional information (pertubations by drug, gene deletion,overexpression of genes,replicate experiments) which is not needed in this thesis.

Preprocessing of the data

(Elena S. Dimitrova 2010) (Berestovsky & Nakhleh 2013)

2.2 PyBoolNet and BoolFilter

3 Results

3.1

3.2

3.3

4 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 .....)