Gene regulatory network inference in human pathogenic fungi

Dissertation
zur Erlangung des akademischen Grades
- doctor rerum naturalium -

vorgelegt dem Rat der
Biologisch-Pharmazeutischen Fakultät der
Friedrich-Schiller-Universität Jena

geschrieben von Dipl.-Bioinf. Robert Altwasser
geboren am 06.07.1984 in Luckenwalde

Diese Doktorarbeit wurde von den folgenden Personen begutachtet:

- Prof. Dr. Reinhard Guthke
- Prof. Dr. Marc Thilo Figge
- Dr. Ronald Westra

Die Dorktorarbeit wurde am 14. Januar 2015 verteidigt.

There is a single light of science, and to brighten it anywhere is to brighten it everywhere.

Isaac Asimov

Danksagung

Ich möchte den Abschluss meiner Doktorarbeit nutzen, um mich bei den Menschen zu bedanken, die meine Arbeit möglich gemacht haben. An erster Stelle steht hier Prof. Reinhard Guthke, welcher mir die Stelle angeboten, und mir die Werkzeuge in die Hand gegeben hat, die vor mir liegenden Aufgaben zu meistern. Es war leicht ersichtlich das ich bei einem Menschenfreund arbeiten durfte, der sich ehrlich um das Wohlbefinden seiner Mitarbeiter bemüht.

Ich hätte diese Arbeit auch nicht bewältigen können, ohne die umfangreiche Unterstützung meiner Arbeitsgruppe. Die lockere Arbeitsatmosphäre hat immer zu einem Gelingen der Arbeit beigetragen. Insbesondere ist hier natürlich Jörg Linde zu erwähnen, der mich während dieser Zeit betreut und angeleitet hat, was seine eigene Arbeit sicherlich nicht leichter gemacht hat. Bei Fabian Horn, der mich ein Jahr lang als Zimmerkollegen ertragen hat, möchte ich mich genauso bedanken wie bei Sebastian Vlaic, der mit mir zusammen Florida unsicher gemacht hat. Auch Eugen Fazius, der während der ersten zwei Jahre um alle Hardware-Probleme gekümmert hat und auch sich auch sonst um eine gute Arbeitsatmosphäre bemüht hat, soll hier nicht unbedankt bleiben. Auch bei Dr. Vito Valiante, der die biologischen Experimente für mich durchgeführt hat, möchte ich mich hiermit bedanken. Die anderen aktuellen und vergangen Kollegen werde ich natürlich auch nicht vergessen, da sie ebenfalls Teil dieses Gesamtprojekts "Doktorarbeit" sind.

Da ich trotz allem Interesse an der Wissenschaft nicht umsonst arbeiten kann, möchte ich mich bei der "Jena School for Microbial Communication (JSMC)" bedanken, die sowohl das Geld für mein Gehalt, als auch für meine Auslandseinsätzung und nötigen Materialien zur Verfügung gestellt haben. Die Jenaer Graduierten-Akademie (JGA) bot mir viele Kurse an, die mich sowohl fachlich weiterbrachten, als mir auch den Austausch mit andere Wissenschaftlern ermöglichten.

Natürlich wäre ich nicht hier, ohne meine Freunde und meine Familie, die in Gedanken immer bei mir waren. Die seelisch-moralische Unterstützung war während der, nicht immer leichten, Doktorarbeitszeit nötig, und soll hier daher nicht unerwähnt bleiben.

Bei allen Anderen, die in den letzten drei Jahren in mein Leben getreten sind, und mich auf dem Weg zum Doktor begleitet haben aber hier noch nicht genannt wurden, möchte ich mich nocheinmal bedanken. Diese Dokument ist nicht das Werk eines einzelnen, sondern das Ergebnis der Arbeit vieler.

Abstract

Pathogenic fungi are a serious threat to people with impeded immune system, especially during organ transplantation and HIV infections. As the number of treatments that include a weakening of the patients immune system increase, so does the number of fungal infections. Often, the infection is opportunistic, meaning the pathogen already lives as a commensal in the host and uses the weak immune system to spread out and starts to colonise different parts of the host. These infections can lead to systemic, life-threatening infections, lowering the survival rate of the often already weakened host.

Two of the most common human pathogens are *Candida albicans* and *Aspergillus fumigatus*. While *C. albicans* is a commensal and part of the healthy human flora, it can turn to an opportunistic pathogen, once the hosts immune system fails to contain it. Conidia of *A. fumigatus* are inhaled by humans every day and removed again by the immune system. In a weakened host, *A. fumigatus* can colonise the lung of the host and spread to other parts of the body, which can lead to fatal results, if no treatment is administered.

The first part of this thesis aims to study the gene regulatory network of *C. albicans* on a genome-wide level, with a scale-free distribution of node degrees. These networks can be used to identify genes with central regulatory functions, called hubs, which are possible drug targets and can be the starting point for future studies. The modeling process included a large set of gene expression data measured by microarrays, the use of prior knowledge and a automatically harvested gold standard for the evaluation of the results. The final model is used to identify several hubs and is also able to reproduce current knowledge.

A focused small-scale gene regulatory network is inferred for *A. fumigatus* while it is treated with the clinically applied drug caspofungin. The chapter describes the process from mapping of the RNA-Seq data over the selection of candidate genes and the harvest of prior knowledge to the application of the NetGenerator tool. A network model of 26 genes is tested for robustness against noise and used to identify a so far unknown cross-talk between to key regulators of major drug response pathways in *A. fumigatus*, which could be experimentally verified by the collaboration partner.

Both, the large- and the small-scale network inference are later compared to give guidance on the correct application depending on the scientific question.

To further study the influence of drug treatment on *A. fumigatus* caspofungin treatment was paired with the use of humidimycin, which does not have antifungal properties on its own, but seems to enhance the effect of caspofungin. Analysis of differential expression and clustering revealed that the combination of the two drugs lowers the number of differentially expressed genes in *A. fumigatus*, giving hints on how the enhancing effect of humidimycin works on the genetic level.

Zusammenfassung

Pathogene Pilze stellen eine ernste Bedrohung für Menschen mit geschwächtem Immunsystem dar. Die betrifft insbesondere Menschen während Organtransplantationen und HIV-Infizierte. Mit der steigenden Anzahl von Behandlungen, bei denen eine Schwächung des Immunsystems einhergeht, steigt auch die Anzahl der Pilzinfektionen. Diese sind häufig opportunistisch, was bedeutet, das der Pathogen bereits als Nutznießer im Wirt lebt und ein geschwächtes Immunsystem nutzt, um sich auszubreiten. Dies kann zu systematischen, lebensbedrohenden Infektionen führen, welche die Überlebenswahrscheinlichkeit des oft bereits geschwächten Wirts weiter senkt.

Zwei der am weitesten verbreiteten Pathogene sind *Candida albicans* und *Aspergillus fumigatus*. Während *C. albicans* gewöhnlich als Teil der gesunden menschlichen Flora lebt, ohne Schaden anzurichten, kann es sich zu einem opportunistischem Pathogen entwickeln, sobald das Immunsystem des Wirts ihn nicht mehr eindämmen kann. Sporen von *A. fumigatus* werden von Menschen jeden Tag eingeatmet und vom Immunsystem wieder entfernt. In einem geschwächtem Wirt kann *A. fumigatus* die Lunge besiedeln und sich auf andere Teile des Körpers ausbreiten. Ohne Behandlung kann dies tödliche Folgen für den Wirt haben.

Der erste Teil dieser Doktorarbeit zielt auf die Untersuchung der genregulatorischen Netzwerke von *C. albicans* auf genomweiter Ebene ab. Dabei wurden Netzwerke mit einer skalenfreien Verteilung der Kantengrade erzeugt. Diese Netzwerke können dafür verwendet werden, Gene mit zentraler regulatorischer Funktion zu identifizieren. Diese so genannten Hubs sind mögliche Zielgene für Medikamente und können der Anfang für zukünftige Studien sein. Die Modellierung enthält die Verwendung von Vorwissen und ein automatisch gesammelter Goldstandard zu Evaluierung der Ergebnisse. Das endgültige Modell wird benutzt um verschiedene Hubs zu identifizieren und ist auch in der Lage, aktuelles Wissen wiederzugeben.

Darüber hinaus wird ein fokussiertes genregulatorisches Netzwerk für *A. fumigatus* erstellt, während es mit dem klinischem Medikament Caspofungin behandelt wird. Hier beschrieben wird der Vorgang von der Kartierung der RNA-Seq-Daten über die Auswahl der Kandidatengene und das Sammeln von Vorwissen zu der Anwendung des NetGenerator Programms. Ein Netzwerkmodel aus 26 Genen wird bezüglich seiner Robustheit gegen Rauschen in den Daten und fehlendes Vorwissen getestet. Dabei wird eine bisher unbekannte Regulation zwischen zwei zentralen Genen gefunden, welche für die Stressantwort gegen Medikamente in *A. fumigatus* verantwortlich sind. Diese Regulation konnte experimentell durch Kollaborationspartner bestätigt werden.

Sowohl die genomweite, als auch die fokussierte Netzwerkinfernz werden anschließend verglichen, um Hinweise für ihre korrekte Anwendung zu geben, abhängig von der biologischen Fragestellung.

Um den Einfluß von Medikamenten auf *A. fumigatus* weiter zu untersuchen, wurde die Kombination von Caspofungin mit Humidimycin untersucht. Humidimycin besitzt selbst keine antifungielle Wirkung, scheint jedoch die Wirkung von Caspofungin zu verstärken. Eine Analyse der differentiell exprimierten Gene und Clustering zeigte, das die Kombination beider Medikamente die Anzahl der differentiell exprimierten Gene gegenüber der Einzelbehandlung mit Caspofungin verringert. Dies gibt Hinweise darauf, wie der verstärkende Effekt von Humidimycin auf Genebene funktioniert.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Pathogenic fungi

The advancements in modern medicine offer new hope for formerly terminal ill patients. Complex medical and surgical treatments increase the life expectancy of patients that suffer from diseases like organ failure, cancer or HIV-infection. During the treatment, it is often inevitable, and sometimes intended, to weaken the immune system of the patient, or to perforate protective barriers of the body. Those breaches do not go unnoticed to potential pathogens. Every day, the human host is attacked by countless spores of different fungi. But not all are agents from the outside. Some fungi are constitute part of the human flora. In healthy hosts, they are not able to cause an infection. However, patients that undergo an extensive medical treatment or suffer from a severe illness are highly susceptible to hospital-acquired (nosocomial) fungal infections [90]. From 1979 to 2000, the number of sepsis cases in the USA caused by fungal organisms increased by 207% [76]. In 2007 the EPIC II study investigated the infections of 14414 patients in 1265 intensive care units (ICUs). It revealed that 19% of pathogens isolated in ICU patients were fungi [125]. The Candida species was by far the most common fungal pathogen in ICU patients, followed by the Aspergillus species. Successful treatment of the infection can be hindered by late diagnosis and the development of drug resistances by the fungus. Other risk factors include, but are not limited to: venous catheter or burns that disrupt the human skin and create an entry, or multiple site colonisation. Also advanced age, malnutrition and Diabetes mellitus are risk factors [87]. In European countries, the trend seems to be ambiguous. Countries like the Netherlands, Iceland and Finland reported an increase in Candida blood stream infections [8, 92, 126]. On the other hand, reports from Switzerland, Norway or Germany [74, 77, 100] showed no increase in the number of infections. A comprehensive view is still out of reach, since most European studies are focused on specific groups of patients or selected hospitals. In developing countries, the HIV epidemic is one of the major factors for invasive fungal infections. Without treatment, over 80% people with HIV-infection contracted an infection by an opportunistic fungus [129].

1.1.1 Candida

The most common form of nosocomial fungal infection was a bloodstream infection by the Candida species [87]. It is a well-recognised cause of mortality among ICU patients. It is difficult to distinguish between a death caused by a fungal infection and the underlying disease. That is one reason, why the attributed mortality in different studies

varies greatly, ranging from 5 to 71%. However, Candida species is capable of a broad spectrum of diseases, including invasive Candidasis [52,91] and hepatosplenic candidasis [120]. By far the most common representative of the Candida species is *Candida albicans*. According to estimates, it can be found in half of the worlds population [50]. It is called an opportunistic pathogen, because most of the time, it lives a harmless commensal as part of the hosts flora. However, should the conditions change, for example by long-term antibiotic treatment or a compromised immune system, the fungus can switch to pathogenic behaviour [132].

Superficial infections on the skin or the mucous membranes can occur also in immunocompetent patients [98]. It is recognised that these infections are often chronic and recurring. As an example, approximately 15% of the population has a fungal infection on the skin or nails of the feet [20].

An important tool that Candida albicans uses to counter the hosts defences is the ability to form hyphae. As a polymorphic fungus, it is able to grow in yeast, hyphal or pseudo-hyphal form [111]. The hyphal form gives *C. albicans* the ability to enter the blood stream, by penetrating the epithelia and endothelia. Once *C. albicans* entered the bloodstream, it can cause a systemic infection by colonising various organs like brain and lungs. Other important virulence factors include: adherence to mucosa and biofilm formation, as it supplies resistance to antifungal therapy [34], iron acquisition from intracellular host sources [112] and the ability to survive in oxygen-limited micro-environments [38]. It is also able to react with hemoglobin [99]. All those virulence factors require the ability to react on changing environmental conditions. *C. albicans* facilitates this via complex pathway, that transmit signal from the surface to cell core. There, the activation of transcription factors lead to altered gene expression as a response to new condition.

Candida is a yeast belonging to the hemiascomycete group. The most popular representative of this group is *Saccharomyces cerevisiae*. One origin of its popularity roots in the fact that it is used for baking and brewing for thousands of years, giving it the name “baker’s yeast”. Apart from that, it can be very easily manipulated on the genetic level. It’s ability to grow haploid makes it comparable easy to create gene knock-outs. Because of this, *S. cerevisiae* became one of the main model organisms for eukaryote organisms in general. It was also the first eukaryote organism to become fully sequenced in 1996 [37]. As an effect, many references for the Candida species root from orthologous genes of *S. cerevisiae*. Many tools and procedures used for the study of *S. cerevisiae* are also adopted for the use in the investigation of Candida species.

In an attempt to investigate the genetics of *C. albicans*, the Stanford Genome Technology Center started sequencing its genome [24]. It took ten years before the assembly of *C. albicans*’s eight chromosomes were released. The length of the chromosomes varies from 0.95 - 3 megabases. In total, *C. albicans*’s genome consists of 16 megabases [24]. The Candida Genome Database [103] makes sequencing data publicly available. Once the genome sequence of *C. albicans* was known, microarrays have been developed to investigate its transcriptome.

Despite *C. albicans* being a model organism among fungal pathogens, it has two special features that makes genetic investigation difficult. First, *C. albicans* is a diploid species

without a sexual cycle including a haploid phase. The creation of knock-out mutants is therefore difficult and tedious. Another interesting property of *C. albicans* genome is that the triplet CUG is translated to serine instead of leucine. This prevents the use of standard reporter genes. The development of new reporters for *C. albicans* and other Candida species which share this property is necessary. Instead of a sexual cycle, *C. albicans* has a parasexual one. The phenotype changes from a white to an opaque state and is controlled by a mating-type loci. The influence of this parasexual cycle in pathogenesis will need to be further investigated as potential virulence factor.

There are also other Candida species capable of infecting the human host, like *C. glabrata*, *C. dubliniensis* or *C. tropicalis* [64]. Together with *C. albicans* and other, non-pathogenic species, comparative studies can unravel the pathogenicity of Candida species [27].

1.1.2 Aspergillus

The fungi of the Aspergillus species can be found in various environments all over the world [14]. It recycles carbon and nitrogen in its ecological niche that consists of soil and decaying vegetation, which is called a saphorytic lifestyle. The decomposition of organic matter is an exothermic reaction, which can increase the temperature of the environment. This leads to the development of heat resistance in many saphorytic organisms, which is beneficial when invading human hosts. *A. fumigatus* developed the ability to grow in temperatures above 30°C [1]. From its ecological niche, it proliferates using small conidia get carried away by air. According to estimates, the human body inhales several hundred of these conidia per day [67]. While these do not pose a threat to humans with intact immune system, immunocompromised patients can suffer a life-threatening systemic infection [14]. After the Candida species, Aspergillus moulds are the second fungal pathogens found most often in ICU patients. Among different pathogens in the Aspergillus species, including *A. niger*, *A. flavus* and *A. terreus*, *A. fumigatus* is by far the most prominent. It is regarded as the most important airborne fungal pathogen.

Common sources of Aspergillus in the ICU are improperly cleaned ventilation systems, water systems and computer consoles [87]. Clinical symptoms are often not specific, making it difficult to recognise the infection. Sometimes, an autopsy is necessary to confirm a diagnosis. Depending on the immune status of the patients, different infection loci can occur [22]. In immunocompetent patients, mucociliary clearance and phagocytic cells prevent infection [15]. An impaired lung function like asthma can lead to bronchopulmonary aspergillosis. Tuberculosis patients are susceptible to non-invasive aspergillomas, if they are repeatedly exposed to conidia. Among others, patients which suffer from leukemia, organ or stem cell transplantation have a heightened chance of invasive aspergillosis, possibly the most severe form of Aspergillus related infections.

2005, Nierman *et al.* published the complete genome sequence of *A. fumigatus* strain [84]. In their study, they used the clinical strain Af293. It consists of eight chromosomes with 29.4 megabases. In their study, Nierman *et al.* compare the genome with those of *A. oryzae* or *A. nidulans*. Even though, these fungi are of the same genus, their evolutionary distance is as far as the one between man and fish [114]. In 2008, Fedorova

et al. published a second *A. fumigatus* genome sequence, this time on the clinical strain A1163. This was done in an attempt to investigate genetic traits such as sexual cycles and virulence. One result of the study was that 8.5% of the genes in *A. fumigatus* are lineage-specific, i.e. genes with limited phylogenetic distribution of orthologous genes in related species. Another important step for genetic investigation was the creation of ku70 [63] and ku80 [21] knock-out strains, which did not show a difference in phenotype, but facilitated easy creation of additional knock-outs.

1.2 Systems biology

The switch from commensal to pathogenic behaviour of *C. albicans* or the impressive adaptive capabilities of *A. fumigatus*, growing in soil and human body alike, are just two examples of how organisms are able to adapt to environmental changes. A major goal in biology is to understand the nature of these changes in phenotype and behaviour.

A basic principle of genetic responses in organisms, is that genes usually do not work "on their own", but interact with each other, forming complex networks of different types. This is necessary to govern the various processes an organism needs to survive in a changing environment. Before the dawn of microarrays and later Next-Generation-Sequencing (NGS), scientists investigated the biology of one gene or protein at a time. From these single information, detailed biomolecular models have been constructed, that are at the same time accurate and reliable. It soon became obvious that investigating each gene one by one is neither practical, nor will it be able to explain the complexity of biological regulation in a cell. The same way, that the whole is greater than the sum of its parts, an organism can not be explained by looking at each part independently. Organisms are complex systems in which all components must be seen in regard to the other components. This is the basic principle of a field in biology called systems biology. To understand the dynamics and structure of whole organisms, even only on the single cell level, requires extensive knowledge of different fields of science, especially mathematics, to cope with probabilities and separate random correlation from significance, and informatics, whose graph theory is the perfect platform to understand the connection between different parts of cells.

A systematic analysis of multiple regulatory components of a cell, for example genes or proteins, requires huge amounts of data. Depending on which level of regulation is to be analysed, different data has to be collected. Information on the genome is transcribed into gene products like RNA and can be translated into proteins. Again, genes do not act alone but influence other genes in their transcription. By studying the expression pattern of various genes, these connections can be unraveled. This knowledge can for example be used to increase the excretion of desired natural products or to inhibit pathogenic traits.

Technologies like the microarray, and later NGS provide this data. After the genome sequence is known, microarrays can be used to examine the transcriptional activity of genes that forms the basis of gene regulation studies. The first study with 1000 human genes was conducted by Schena *et al.* in 1996 [101]. Two years later, Eisen *et al.*

published the first genome-wide study of expression patterns [29]. Later, fungi-centered studies followed [51, 82]. The invention of different platforms of microarrays lead to a low comparability of results. Experience quickly showed that the use of different methodologies concerning sample preparation lowers the comparability of transcription data severely.

2009, Wang *et al.* developed the RNA-Seq technology [128]. It has several advantages over microarray technology, for example, the genome sequence does not need to be known beforehand. It has also been shown that the comparability of different RNA-Seq studies is much higher than among microarrays [83]. The sensitivity of RNA-Seq enables it to detect even small changes in expression.

Using measurements of genes, and later proteins, and their expression to describe complex models is called a "bottom-up" approach. They lead to hypothesis that involve combination of known subsystems and predict the behaviour of inter- and intracellular processes of an organism [31]. The "top-down" approach includes the search for molecular dynamics that can than be verified by new experiments. Due to the limited amount of data available, only specific problems, like drug-response, can be addressed. The regulatory networks in organisms span over various scales like molecules, cells, organ, organism. Some go beyond single organisms in an attempt to model for example host–pathogen interactions [49]. In studies of influenza infections [59], it might even be desirably to integrate knowledge about flight patterns of birds and humans, or transmission efficiency of viruses. As the heterogeneity of the data dramatically increases, it gets more and more difficult to combine the information. Often, the data is only available via supplementary tables of papers. While the data can be visualised using free available tools like Cytoscape [105] or Ondex [61], it can not be combined with orthologous data. In 2011, Kozhenkov *et al.* developed a tool to integrate multi-scale data for that purpose [62].

Despite new technologies, the amount of data available for regulatory network modeling is still insufficient. Especially the combination of transcriptome and proteome date is still difficult [83]. It is clear, that the relation between gene expression and protein production is not a linear one [3]. Because of this, influence networks describe regulation interactions directly between genes. This reduces the amount of data needed but leads to a loss of information. Additionally, the use of heuristics and computer simulations are often necessary among system biological research.

A concept often found in network modeling is that of sparseness [135]. It implies that a regulatory network contains as little connections as possible in order to achieve the necessary regulation. This property is especially important in network modeling, since the number of predictors is usually very high. Selecting only predictors with a high correlation to the measured data lessens the probability of including redundant or noise features. The decrease of connections in the network also makes the interpretation of high dimensional data easier.

1.2.1 Prior knowledge

An approach to increase the amount of available data is to include data from different sources, so called prior knowledge. It has been successfully applied in network inference multiple times [48,133]. Prior knowledge consists of information from other data sources than those directly used in the modeling. This includes regulatory information from other experiments, literature or data bases. Often, the reliability of these sources remains unclear. To deal with unreliable data, prior knowledge is often integrated "softly". The idea is not to "force" information into the model, but give the algorithm favorable interactions. To that end, each prior knowledge has a weight, which represents the reliability of the source. If those interactions do not fit the data, the algorithm can still choose not to implement the prior knowledge. Christley *et al.* could show that offering false prior knowledge to an inference attempt does no decrease the predictive power [19]. The estimation of the prior knowledge weight remains difficult and adds another parameter, that has to be estimated in the model.

1.2.2 Scale freeness

Another desired network property is the so called scale freeness. It was described by Barabási and Albert in 1999 [12]. They investigated the topology of various real-world networks such as co-authorship in science, web graphs or genetic regulatory networks. Barabási and Albert soon realised that the connections between nodes were not equally distributed. Among all nodes were some, that had significantly more connections than one would expect by chance. To be precise, they found that the probability, that a given node has a certain number of interactions, follows a power law distribution: $P(k) \sim k^{-y}$. $P(k)$ is the probability that a node interacts with k neighbors. This results in networks, were most nodes have a very small number of interactions and only a few are highly connected. Those nodes are interpreted as central regulators, called hubs.

The first large scale protein interaction models made it possible to relate the topographic property of a gene or protein with its function [53]. These models were not done using reverse-engineering using transcription data, but connecting already verified interactions. They also showed the scale freeness and the occurrence of central regulatory genes. For those so called **hubs**, Jeong *et al.* coined the centrality-lethality rule. It states that the more connections a gene has, the more essential it is for the organism i.e. the more a deletion of this gene cripples the ability of the organism to grow or proliferate. Jeong realised that this property gives the organism robustness towards mutation, since a knock-out mutation of a random gene will less likely be fatal. It also increases the adaptability of the organism, since expression change in a few genes are sufficient to alter the phenotype drastically.

Several explanations for this phenomenon have been made, one was based on simple statistics: If every connection has the same chance of being essential, genes with many connections have higher chance of possessing an essential connection [47]. This neglects the general topology of the network and focuses on basic statistic. This view was challenged by the theory that hubs increase the connectivity of the network, and mediates

between several less connected genes [53]. This can be observed by measuring the network diameter before and after deleting central hub genes [2]. This assumes, that the viability of an organism is based on its connectivity between several parts of the genome.

On the other hand, Yu *et al.* argued against any correlation between centrality and lethality [136]. He presented evidence that he could not observe this relation in his protein dataset. Rather, he argued, was the centrality-lethality rule an artifact, caused by an investigation bias towards essential and well-studied proteins, i.e. the more information is available for a protein or gene, the more likely it is to be used in further studies. Following that, [86, 138] argued that there certainly is a correlation between centrality and lethality, but with a different explanation that the previously mentioned. They argued that the importance of hubs is not based on connectivity over large parts of the network, but because of their role in "essential complex biological molecules". Those are clusters of tightly connected genes that have a similar biological function and some form large multi-protein complexes, like regulation of transcription.

1.3 Network inference

A general concept of reverse engineering of gene expression is, that the expression pattern of a gene is the result of the expression of the other genes in the organism. This is of course a simplified view, since genes do not regulate each other directly, but via complex regulatory pathways, and the connection of different pathways is not always linear. As mentioned before, the available data is often insufficient to generate a multi-layer network model. The assumption, that genes with correlated expression pattern are similar regulated is a reasonable thought and was already successfully applied to predict gene regulatory networks [48].

There are different mathematical models to simulate the gene expression pattern. One is based on the correlation of expression pattern [110]. The fundamental idea is that statistical correlation between the expression pattern of two genes, that can not be explained as artifact of expression profiles of other genes, are assumed to interact. Often, a threshold is applied on the correlation. The higher the correlation is, the more certain the the prediction. An early limitation was the fact that the networks were always undirected, i.e. there is no way to tell source from target gene of an interaction. This changed with the introduction of time–delayed network inference in mutual information networks [137].

Another modelling approach is used by Boolean networks [58]. In a Boolean network, a node can have the state 0 or 1 (e.g. expressed or not expressed). The nodes are connected by logical Boolean operators like AND, OR or NOT. Since the gene states are always discrete, the continuous expression data has do be transformed to binary data. This limits their predictive power, since a lot of information is lost in the simplification of expression. Compared to other models, their predictions are easy to interpret.

A probabilistic approach is the network modeling via Bayesian networks [33]. The idea is to regard gene expression as random variables, that follow a certain probability

distribution. The connections between the nodes is estimated via Bayes rule[1]. They are very well able to deal the randomness and noise that accompanies every gene expression measurement. Bayes rule makes it comparable easy to include prior knowledge. Every modeling process starts with the selection of a template, for which the network probabilities are calculated. Later, different templates and their probabilities are evaluated. The selection of a template is a necessary weak spot in this method. Since the number of possible network structures increases exponentially, enumerating all possible templates is not feasible and heuristics have to be applied.

The simulation via differential equations is a quantitative approach. Here the expression of a gene is described as the direct function of all other genes, plus an outside perturbation. I want to mention explicitly, that "outside perturbation" in this case means outside of the model, not necessarily outside of the cell. It also includes for example the influence of genes that are not part of the gene regulatory network. The mathematical description of a linear differential equation is:

$$\frac{dx_i}{dt} = \sum_{j=1}^{N} \beta_{i,j} x_j + b_i u \tag{1.1}$$

The expression profile x_i is multiplied by $\beta_{i,j}$, which is an element of the interaction matrix B, and describes the influence of predictor x_j on x_i. Additionally, u refers to the perturbation and b_i its influence on x_i. The interaction matrix B later describes the model and its connections. In practice, there are a lot more genes than measurements, which makes the model under-determined. There are infinitely many solutions for this system, which makes it ill-posed. To make this system solvable, heuristics and constraints are applied, for example a threshold on the sum of coefficients or that the relationship between genes is linear.

1.3.1 Linear regression

Differential equations can be approximated by ordinary difference equations (ODEs). The idea of linear regression is based on these ODEs with the assumption that there is a linear relationship between one gene x_i and the expression of all other genes:

$$x_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} \beta_{i,j} x_j \tag{1.2}$$

where N is the number of examined genes and $x_j = x_j(1), \ldots, x_j(M)$ is the expression of gene j in the experimental condition 1 to M. $\beta_{i,j}$ is the coefficient matrix, describing the influence of gene x_j on the expression of gene x_i. The network is defined via the coefficient parameters stored in β. Each $\beta \neq 0$ represents an interaction between two genes. Positive values are activating, negative values repressing. The restrain on linear models is often not enough to find unique solutions to the equation, so often additional

[1] $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

constrains are applied. One is to have many $\beta = 0$, to get sparse networks. Sparse networks are generally more reliable, since only predictors with a strong impact in the measurements are considered for the model. This makes it less susceptible for noise in the data. Another issue is the so called over-fitting, which occurs when there are a lot of degrees of freedom compared to the measurements. It can happen that the algorithm selects predictors whose high correlation to the target is only coincidence, or have only a very small impact on the expression of the target.

1.3.1.1 Ridge regression

One of the most common methods to solve ill-posed problems is called "Tikhonov regularisation", also known as ridge regression [89]. In order to give the ODE a single solution, it puts a restriction μ on the L_2-norm[2] of coefficients:

$$\sum_{\substack{j=1 \\ j\neq i}}^{N} \beta_{i,j}^2 \leq \mu \tag{1.3}$$

The Residual Sum of Squares (RSS) is then minimised:

$$\arg\min_{\beta} \sum_{i=1}^{N} (x_i - \sum_{\substack{j=1 \\ j\neq i}}^{N} \beta_{i,j} x_j)^2 + c \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \beta_{i,j}^2 \tag{1.4}$$

the variable λ puts a penalty on the sum of coefficients, which forces the algorithm to shrink them. Since the sum of coefficient is squared before the threshold is applied, it more is preferable to decrease the coefficient with high values. In practice, this often leads to networks with all possible predictors having low values. There is no parameter selection, which makes the model difficult to interpret.

1.3.1.2 LASSO

If sparseness of a model is an issue in the network modeling, as it is mostly the case in regulatory network inference, the LASSO may be a more suitable choice. LASSO stands for Least Absolute Shrinkage and Selection Operator and was introduced by Tibshirani *et al.* [116] in 1994. It works very similar to the ridge regression by using a threshold μ to keep the coefficients small. Here, this threshold is applied on the L_1-norm of coefficients

$$\sum_{\substack{j=1 \\ j\neq i}}^{N} \mid \beta_{i,j} \mid \leq \mu_i \tag{1.5}$$

and minimises:

[2] Lp-norm $(x.) = (\sum_j |x_j|^p)^{1/p}$

$$\arg\min_{\beta} \sum_{i=1}^{N}(x_i - \sum_{\substack{j=1 \\ j\neq i}}^{N} \beta_{i,j}x_j)^2 + \lambda \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} \mid \beta_{i,j} \mid \tag{1.6}$$

The algorithm treats all coefficients equal, disregarding their absolute value, when it comes to parameter shrinkage. This often leads to smaller coefficients being removed from the model first, filtering the parameters to those with the highest influence on the model. While this is desired in most regulatory network inferences, it can also lead to problems. When correlation among several different predictors is very high, like for example genes with similar biological mode of action, LASSO tends to select only one gene and omits the others.

This can be problematic, since genes in the same functional cluster often have similar expression patterns and therefore a high correlation. These clusters may stay be hidden, since LASSO only selects a few of them, unable to uncover the connection.

2006, Zou *et al.* enhances the algorithm with the weighting parameter $\omega_{i,j}$ [139]. It weights the coefficients in equation 1.5 individually, so the user can influence the parameter selection

$$\sum_{\substack{j=1 \\ j\neq i}}^{N} \omega_{i,j} \mid \beta_{i,j} \mid \leq \mu_i \tag{1.7}$$

This version is called adaptive LASSO and is used to incorporate prior knowledge. Interactions that the prior knowledge suggests, receive a lower weight and do have less influence on the calculation of the threshold, making it less likely to be omitted from the model.

1.3.1.3 LARS

The calculation of a LASSO solution is computationally demanding. In 2004, Efron *et al.* [117] presented the Least Angle Regression (LARS). It is a less greedy version of forward selection methods. The algorithm starts with selecting the predictor x_j with the smallest angle between the predictor and the response variable x_i. Then LARS proceeds in that direction until the angle between x_j and the vector of the residual $x_i - \beta x_j$ is smaller than the angle between the residuals and other predictors. At the point, where another predictor x_k enters the model, LARS moves in the direction of the least-squares fit of (x_j, x_k) until a third predictor becomes part of the model and so on. Figure shows the steps LARS takes for an example of two coefficients.

LARS can be modified that it produces similar outputs as LASSO, while being computationally less demanding. A reason for this is that, once a predictor entered the model, LARS keeps it part of the model. This means that the LARS algorithm reaches the full model after at most m steps, with m being the number of predictors. LASSO on the other hand lets predictors leave and enter the model multiple times. Therefore, calculation of the full model can take more than m steps.

1.3.1.4 Combination of ridge regression and adaptive LASSO

One of the most important network properties is its size, i.e. the number of connections between the genes. It has influence on the network topological properties like scale-freeness and of course sparseness. In the LASSO algorithm, the number of predictors for a target is indirectly regulated by the μ_i variable in formula 1.7. It is an upper limit to the absolute sum of coefficients for the target gene x_i. As described above, the LASSO algorithm is very competent at selecting predictors in the model, yet is sometimes to greedy and misses predictors that belong to the same functional cluster. Ridge regression on the other hand is able to identify these functional clusters and achieve a good fit, but generally selects to much predictors for a model to be considered sparse. In order to use the advantages of two worlds, Gustafsson *et al.* combined the algorithms [40, 41]. Following his approach, I first computed the solution for the ridge regression according to formula 1.4. For each gene, I calculated the threshold μ_i^{ridge}, which is the sum of coefficients for a given gene i in the solution:

$$\mu_i^{ridge} = \left(\sum_{\substack{j=1 \\ j \neq i}}^{N} (\beta_{i,j})^2 \right)^{\frac{1}{2}} \tag{1.8}$$

μ_i^{ridge} represents the calculated influence the predictors should have on the target. It now serves as an upper limit μ_i^{lasso} for the LASSO regression to decrease the number of predictors in formula 1.7. In practice, we found that this upper limit is still to high, as still to many predictors are part of the model. Gustafsson *et al.* suggested another scaling factor c with:

$$\mu_i^{lasso} = c\mu_i^{ridge} \tag{1.9}$$

He fixed c to the value of 0.1 and found the results reasonably sparse and changes of c do not cause large deviations in the results.

1.3.2 Mutual information

A different approach to infer networks are the *mutual information networks* [78]. They derive the network structure by calculating the mutual information of different expression patterns. Mutual information is a non-linear measure of dependency and therefor provides a natural generalisation. In the information theory, the mutual information between X and Y it is defined as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \left(\frac{p(x,y)}{p_1(x)p_2(y)} \right) \tag{1.10}$$

From this, a symmetric *Mutual Information Matrix* (MIM) can be constructed

$$MIM_{i,j} = I(x_i; x_j) \tag{1.11}$$

Here, the element i, j represents the mutual information between x_i and x_j. When the mutual information is above a certain threshold, an interaction is assumed. This approach was called *relevance network* by Butte *et al.* in 2000 [16]. This method does not eliminate indirect interactions between genes. If, for example, gene x_1 regulates the genes the genes x_2 and x_3, the mutual information between (x_1, x_2), (x_2, x_3) and (x_1, x_3) would be high. Since the algorithm sets edges between nodes with high correlation, it will create a connection between x_2 and x_3 as well.

1.3.2.1 ARACNE

In 2006, Margolin *et al.* presented the *Algorithm for the Reconstruction of Accurate Cellular Networks* (ARACNE) [75]. It is based on the Data Processing Inequality, meaning, if gene x_1 interacts with gene x_3 through gene x_2, then

$$I(x_1; x_3) \leq min(I(x_1; x_2), I(x_2; x_3)). \tag{1.12}$$

After assigning an edge between two nodes based on their mutual information, it tests each interaction for statistical significance. If $I(x_i; x_j) < I_0$, a given threshold, there will no edge be inferred between x_i and x_j. This approach has been extended by Zoppoli *et al.* to the Time-Delayed ARACNE [137]. It offers the possibility to include time-series information into the modeling process. By determining the time of initial change of expression, it is able to detect time-delayed dependencies. The resulting network is directed, in contrast to the original ARACNE. However, the additional complexity of the calculation makes it unsuitable for large scale network inference.

1.3.2.2 CLR

2007, Faith *et al.* introduced the *context likelihood of relatedness* (CLR) algorithm [30] as an extension of the relevance network. It derives a score $z_{i,j}$ for each pair of nodes x_i and x_j related to the empirical distribution of the mutual information values

$$z_{i,j} = \sqrt{z_i^2 + z_j^2} \tag{1.13}$$

with

$$z_i = max\left(0, \frac{I(x_i; x_j) - \mu_i}{\sigma_i}\right) \tag{1.14}$$

μ_i is the sample mean and σ_i is the standard deviation of the empirical distribution.

1.3.2.3 MRNET

The MRNET by Meyer *et al.* [78] uses the maximum relevance/minimum redundancy feature selection method to infer the networks. This method performs filter selection in supervised learning problems. For a set of input variables V and output Y, the method ranks V according to the mutual information with Y (maximum relevance) and the average mutual information with the previously ranked variables (minimum redundancy). The idea is that direct interactions have a less redundant information than indirect ones and because of this, should be ranked better.

1.4 Thesis proposal

The increasing number of drug-resistant strains among pathogenic fungi is a serious threat for immunocompromised people all over the world. Developing new treatments and enhancing the effectivity of current drugs are keystones in tackling these challenge. To find new drug targets is a major task in systems biology and bioinformatic methods a valuable assets in this work. We know that in the genetic regulation of organism, some genes are more important than others. The identification of hubs in gene regulatory networks requires to reconstruct the topology of biological network as precise as possible. Two well recognised properties of these biological networks, not only on the genetic level, are sparseness and scale-freeness. A robust method to create and evaluate gene regulatory networks of pathogenic fungi, that is also able to include current knowledge into the modeling process, is necessary for a systematic search of new drug targets.

Often, it is not really understood, how currently applied clinical drugs work on the genomic level of the pathogen. This is especially dangerous when pathogens start to show unpredictable reactions to the treatment or start to develop resistances. A systems biology study not only helps deeper understanding of the genetic effect of a drug to counter resistances, it also gives valuable hints on how to enhance the effect of the drug altogether. The emergence of RNA-Seq data allows for a focused and reliable prediction of the regulatory processes in a pathogen during the application of antifungal drugs. Yet it is often not clear what workflow should be followed in order to get results that are robust and statistically meaningful. Tests in the laboratory are expensive and time consuming, so a bioinformatic workflow in the analysis is necessary before biological testing begins.

Network modeling on large- and small-scale often follows different biological questions and computational requirements and therefore needs different approaches in order to achieve results. The variety of inference methods is hard to keep track of different developments, increasing the need for standard procedures in the analysis of different datasets.

Not always is the amount of data sufficient for network reconstruction when investigating the influence of different drugs. And it is also not always necessary, as the analysis of differentially expressed genes can already be of great help when trying to get first insight of how drugs work. Humidimycin does not have antifungal properties on its own, but seems to enhance the effect of Caspofungin, a clinically applied drug. Knowledge about the global genetic effects the combination of these two drugs can help to increase the effectivity of the antifungal treatments.

In face of these circumstances, this thesis addresses the following questions:

1. Given transcriptomic data from different experiments, prior knowledge of different sources and an automatically harvested gold standard, is it possible to infer a gene regulatory network that is sparse and follows a scale-free distribution of node degrees, in order to identify hub genes?

2. Given RNA-Seq data from a drug study, including knock-out mutants of key regulators, is it possible to extract prior knowledge from the knock-out data, and

infer a focused gene regulatory network that predicts gene regulations that can be verified in the laboratory?

3. What are conceptual differences between large- and small-scale network inferences?

4. Given RNA-Seq data from a study of different drugs and their combination, can bioinformatic analysis give hints on the genetic influence of the treatments?

The first question was investigated for *C. albicans* while in the second and fourth question, *A. fumigatus* served as model organism. These questions can be posed for any organism and the approaches should be able to handle any given organism.

1.5 Outline of the thesis

The first question is addressed in the second chapter of this work. After introducing the question of study, data and methods, the results of the modelling is presented. First tests were run on smaller sub-models containing only genes that are part of the gold standard to investigate the influence of the prior knowledge. Next, full-genome networks are presented with the help of different sources of prior knowledge, the combination of all prior knowledge sources as well as models with no prior knowledge at all. The final model is investigated towards sparseness and scale-freeness and hubs are identified. The results are also compared to three different mutual information network inferences. This work is also the subject of publication [4] of 2012.

The second question is the topic of the third chapter. Again, the first third of the chapter is used to introduce the question in more detail, as well as the data and the methods that were used to answer it. The data from a RNA-Seq study of *A. fumigatus* and different knock-out mutants under Caspofungin treatment is presented. It also includes the methods to extract prior knowledge from knock-out mutants as well as the literature used to extend the prior knowledge. The second third presents the result starting from the investigation of differential expression and clustering of genes. The gene selection is explained in detail as are the candidates for the modeling. After different models are inferred using the NetGenerator tool, the final network is selected using model error and number of implemented prior knowledge. After the interactions in the model are tested for robustness, hypotheses are extracted and tested in the laboratory *via* western blotting and qRT-PCR. Eventually the results are discussed. To make this work public a manuscript has been drafted and is about to be submitted.

The comparison of the large- and small-scale approach is part of chapter four. This includes a repetition of the analysis of the previous chapter with the method presented in chapter two. The results are discussed differences and recommendations for applications are given.

The content of the fifth chapter is the comparison of RNA-Seq data from *A. fumigatus* under the influence of Humidimycin, Caspofungin and the combination of both. Investigation of differentially expression and subsequent clustering is used to show the difference

of global gene expression. Differentially expressed genes are studied using gene enrichment analysis. This work is also part of a manuscript that has been drafted and will be submitted soon.

The last chapter summarises the results of the previous chapters and draws final conclusions.

2 Full-genomic network inference on *C. albicans*

2.1 Introduction

Since the inference of a full-genome network model requires a lot of data, the first large-scale model inferences were applied on model organisms like *S. cerevisiae* in 2005 [41] by Gustafsson *et al.* and *Escherichia coli* by Faith *et al.* in 2007 [30]. Gustafsson used an ODEs-based approach called LASSO (See chapter 1.3.1.2 for details) and proofed its capability to model large-scale biological systems. In his thesis, he himself stated that the "inferred system contains lots of errors but ... is more right than wrong" [39]. He also mentions the lack of "golden truth" to benchmark the models. Instead, he uses topological properties and biological annotation from data bases to evaluate his networks.

Faith *et al.* presented an inference method based on mutual information, called *CLR*. He also stated that the lack of experimentally determined interactions in combination with corresponding gene expression data makes it difficult to judge the quality of the network. He was able to find 3216 experimentally determined *E. coli* interactions and, independent from that, 445 microarrays.

Four years later, non-model organism *C. albicans* was subject of full-genome studies [69]. It is the first human fungal with a full-genomic network model. Here, the ODE based adaptive LASSO algorithm was applied (See 1.3.1.2 for details). It is able to implement prior knowledge into the network. It provided useful insight into the interactions of genetic interactions, but it does not follow a power law distribution of node connections (See chapter 1.2.2).

Since all network inference projects have to cope with the problem of how to evaluate the quality of the network. Along with this comes the question of the strengths and weaknesses of different modeling approaches and how to compare them. To address this questions, Stolovitzky *et al.* started the "Dialogue on Reverse-Engineering Assessment and Methods" (DREAM) [109]. It contains a conference, specifically addressed to network inference assessment, as well as the DREAM challenge, which started in 2007 and was called DREAM2. In the DREAM challenge, the DREAM team provides expression datasets from artificially created networks. The topology of the network is undisclosed and the teams that participate in the challenge are asked to uncover it with the help of the expression data provided.

2.2 Data & Methods

2.2.1 Data

2.2.1.1 Microarray data

When collecting data for a large-scale analysis of microarray data, it is often necessary to include data from different sources. One of the biggest collection of microarray datasets for *C. albicans* was published by Ihmels *et al.* in 2005 [51]. It contains the expression data of 6167 open reading frames (ORFs) in 244 expression profiles and combines the work seven laboratories. The conditions, under which the samples were taken range from drug exposure to application of mating pheromones. Since the set up and conditions of the experiments are so diverse, the dataset is not complete. There are 16.7% missing data points, which have to be imputed, since the applied network inference method can not handle missing values. I imputed missing data with the remaining values in the expression profile. 411 ORFs and 46 expression profiles have more than 50% missing data points. Imputing values on more than 50% missing data is highly unreliable, so I omitted the respective expression profiles. After the filtering and imputation, I was left with 198 expression profiles for 6167 ORFs. I used the Local Least Squares imputation, which is part of the pcaMethods package [106] for the statistic language R [115].

2.2.1.2 Gold standard

We used text mining to harvest as much information about the gene regulation in *C. albicans* as possible. We call this information *gold standard*, as it contains interaction we consider "correct", in order to evaluate the results of our network inference. We downloaded around 9,000 open access research paper about *C. albicans*. Buyko *et al.* applied their **JReX** [17] algorithm, a high-performance machine-learning relation extraction system. Providing syntactic and semantic information, JReX was able to identify 1,016 interactions between 509 genes. 503 of them are also part of the expression set and are now called *gold genes*. The reliability of such an automatically collected set of interaction is disputable [60], yet there is still no manually curated gold standard available for *C. albicans* and therefore this approach seems justified.

In an attempt to overcome this collection problem and offer a quick and easy way to access fungal specific annotation, different databases have been created. Notably *FunTF* by Shelest *et al.* [102] and *FungiDB* by Stajich *et al.* [107]. The available data contains genome sequences and annotation for 18 species of several fungal classes as well as cell cycle microarrays and RNA-Seq data. To further assist in *in silico* studies, it also offers an analysis pipeline.

2.2.1.3 Prior knowledge

The amount of time points we have is still very small compared to the number of predictors we try to estimate. To compensate that, I included four different prior knowledge sources (See chapter 1.2.1), collected by Jörg Linde (HKI, Jena).

FAC: 249 interactions been 226 genes.

From the TRANSFAC database [134], physical transcription factor – target gene interactions were harvested. This is a human curated database holding regulatory interactions for a number of organisms including fungi. We downloaded all fungi related interactions and blasted the protein sequence of transcription factors and target genes against the *C. albicans* genome. The necessary sequence similarity was 25% and an E-value had to be smaller than 0.001.

TRANS: 2689 interactions between 1502 genes.

Deriving from *S. cerevisiae*-orthologous genes is a dataset based on transcriptional relations. The data was acquired from the work of Balaji [10] who created a regulatory network based on transcription factors. Again, we mapped the orthologous genes to *C. albicans*.

BIND: 6333 interactions between 2288 genes.

The Biomolecular Interaction Network Database (BIND) [9] is an archive for biomolecular interactions and pathways. The data is gathered through individual submission, Protein Data Bank (PDB), as well as large-scale network inferences.

PPI: 6674 interactions between 2290 genes.

This dataset consisting of protein–protein interactions. It was acquired from orthologous genes of *S. cerevisiae* which were taken from the MPACT [44] section of the CYGD database at MIPS. For every protein–protein interaction, we identified the corresponding orthologous genes in *C. albicans* and added the pair of source genes to the prior knowledge.

2.2.2 Methods

For my investigation of the transcription data, I used the combined approach of ridge regression and the adaptive LASSO, as presented in chapter 1.3.1.4. First, I calculated the coefficients for the ridge regression and used the sum of coefficients for each gene as upper limit for the sum of coefficents for the LASSO solution. A parameter to be estimated is c, which indirectly determins the number of connections in the network. It is a factor multiplied with the solution of the ridge regression, allowing more or less coefficients to be included into the model. In order to test different network sizes, I tested 24 different values of c.

The second crucial parameter in the LASSO algorithm (See formula 1.6) is λ, the influence the prior knowledge has on the network inference. It lowers the penalty for adding prior knowledge interactions to the model. The smaller the value of λ is, the less penalty a prior knowledge interactions receives, which makes it more likely to be selected. If $\lambda = 1$, prior knowledge interactions gain no benefit compared to other interaction, which has an equal effect to adding no prior knowledge at all. In order determine a good level of influence, I started a grid search over 10 values ranging from 0.1 to 1. To check

the quality of the network, I used the *F-measure* [94] in search for the best accordance to the gold standard. The F-measure incorporates two different, often contradicting, aspects of model design. One is the completeness of correct interactions, represented by the *recall*[1]. The second is the ratio of correctly identified interactions compared to all identified interactions, called *precision*[2]:

$$F = 2 * \frac{precision * recall}{precision + recall} \tag{2.1}$$

2.3 Results

2.3.1 Parameter estimation

2.3.1.1 Prior knowledge

At first, I investigated, how many common interactions can be found among the different prior knowledge sources. As shown in figure 2.1, the general overlap between the prior knowledge sources is very low. Only BIND and PPI have 4337 common interactions. This is most likely because both data sources investigate very similar properties. The next biggest overlap is between TRANS and FAC with only 80 interactions. If the little overlap indicate a low compatibility of the data or a widespread use of different sources, remains to be seen.

Since the central measure for the correctness of the network is the gold standard, I calculated the intersect between it with the prior knowledge sources. The results can also be seen in figure 2.1. The overlap is very low. PPI has the most common interactions with the gold standard, which is not surprising, since it is the biggest source. FAC, being the smallest prior knowledge, has only 14 interactions in common with the gold standard.

I studied an expression matrix with 6167 genes and 198 experiments and a total of 15,945 prior knowledge interactions. In order to investigate how much influence the prior knowledge should get in the network I studied different values of λ. To increase the accuracy, I limited this test to a subset of the original transcription data, consisting only of the 503 gold genes (See chapter 2.2.1.2 for details). Testing the 10 possible values for λ on the whole dataset would also be computationally very demanding.

The result of the weighting process is depicted in figure 2.2. It shows that the more influence we give to the prior knowledge, the better the F-measure of the model. Because of this, I decided to use $\lambda = 0.1$ as weight.

I refrained from giving the prior knowledge higher influence. The information is mainly gathered from orthologous genes of *C. albicans*, mostly *S. cerevisiae*. Therefore, I do not consider the reliability of the dataset very high. Even with a soft integration, incorrect prior knowledge with a high influence can still lead the network into the wrong direction.

[1] $recall = \frac{TP}{TP+FN}$
[2] $precision = \frac{TP}{TP+FP}$

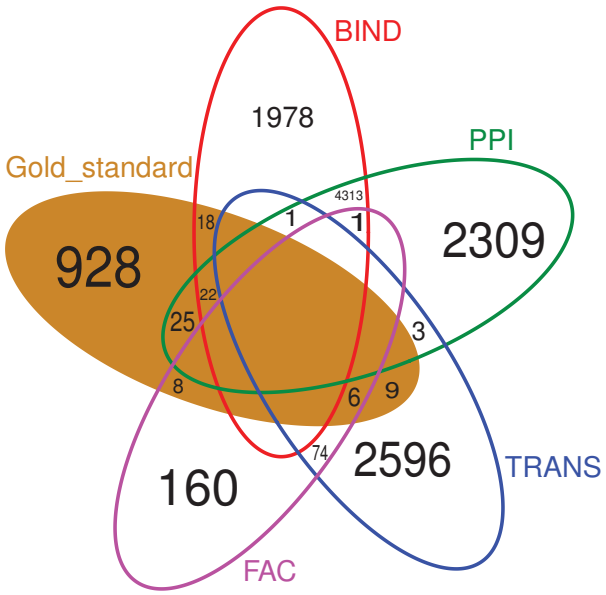

Figure 2.1: **Overlap of prior knowledge with the gold standard.** Despite BIND and PPI, there is only little overlap between the various prior knowledge sources. No prior knowledge source has many interactions in common with the gold standard.

2.3.1.2 Network size

I applied the network inference method of combining the ridge regression and the adaptive LASSO. An important variable is the scaling factor c. Gustafsson fixed the parameter at 0.1. To further investigate the effects of c, and to have an influence on the network size, I still performed a grid search over 24 different values. They ranged from 0.00001 to 0.5. I used the complete microarray dataset for the investigation and compared the network to the gold standard using the F-measure.

First, I modeled a network using no prior knowledge. The result of the F-measure analysis can be seen in figure 2.3. The best score was found at a c value of 0.2 by a model containing 6867 interactions between 6167 genes. The general F-measure is very low, since I compare the model to the gold standard, which is much smaller than the model.

Then, I inferred network models for each prior knowledge source individually. The results can be seen in table 2.1. The FAC prior knowledge gave results very similar to the one without any prior knowledge, as far as the F-measure is concerned. This is to be expected, since FAC is the smallest prior knowledge (29 interactions) and should therefore have only little influence. Also, it has the smallest total overlap with the gold standard (14 interactions), so the little improvement is no surprise. The network size is also close to one without prior knowledge, with 6886 interactions.

With 47 interactions, the overlap between the gold standard and PPI is higher, since PPI has 6674 interactions. More of a surprise was the high increase in the F-measure.

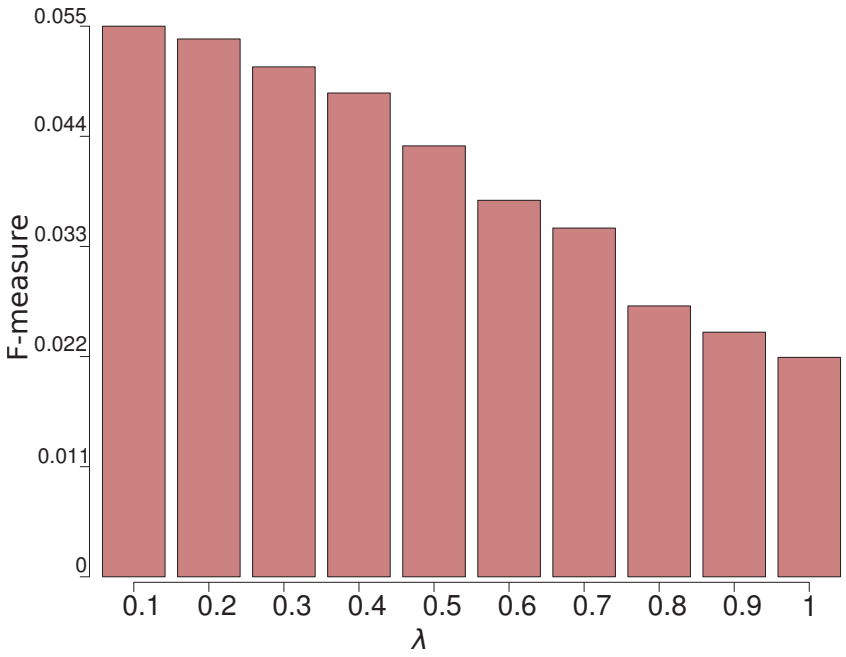

Figure 2.2: **Influence of the prior knowledge on the F-measure for a network consisting of 503 genes**. The lower the value of λ is, the more influence has the prior knowledge. $\lambda = 1$ gives no influence for the prior knowledge. It shows that the more influence the prior knowledge has (small λ), the better the F-measure. Giving no influence to prior knowledge ($\lambda = 1$) gives the worst results.

Table 2.1: **Results of the genome-wide network inference.** The first five columns show the results for LASSO and LASSO with different prior knowledge sources. The sixth column shows the LASSO inference with ALL four sources of prior knowledge and the column the results when the gold standard is given as prior knowledge. The last row shows the coefficient of confidence for the fit of the node degree to the power law distribution.

	LASSO	LASSO +FAC	LASSO +PPI	LASSO +TRANS	LASSO +BIND	LASSO +ALL	LASSO +GOLD
F-measure	0.0018	0.0015	0.0053	0.0058	0.0067	0.0064	0.0202
# of interactions	6,867	6,886	6,167	6,167	6,167	6,167	6,167
R^2 to power law	0.954	0.945	0.929	0.943	0.933	0.937	0.933

With 0.0053 it is about 4 × higher compared to FAC or no prior knowledge. Close to this result comes TRANS, which has an F-measure of 0.0058, and BIND, which produces the best results with 0.0067.

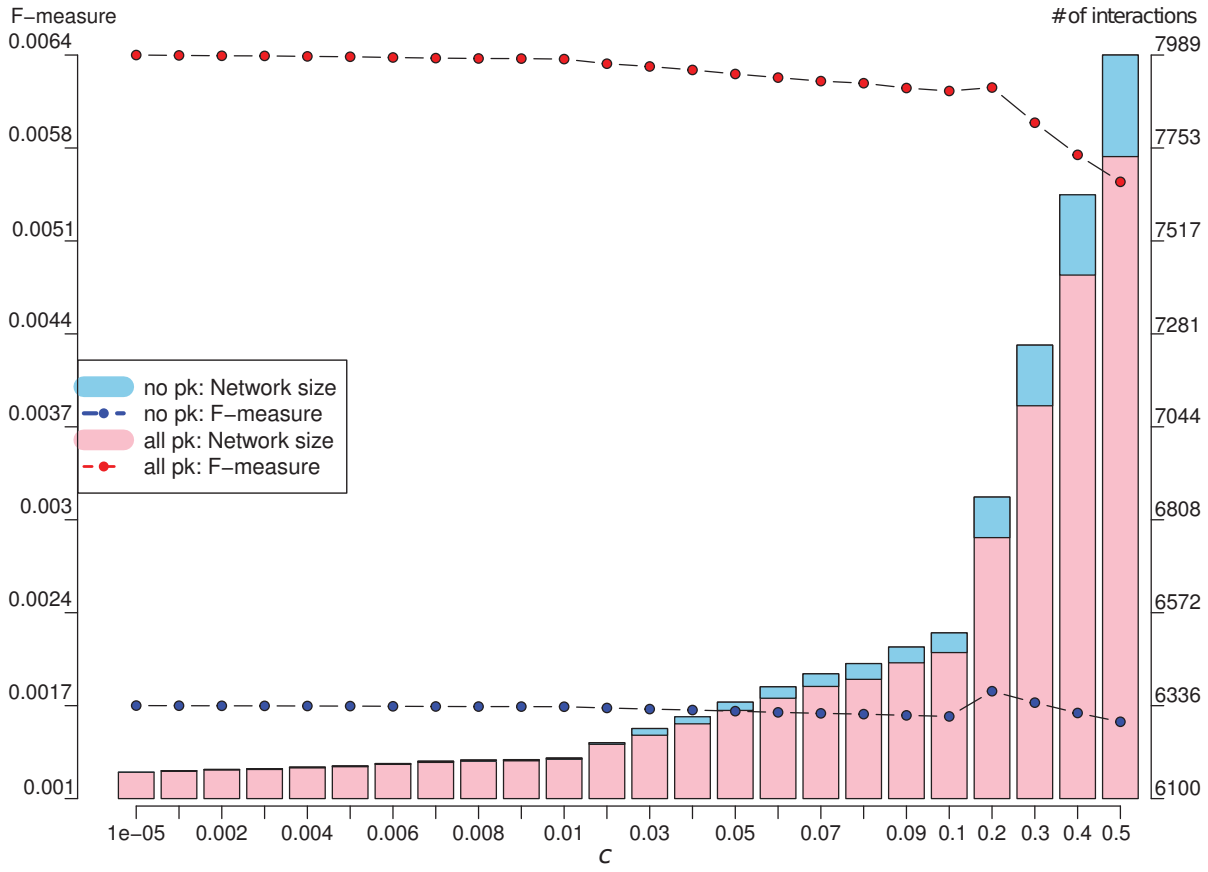

Figure 2.3: **Result of the large-scale network inference.** This plot shows the results for the network inference using no prior knowledge (**no pk**) in shades of blue and ALL prior knowledge (**all pk**) in shades of red. The circles show the F-measure for different values of c while the bars indicate the network size. The maximum F-measure was achieved at a c value of 0.2 for the model without prior knowledge and at 0.00001 for the model with all prior knowledge.

I also created a model using all different prior knowledge sources (ALL) to study their combined effect. The results were close to those of the "big" prior knowledge sources (PPI, TRANS and BIND). The optimal F-measure of 0.0064 was again reached at the smallest c value, giving it 6167 interactions. Despite the slightly lower F-measure, I chose this network as the final model for further investigation.

To further investigate the overlap of prior knowledge and gold standard on the network inference, I also inferred a network model using the gold standard directly as prior knowledge. Again, the smallest network has the highest F-measure, with a value of 0.02. It is not surprising that this model has the highest F-measure of all inference attempts. Yet the intensity of the increase is remarkable. The gold standard contains only 1016 interactions between 503 genes. This makes it the second smallest "prior knowledge" used. Nevertheless, the F-measure is more than 14 × higher than compared to the model without prior knowledge and 3 × higher than the second best, BIND. This clearly demonstrates the big influence the gold standard, and especially the accordance with the prior knowledge, has for the model evaluation. What matters is not so much the size, but the quality of a given prior knowledge in regard to the gold standard.

Looking at the figure 2.3, it is obvious that the F-measure generally decreases, the higher the c value becomes, i.e. the bigger the network is. This is also the reason that most models have their highest F-measure at the smallest network size. This can easily be explained, since, when adding another interaction, it is more likely that this interaction is not included in the gold standard. Therefore, the bigger the network becomes, the smaller the overlap with the gold standard. In some models, there is a break in this trend at $c = 0.2$, where the F-measure shows a peak in figure 2.3. I observed this peak to have the highest F-measure in the models without prior knowledge and FAC. PPI and ALL also show this peak, also less distinct, as it is not the highest value in the graph. BIND and TRANS do not show it, nor does the network inference with the gold standard as prior knowledge.

Further investigation on this matter revealed that this peak is always caused by the same incidence: the discovery of the interaction *orf19.4759* → *orf19.1770*. This interaction is not part of any prior knowledge dataset, but it is part of the gold standard, which is why its discovery cause an increases of the F-measure. The increase is relative to the amount of already correctly identified gold standard interactions. At $c = 0.1$, the model without prior knowledge found only six gold standard interaction, whereas the model based on all prior knowledge found 23 interactions. Since the relative increase of quality is higher in the first model, the peak there is more distinct, even though the absolute F-measure is smaller. Despite identifying a high number of gold standard interaction, PPI and TRANS do not find *orf19.4759* → *orf19.1770* at any tested model size. Especially they do not discover a new gold standard interaction at $c = 0.2$, which is why there is no peak visible at this size. Despite not showing a peak at $c = 0.2$ either, the model based on the gold standard as prior knowledge (GOLD) discovers two further interactions at this point[3]. However, at $c = 0.1$ GOLD already identified 72 gold standard interaction, and the relative increase in quality is to small to compensate for the increase of network

[3] $orf19.6798 \rightarrow orf19.6109$ and $orf19.3829 \rightarrow orf19.6081$

size. The interaction *orf19.4759* → *orf19.1770* is already part of the network at $c = 1$.

But why do these changes happen at $c = 0.2$? Compared to any other tested c value, this one has its biggest increase of allowed error compared to its predecessors, making it more likely for changes to occur. Actually, most networks identify many of those interactions already at the smallest tested network size 6167 at $c = 10^{-5}$. Additional findings later on, if there are any, have a rather small impact on the F-measure. Because of this, many networks have their optimal network size at 6167 interactions.

2.3.1.3 Mutual information networks

To compare our results to other state-of-the-art algorithms, I created the network inferences using mutual information (MI) networks (see chapter 1.3.2). Specifically, I used ARACNE, CLRNET and MRNET from the *minet* R package [79] with default parameters.

Again, I face the question of the network size. The first choice is to select every interaction, for which the algorithms identified a confidence ≥ 0. CLR and MR returned networks with over 15 million connections between all 6167 genes, whereas ARACNE returned a network with 39,986 connections. Not surprisingly, the F-measure is very low. CLR and MR reached a value of $6 * 10^{-5}$ and ARACNE $9 * 10^{-4}$.

To increase the comparability between the MI-based and regression-based methods, I shrunk the larger MI based models to equal sizes as the LASSO-based model. I selected the 6167 most confident interactions of and calculated the F-measure again. The quality increased dramatically, CLR reached a value of 0.0025, MR 0.0016 and ARANCE 0.0019. This places these models among the LASSO without prior knowledge but below those with PPI, TRANS or BIND prior knowledge.

However, the MI models of shrunken size did not connect all genes. In the MR model, only 3419 genes were connected to any other gene. In the CLR the number decreased to 2147 and ARACNE connected only 2022 genes with each other. The LASSO implementation calculates the connections for every gene individually so every gene should have at least one interaction partner. The MI implementations calculate the connections globally, so some genes can be left without any connection.

All these algorithms deliver weighted adjacency matrices of the genes in the model. The values range from 0 to 1 and are confidence measures on the presence of a relationship between two genes. The interactions are undirected, since the adjacency matrix is a diagonal matrix, in contrast to the linear regression networks and the gold standard itself. To make the results comparable, I split each undirected interaction $x_1 \leftrightarrow x_2$ into two directed interactions $x_1 \rightarrow x_2$ and $x_2 \rightarrow x_1$.

2.3.1.4 Hubs

An important goal of large scale network inference is the identification of hubs, central regulatory genes with a high connectivity. The first question that arises is: How many connections should a gene possess, to be considered a hub? Han *et al.* [46] suggested that a hub should have at least six interactions with other genes. Since our model contains

Table 2.2: **Results of the genome-wide network inference with mutual information methods.** The first three columns show the results for all interactions, for which the algorithms found a correlation. In the last three columns, the number of interactions has been reduced to the same number as the LASSO model of the inference based methods (6167). This increases the comparability of results. Reducing the confidence to the 6167 most confident ones leaves some genes disconnected from the network. The last row shows the number of genes with at least one connection.

	CLR	MR	ARACNE	CLR small	MR small	ARACNE small
F-measure	0.00006	0.00006	0.0009	0.0025	0.0016	0.0019
# of interactions	15,686,064	15,329,450	39,986	6,167	6,167	6,167
# of genes	6167	6167	6167	2147	3419	2022

more nodes that the one of Han *et al.* I decided for a minimum of seven, in order to identify hubs. The use of fixed numbers is of course difficult. The final decision of how many connections a gene should have to be considered a hub should depend on the total number of genes and the number of connections in the network.

In the LASSO-based network with ALL prior knowledge, I identified 126 genes with an out-degree of seven or more. To annotate them I used the Candida Genome Database [104]. Despite ongoing research, only a few genes are annotated, and 31 hub genes I identified do not have a functional description. Often, the annotation originates from orthologous genes of *S. cerevisiae*, which makes them less reliable. According to the literature search, the potential hub genes come from various different functional categories. At least 16 of them are known to susceptible to antimycotica like *amphotericin B*, *azoles* or *Caspofungin*, making them potential drug targets. See table 2.4. Details about Caspofungin can be found in chapter 3.1.1.

The three genes with the highest out-degree are *FET31* (29), *BNI5* (28), and *orf19.1300* (25). The fact, that one of them is only known by its *orf*-ID demonstrates the lack of annotation. *FET31* [72] is especially interesting, since it is a putative iron transport precursor, which makes it reasonable to assume to have a central regulatory role. The fact that it is also susceptible to antifungal drugs makes it an interesting drug target. However, heterozygous null mutants remain viable, which lowers its usefulness as a drug target.

BNI5 is part of the cytokinesis and septin ring assembly, which is about everything interesting about it. *orf19.1300* is better studied, despite the fact that is misses an alias. It is a putative membrane protein and is part of the process controlling filamentous growth. Homozygous and heterozygous null mutants remain viable, yet homozygous null mutants have abnormal appearance. Not much more is known about these genes.

Whether they really have central positions in the regulatory processes can only be determined by experimental validation.

Figure 2.4 demonstrate how the use of different prior knowledge leads to different interactions in the final model. The hub gene *PSA2* is responsible for the nucleotidyltransferase activity and biosynthetic processes in *C. albicans* [104], while *TKL1* is involved in the transketolase activity. The figure shows which interaction was estimated given a certain kind of prior knowledge.

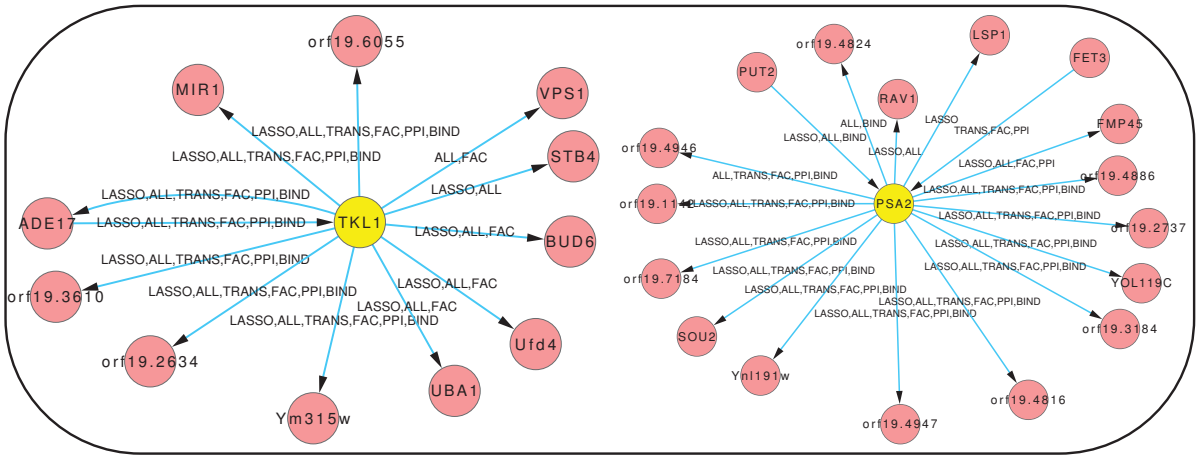

Figure 2.4: **Predicted hubs *PSA2* and *TKL2*.** The labels on the edges show, which prior knowledge had to be included into the inference algorithm for that specific interaction to occur. LASSO means this edge was found without any prior knowledge.

I realised that the hubs in the gold standard differ greatly from those in the networks. To determine an appropriate hub size, I calculated the quantiles of the out-degree distribution. The 75% quantile is reached at a value of 3, so I selected an out-degree of 4 as the selection criterion, since I want a rather strict hub definition. This lead to 83 hub genes. From those, only one (*WOR1*) was also identified as hub gene by the network model without prior knowledge. Compared to that, the network model with all prior knowledge has four hub genes in common, see table 2.3.

This shows that there is indeed a significant difference between the gold standard and the models based on transcriptional analysis, disregarding if they have been inferred with or without prior knowledge.

Table 2.3: **Four hub genes that the gold standard and the network inference with all prior knowledge have in common.** *WOR1* was also identified by the model without prior knowledge.

name	description
ACT1	Actin; transcript regulated by growth phase, starvation
ADE2	role in adenine biosynthesis; required for normal growth and virulence
WOR1	Transcription factor required to establish and maintain the opaque state
BNI1	Formin; role in cytoskeletal organization, cell polarity

Table 2.4: **16 hubs which are sensitive to antifungal treatment.** The rows contain the name, the out-degree and the description of the gene. The annotation was taken from the Candida Genome Database [104]. Susceptibility to antifungal agents are highlighted.

FET31	29	Putative iron transport multicopper oxidase precursor; **flucytosine** induced; **Caspofungin** repressed
SOG2	16	Domain protein of RAM cell wall integrity signalling network; role in cell separation, **azole** sensitivity; required for hyphal growth; lacks orthologs in higher eukaryotes
HIP1	13	Similar to amino acid permeases; alkaline upregulated; **flucytosine** induced; fungal-specific (no human or murine homolog)
ARX1	13	Putative ribosomal large subunit biogenesis protein; downregulated during core stress response; decreased expression in response to **prostaglandins**
UTP22	11	Putative U3 snoRNP protein; decreased expression in response to **prostaglandins**; heterozygous null mutant exhibits resistance to **parnafungin**
NOG1	11	Putative GTPase; mutation confers hypersensitivity to 5-fluorocytosine, 5-fluorouracil, and tubercidin; decreased expression in response to **prostaglandins**
TYR4	10	Putative zinc finger DNA-binding transcription factor; **fluconazole**-downregulated; expression regulated during planktonic growth

Continued on next page

Table 2.4 – continued from previous page

APT1	9	Adenine phosphoribosyltransferase; **flucytosine** induced; repressed by nitric oxide; protein level decreased in stationary phase yeast cultures
Hmg2	8	HMG-CoA reductase; enzyme of sterol pathway; inhibited by **lovastatin**; gene not transcriptionally regulated in response to **lovastatin** and **fluconazole**
Cor1	8	Putative ubiquinol-cytochrome-c reductase; **amphotericin B** induced; repressed by nitric oxide; protein level decreases in stationary phase cultures
Taf19	8	Putative TFIID subunit; mutation confers hypersensitivity to **amphotericin B**
OPT8	8	Possible oligopeptide transporter; induced by nitric oxide, **amphotericin B**
Imp4	8	Putative SSU processome component; decreased expression in response to **prostaglandins**
ASR1	7	Putative heat shock protein; transcription regulated by cAMP, osmotic stress, **ciclopirox** olamine, **ketoconazole**; stationary phase enriched
OPI3	7	Phosphatidylethanolamine N-methyltransferase of phosphatidylcholine biosynthesis; downregulation correlates with clinical development of **fluconazole** resistance; **amphotericin B** and **Caspofungin** repressed
AGP2	7	Amino acid permease; hyphal downregulated; regulated upon white-opaque switching; induced in core **Caspofungin** response, during cell wall regeneration, or by **flucytosine**; fungal-specific

2.3.1.5 *GAL* sub-network

In my network, I was also able to identify a known studied sub-network. The so called *GAL*-network has been extensively studied in *S. cerevisiae* [54,73] and is well conserved in the yeast clade. It was used to describe transcriptional rewiring between *C. albicans* and *S. cerevisiae* [97]. The GAL-network plays a major role in the degradation of galactose. *GAL10* transfers β-D-galactose to α-D-galactose. This is converted to α-D-galactose 1-phosphate by *GAL1*. *GAL7* then transfers α-D-galactose 1-phosphate to α-D-glucose 1-phosphate. The regulatory cascade $GAL10 \rightarrow GAL1 \rightarrow GAL7$ is correctly predicted by the inferred network models from all 6167 genes, as can be seen in Figure 2.5. This is especially important, as these interactions are not part of any prior knowledge and only the interaction $GAL1 \rightarrow GAL7$ is part of the *gold standard*.

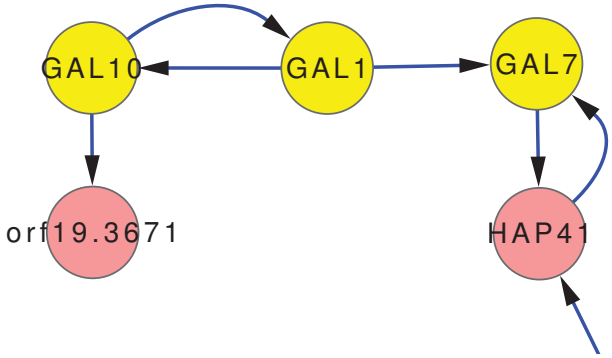

Figure 2.5: **Sub-network of GAL-genes**. Modelled using LASSO and all prior knowledge. Predicted relations are not part of the prior knowledge. The connection of *GAL10*, *GAL1* and *GAL7* is well studied in many yeast forms.

2.3.1.6 Scale-freeness

A desired property of the inference model is scale-freeness (see chapter 1.2.2 for details). To measure this property, I used Cytoscape [105] to fit a power law to the degree distribution of the model, and calculate the R^2. It is a coefficient of confidence, and describes the fit between data points and the model. The results can be seen in table 2.1. The coefficient for all LASSO based models is above 0.9. A graphical representation can be seen in figure 2.6. The model with no prior knowledge and all prior knowledge have the highest R^2, while the TRANS prior knowledge lead the model to the lowest value. This proves that the algorithm is able to infer scale-free network models.

To investigate how much influence the prior knowledge has on the scale-freeness, I calculated the R^2 scores for them as independent networks (see table 2.5 for results). All reached high values, except for TRANS, which had to settle with a score of 0.6. The combination of all available prior knowledge also had a high R^2 score of 0.9. BIND and PPI are the biggest prior knowledge networks by far and so already show the scale-freeness of the underlying biological network, since an "investigation bias" is more likely to occur in networks with a little number of genes.

The gold standard itself has a relative high R^2 of 0.87. Every gold gene is part of an interaction pair, but that does not mean, all gold genes have a relation with each other. This can lead to a scattered and disconnected network. To investigate this, I looked for the biggest connected sub-network within the gold standard. In fact, 402 genes are directly or indirectly connected with each other. Only 42 genes have only one interaction partner. This shows that the gold standard is already well connected.

2.3.2 Parallelisation

Despite increasing computational power of modern cluster systems, genome-scale network inference is still time consuming. It is therefore often advisable to use parallel computing if possible. To do so, it is important to identify the time consuming parts

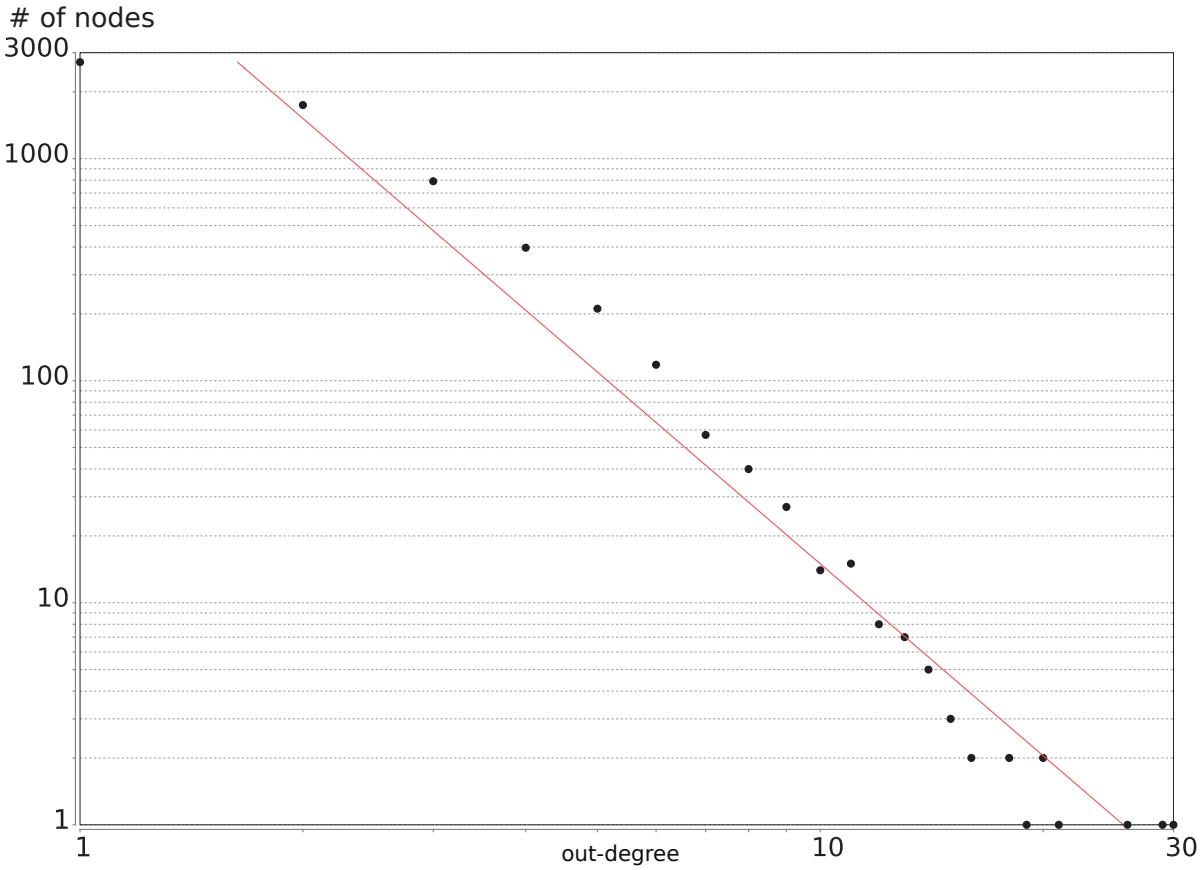

Figure 2.6: **Power law distribution of the nodes in the LASSO model.** The dots show the number of genes with a certain degree. The red line is the fitted power law distribution. The axises are logarithmic.

Table 2.5: **Scale-freeness of prior knowledge and gold standard.** The first two rows show the number of genes and interactions in the dataset. The last row the value of the R^2 score showing the coefficients of confidence that the node degree follows a power law distribution.

	FAC	TRANS	BIND	PPI	ALL	GOLD
genes	226	1502	2288	2290	3051	509
interactions	249	2689	6333	6674	11523	1016
R^2	0.857	0.615	0.917	0.891	0.908	0.871

of the calculation and extract those parts that are independent from each other. In case of this study, the calculation of the coefficients for each gene is the most time consuming part of the calculation. Fortunately, the calculation of the coefficients of one

gene is independent from the coefficient calculation of other genes. The penalty matrix and transcription information is distributed among multiple computer cores and each calculates the coefficients for the one gene. Later, the coefficient values are collected and a coefficient matrix is created. Each row contains the coefficients for the ODE for a specific gene. This matrix later determines the final network.

Even with this parallelisation, calculation power remains a restricting element. The price of computer equipment decreases and the use of cluster computing is achievable. Nevertheless, the amount of data grows faster than the computation power, increasing the demand for parallelisation and efficient programming.

The software used in this thesis is the *R* package *multicore* [122].

The calculations were done on a SUN Fire X4600 Server M2. It has 8 CPUs with 4 cores each and are clocked on 2.2 GHz. The available RAM is 128 GB. It took about five days for one complete run of the algorithm, including modeling and cross validation to finish.

2.4 Discussion

2.4.1 Weighting & evaluating the prior knowledge

Collecting prior knowledge for the network inference is time-consuming and raises a lot of methodical questions. However, the benefit is undeniable, since they have a great potential of increasing the predictive power of the algorithm, as seen in table 2.1.

The problems of combining different prior knowledge sources do not end with evaluating the confidence, or contradicting information, but also have a more substantial aspect. All the networks inferred using ODEs are directed. This means that a predicted interaction consist of a source a and a target b. The interaction has only one direction $a \rightarrow b$ with a certain sign, representing activation or inhibition. Most of the prior knowledge, however, is undirected, meaning we can not say whether $a \rightarrow b$ is true or $b \rightarrow a$ (or both). Since the prior knowledge is incorporated softly (a suggested interaction can still be skipped by the algorithm) I decided to split every undirected interaction $a \leftrightarrow b$ into two independent interactions $a \rightarrow b$ and $b \rightarrow a$. This can be justified by the observation of Christley *et al.* [19], who noted that the incorporation of incorrect prior knowledge does not decrease the predictive power significantly.

The key to this effect lies in the soft integration of prior knowledge. This way, every prior knowledge interaction is merely a suggestion, that is ignored if it does not fit to the data. This is also a reason why an undirected interaction can be split into two contradicting prior knowledge interaction can be used as well. The algorithm will simply chose the one that better fits the data.

In contrast to this, the use of ALL prior knowledge gives the model a lower F-measure than the use of BIND alone. This can happen, when the prior knowledge guides the algorithm into the “wrong” direction. If certain interactions are chosen early in the inference process, other interaction become less likely later on. A possible way to encounter this problem is weighting the prior knowledge sources individually. With further

investigation of the prior knowledge sources, one may be able fine-tune the influence of the prior knowledge. The question remains, how to evaluate this consciously.

The origin of mismatching prior knowledge and gold standard may also be that the prior knowledge mostly comes from *S. cerevisiae*, whereas the gold standard is based on *C. albicans* specific sources. Many interactions are well conserved among these species. Nevertheless, there are substantial differences in the genetic regulation between *S. cerevisiae* and *C. albicans*. This is also one reason I did not weight the prior knowledge more than I did, as it can lead to false conclusions.

2.4.1.1 Gold standard

The results of the modeling with ALL prior knowledge show that the suggested interactions sometimes fit well to the transcription data, even though they do not fit to the gold standard. Otherwise they would not have been considered in the first place. Here we see another possible source of error: What if the gold standard does not fit to the transcription data? This means, what if the genes, that are connected in the gold standard do not correlate in their transcriptional expression in the dataset? The algorithm would not be able to identify these interactions, if not guided by prior knowledge. The highest F-measure was achieved by a model that has the gold standard itself as input. While some might argue that this does not result in additional knowledge, it can still be beneficial. It can present a template of "correct" interactions to guide the inference along with it.

As seen in figure 2.2, more influence of prior knowledge increases the predictive power, given an overlap of prior knowledge and gold standard. The predictive power can also be increased by selecting smaller networks, which can have multiple reasons. The most important factor is the comparably small gold standard network. With only 1016 interactions between 503 genes, the genome-wide networks can never have high F-measures even if they include every interaction of the gold standard. The current size of the gold standard is still insufficient to reliably evaluate genome-scale network inferences. However, since new information about interactions and annotations for genes are published regularly, the size and quality of the gold standard will grow over time. This shows the necessity to find automatic and reliable ways to extract these information from the literature. Databases like FungiDB and the Candida Genome Database are still far from comprehensive. Automatic text mining software like **JReX** offers a mature way to harvest literature information available. The machine learning algorithm that drives the pattern recognition is still difficult to use and requires lots of work in fine-tuning and providing the necessary key words. Even though, the false positive and false negative rate remain comparably high.

The gold standard used is a mixture of directed and undirected interactions. Since the gold standard was also compared to mutual information networks, which are undirected by their nature, I also split every gold standard interaction $a \leftrightarrow b$ into two interactions $a \rightarrow b$ and $b \rightarrow a$. Of course, this approach can lead to false positives, i.e. that I count interactions as matches even if the gold standard originally had contradicting information. As mentioned before, in the case of genome-scale network inference, the aim is

more about finding hubs and investigating network topology than investigating individual interactions. Therefore, identifying a connection between two genes is a valuable feat, even though the direction may be wrong.

2.4.2 Scale-freeness

Despite the importance of large scale power law distributed networks, no such network was modelled for a pathogenic fungi so far. In my study, I was able to infer a network, whose coefficient of confidence (R^2) to the power law distribution was 0.96. This was possible by combining two algorithms, LASSO and ridge regression.

Since ridge regression does not make a parameter selection, i.e. does not shrink the coefficients of the model to zero, it always produces fully connected networks, which makes it an excellent tool for regularization, but not for model selection. LASSO on the other hand has a bias for coefficient values becoming zero. The models tend to be smaller, and the coefficient to more significant. This works along with the sparseness criterion, which assumes that the number of interactions in a network should be as small as possible, while still giving a reasonable data fit. Using ridge regression to estimate the constraint for the LASSO makes it possible to produce scale-free networks.

Using the combined approach of ridge regression and LASSO without prior knowledge certainly offers a rather unbiased approach on the network structure. By adding prior knowledge to the inference, we most certainly influence the networks structure, which is of course intended. However, Yu and colleagues [136] described the effect that becomes visible when we start to add prior knowledge. The prior knowledge is not equally distributed among all genes. In fact, the prior knowledge sources themselves follow partly a power law distribution of connections. I took the connections in every source of prior knowledge as an independent network and counted the degrees of every gene. When I calculated the coefficient of confidence to a fitted power law distribution we observed very high values. The BIND dataset has a R^2 of 0.92, which is the highest value of all prior knowledge. PPI follows with 0.9 and FAC with 0.86. TRANS has the lowest value with 0.62. The combination of all prior knowledge sources gave a R^2 of 0.91, which is better than some prior knowledge sources alone.

One might argue that this reflects the power law distribution of the underlying biological network, instead of following Yu's argumentation, that there is an investigation bias towards well studied genes. This claim would only hold true if all genes would have equal chance to be investigated, so that an inherent power law distribution would show itself. Despite the fact that the number of investigated genes increases constantly, there is an obvious focus on genes that are presumed to be essential. The general lack of annotation for example is a serious problem when working on regulatory networks. Also, well studied genes are more helpful in understanding the function of a cluster or hub structure, so these are of course more often part of the investigation.

The gold standard R^2 score of 0.87 is comparably high. Again, whether this reflects the "true" nature of connections, or an investigation bias, can not be told. However, one should consider, that the *gold genes* are collected from individual, independent data sources, which can lead to a scattered, unconnected network. This may not reflect

the correct relation between the genes. The genes in the gold standard were selected because they were part of a published investigation. Nevertheless, as the results show, the gold standard still has a high number of interconnected nodes. This rather hints to the existence of an investigation bias. The investigation of regulatory interactions is still in an early stage with only a little number of totally investigated interactions. This increases the likelihood of such a bias.

Including this, prior knowledge seems to give the inference a bias towards power law distribution of connections. One might argue that this is the cause for the final model to have this property. However, the inference that is not assisted by any prior knowledge at all also shows a power-law distribution for their node degrees. The coefficient of 0.95 is actually the highest among the networks, higher than any of the prior knowledge coefficients, as shown in table 2.1. The networks that have been inferred with the FAC and TRANS prior knowledge all have a degree distribution that correlates with the power law by values of 0.95 and 0.94. The BIND and PPI assisted network had a R^2 score of 0.93, as well as the network that had the gold standard as prior knowledge.

This shows that the node degree distribution of the prior knowledge does not necessarily correlate with the one of the final model. While the TRANS dataset itself has by far the lowest R^2 score, the value for model using it is in the middle of all results. BIND on the other hand has the highest R^2 by itself, but the model associated to it is sub-average.

2.4.3 Hubs

One question of this study is whether the algorithm is able to identify hubs in the regulatory network. As mentioned before, the definition of a hub is rather blurry. My decision to consider genes with an out-degree of at least seven as hubs is based on the work of Han *et al.* Choosing the correct value is an important step in every hub study. While I selected an out-degree of seven or more, other values are also possible. The in-degree can also be considered for hub selection, but I found that not useful in a search for possible drug targets. The high number of in-degrees gives them an expression profile that is difficult to predict. A high out-degree means that those genes have a large influence on the regulation of many other genes. These genes are often the beginning or important bottlenecks in signalling cascades. Knocking out these major regulators can impede or cancel the stress response or major metabolic pathways.

The hubs predicted here were checked for their drug susceptibility. Those genes are of great interest, since they do not only have a huge impact on gene regulation, but also because we have influence on them. This means the results can be directly applied in treatment for the benefit of the patients.

Identifying an iron transport precursor as the gene with many interactions seems possible. Of course, this is a prediction, with the exact number of interactions being rather unreliable. Nevertheless, the iron transport is regulated by a major pathway, inhibiting it impedes the viability of *C. albicans* substantially. However, experiments show that *C. albicans* stays viable even when the gene is knocked-out, which makes this gene no suitable drug target. But this only considers the genes that already have an

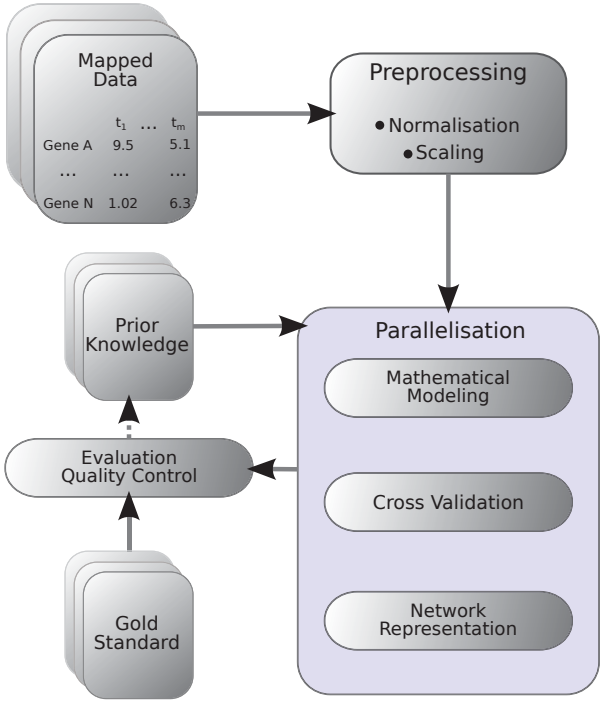

Figure 2.7: **Workflow of the *Candida* study.** After preprocessing, the expression data and the prior knowledge was used to create mathematical models. These models were evaluated and new hypothesis could be constructed.

annotation, and those are few. After all, looking only at those genes, that are already annotated gives only little insight. More fruitful may be the study of genes without annotation, but high number of connections.

As mentioned before, the results of genome-scale network inference should not be considered exact. The aim of this inference is rather to point out potential sites of high regulatory activity. Robustly differentiating what is the most important regulator and what is the second most important is beyond the scope and capabilities of this approach. The final decision can only be made using experimental validation in the laboratory. Between the genome scale network inference and the experimental validation stands the small scale network inference. It takes a closer look into the regulatory interactions, requiring less data and offering a higher reliability of the predictions.

2.5 Conclusion

The rising amount of transcription data available leaves the question of how to analyse them. Often, the data comes from different laboratories, was harvested under multiple conditions or various platforms. Despite many attempts to offer easy and machine readable access, collecting data compendium often still consists of a lot of manual work. The datasets themselves are often not comprehensive and inhomogeneous. Carefully and elaborate preprocessing is important in order to combine multiple datasets without

adding a bias. Selecting an adequate approach to model the data may be one of the most important steps in the inference process. Finally, choosing a model and evaluating it is time consuming and should not be underestimated. There is no unique set of answers for these decisions, since it mainly depends on what kind of questions one tries to investigate.

The main question studied here is: Is it possible to create a prior knowledge supported genome-scale network inference with a scale-free degree distribution, evaluated by an automatically harvested gold standard in order to identify central regulatory genes. The current thesis presents a successful way to tackle this task. The difference equation systems, used by the adaptive LASSO is able to cope with the lack of data and return robust results. The combination of the ridge regression for model fit and the LASSO for parameter selection proves to be able to create scale free networks. It is challenging to estimate so many interactions of genes using so little data. Using parallelisation and high-end server architecture, I was able to solve the equation systems in reasonable time. The general workflow is depicted in figure 2.7.

The incorporation of prior knowledge increased the predictive power of the algorithm. Even small sources of prior knowledge could increase the F-measure in the models. More comprehensive prior knowledge was able to increase the quality of the model drastically. Combining all prior knowledge sources to one big set, however, slightly decreased the the quality again. This most likely comes from contradictions inside the different sources. Further investigation on the fine tuning of influence of different prior knowledge sources may not only prevent such decrease in quality, but also improve the predictive power to levels higher than any source alone.

I found that the bottleneck in the evaluation of the model is the gold standard. In order to measure the predictive power, a comprehensive gold standard is necessary. The gold standard in this study is far from comprehensive. Despite hard work from Buyko *et al.*, only 1016 interactions between 509 genes could be extracted from literature. This lead to the preference of smaller networks in the model evaluation. Most of the genome-scale networks had their maximum F-measure at their smallest network size, which consists of 6167 interactions. These networks found most gold standard interactions at the beginning and barely any more after that. This fact also leads to the occurrence of a peak at $c = 0.2$, since with so few gold standard interactions to find, even one or two new findings can cause huge changes in the F-measure, as seen in figure 2.3. A larger gold standard with more interactions may lead to more sensitivity in the measurements. The small overlap of the gold standard and the prior knowledge is also problematic, since the prior knowledge does not seem to guide the inference in a direction that leads to a better F-measure. As new publications on fungal pathogens are submitted every day, the size and quality of a reliable gold standard may increase in the future. In fact, as a consequence of this findings, a new PhD-project will start in autumn 2014, speciffically aiming at improving the quantity and quality of the gold standard and the prior knowledge.

Apart from the gold standard interactions, the algorithm was also able to identify the *GAL*-sub network. The correct prediction of the regulation cascade $GAL10 \rightarrow GAL1 \rightarrow GAL7$ is a strong sign of the modeling power of this approach. Especially, since no prior

knowledge source contain any of these interactions.

The scale-freeness of the model enabled me to identify hubs in *C. albicans*. Studying the annotation for drug response and viability of knock-out mutants shows that, while many are susceptible to drugs, *C. albicans* is still able to compensate the change or lack of expression. There is more to the importance of a gene as drug target than just a high number of connections. Eventually, only experimental validation is able to check the results of computer modeling. Large-scale networks usually produces a large amount of hubs with only little reliability. Therefore, it is advisable, to continue the study with small scale networks, focusing on the potential hubs and their interactions.

3 Small-scale network inference on *A. fumigatus*

3.1 Introduction

Before the upcoming microarray technologies, expression changes were seen as linear pathways of causality [113]. With Microarrays, the study of complex regulatory networks became possible. The transition from sequential information to functional information is one important step in the inference process and genome-wide network inference is a powerful tool to identify potential regulators in the genetic network of an organism. The precision of the analysis is often not sufficient to get a detailed look at the individual connections. For a more detailed and robust study of regulatory connections, it is advisable to focus the investigation on a subset of genes [28].

3.1.1 *Aspergillus fumigatus* stress response

A. fumigatus is able to colonise various different habitats and in order to survive, it has to adapt to various changes in environment. A significant change in environmental conditions occurs for example when the conidia of the fungus colonise the human body. Among other factors, it comes with a change in temperature, pressure and pH-value and also an activation of the host's immune response, which poses a serious thread to the fungus. Inhaled conidia are often trapped in the mucus and transported out of the lung by beating of the cilia [11]. It has been shown, that *A. fumigatus* is able to release factors that can slow the beating of the cilia and damage epithelial cell to create an entrance into the body [5]. Another defense mechanisms of the human body consists of type I and II pneumocytes. *A. fumigatus* is capable using the respiratory epithelium cells as refuge from these phagocytes and may even start germination, as several different studies found out [88, 130]. Once the fungus managed to colonise the human host, it may be discovered and anti-mycotic treatment starts.

In order to overcome all these obstacles, *A. fumigatus* needs to quickly adapt to the changing environment. First, *A. fumigatus* has cell surface proteins that sense the environmental conditions. Once these proteins notice a change in the environment, they start a signalling cascades that eventually lead to activation of transcriptional regulators. These regulators cause a change in the expression of specific target genes and thereby an adaption to the new environmental state by producing (or stop producing) certain proteins. It is important for the survival of the fungus that these cascade mechanisms work quickly and target accurately the correct regulatory interactions. These signalling

cascades and their respective regulatory interactions have been well studied in the last decades [45,95]. Not only toward their roles in metabolic regulation, but also concerning drug adaption. It is important to understand that these cascades are not strictly linear and independent, but are interconnected and complex.

The cell wall of *A. fumigatus* consists of an unique polysaccharides composition [35]. This makes it a suitable target for antifungal drug treatment, since it distinguishes the fungal cells from human cells. Of central importance to this membrane formation is the protein Fks1, which synthesises β-1,3-D-glucan [18]. These glucans and chitins are essential components of the cell wall in many fungal species. An interesting adaption mechanism is the capability to compensate for a loss of glucans by inducing a chitin biosynthesis, and *vice versa* [108, 124, 127]. Details of the regulation of this mechanisms are still unclear.

A group of drugs that targets Fks1 and its glucan synthesis is Echinocandins [18]. The first clinically applied Echinocandins is Caspofungin [13]. It is used successfully against *Candida* sp infections, but shows less effectiveness against invasive mycoses, like aspergillosis. It was not able to decrease the mortality rate of patients suffering from invasive aspergillosis since it was introduced in clinical trials [56]. This ineffectiveness is mainly caused by two factors: One is the occurrence of resistant strains, the other is a so called "paradoxical effect". This paradox effect consists of a lack of effectively, when the concentration of Caspofungin passes a certain threshold [66, 96]. According to different studies, the glucan / chitin ratio in the cell wall may be responsible for this effect. The glucan / chitin balance itself is regulated by various different signalling pathways, which operate on protein kinase C (PKC), calcineurin, HOG and the mitogen activated protein kinases (MAPKs) [32, 80, 81, 127].

In the *A. fumigatus* genome, four putative MAPKs have been identified: MpkA, MpkB, MpkC and SakA. MpkA is responsible for the regulation of the cell wall integrity (CWI) pathway. It is tightly related to the activity of cell wall acting compounds and reactive oxygen species [123]. SakA is a putative ortholog of Hog1 in *Saccharomyces cerevisiae*, which governs the high osmolarity glycerol (HOG) response. SakA seems to regulate the response to oxidative stress and high osmolarity [26]. *sakA* knock-out mutants show a significant increase in sensitivity to Caspofungin. Both, the CWI as well as the HOG pathway seem to play a major role in the drug response of *A. fumigatus*, especially concerning the paradoxical effect. Yet little is known if they interact with each other, and how they do it.

3.2 Material & Methods

3.2.1 NetGenerator

In 2005, Guthke *et al.* [43] presented the basic idea of the NetGenerator heuristic, which was implemented in 2007 by Töpfer *et al.* [119]. NetGenerator uses systems of linear or non-linear differential equations to model dynamic changes of gene expression. From dynamic time-resolved data, NetGenerator creates an interaction network that represents

the most significant interactions between those genes. It determines the coefficients of the model *via* a heuristic, that simulates the dynamic behaviour fitting to the time-resolved gene expression data. It has been successfully applied in modeling human [118] and fungal regulatory networks [42, 70, 71].

In 2013 Weber *et al.* published Version 2.0 of the algorithm [131]. The added features include the capability to work with data from multiple experiments and under multiple stimuli. The way in which prior knowledge is included has also been improved. In their publication, Weber *et al.* explicitly address the occurrence of cross-talk as object of study, which is a topic in this thesis as well.

The expression change of a gene i at the time-point t is defined as:

$$\dot{x}_i(t) = \sum_{j=1}^{N} \beta_{i,j} x_j(t) + \sum_{j=1}^{M} b_{i,j} u_j(t) \tag{3.1}$$

$\beta_{i,j}$ is the influence of a predictor j on the expression of gene i. The second term of the equation describes external influence on the gene expression. u_j is the j^{th} of M external perturbations, multiplied by $b_{i,j}$, which describes the strength of influence of u_j on gene i. The sign of the coefficients β determines if the influence is repressing ($\beta_{i,j} \leq 0$) or activating ($\beta_{i,j} \geq 0$). If $\beta_{i,j} = 0$, there is no influence of gene j on gene i. Typically, the expression of each gene is modeled by one single differential equation. NetGenerator is also able to model one gene with multiple equations, which is called increase of dynamic order. It enables NetGenerator to simulate even more complex behaviour, especially a time-delay of interaction.

While determining the network structure, the algorithm applies certain rules to achieve sparse networks and avoid over-fitting:

- Addition of a predictor must lead to significantly improved model fit,
- Removal of a predictor must not significantly decrease model fit,
- The number of predictors in the model must be smaller than the number of time points in the corresponding time series,
- The number of interactions must not exceed a user-set limit

It is difficult to estimate what is a "significant" increase or decrease in model fit. The default parameters consider a change by the factor of 0.2 as significant.

Due to the elaborate nature of calculation, large- or medium-scale modeling is not tractable. In order to find unique solutions for the differential equations, NetGenerator needs a minimal "difference" in the expression profiles of the genes. If the similarity of expression profiles is to high, the algorithm can not safely decide which gene is responsible for a certain regulation. By default, the maximum allowed correlation between two expression profiles is 0.95. This limits the maximum number of genes in the network, if no sufficient number of data points is provided. To increase the amount of data available, NetGenerator is able to interpolate values between data points using cubic splines.

During the determination of network structure, it is also possible to include prior knowledge (see chapter 1.2.1). The algorithm tries to makes all sub model structures

consistent with the offered prior knowledge. The integration of prior knowledge can be done using two ways: fixed and flexible. Interactions of the fix prior knowledge has to be part of the final model and can not be ignored by the inference. This should only be used with the most reliable of knowledge sources, and the user should be sure that the expression data also supports the demanded interaction. Otherwise, NetGenerator is not able to infer a network with the provided parameters. One should especially check the fixed prior knowledge for contradicting information. The flexible prior knowledge on the other hand is not required to be a part of the final model, as the algorithm can skip it, if it does not fit the data. Each interaction can be individually weighted for its confidence and in addition to that, there is also a general parameter defining the overall influence of the prior knowledge.

NetGenerator tries to fit the model by minimising the model error i.e. the difference between the simulated expression values and the actual measurements:

$$model\_error = \sum_{t=1}^{T} \sum_{i=1}^{N} (x_i(t) - \hat{x}_i(t))^2 \tag{3.2}$$

$\hat{x}_i(t)$ represents the simulated expression of gene i at time point t, which is subtracted from the measured value for that gene at that time point ($x_i(t)$).

In general, adding new interactions to the model decreases the model error, at the cost of an increased complexity. This leads to a conflict between modeling power and sparseness (simplicity) of the model. NetGenerator offers a parameter to set the allowed model error in the inference process. Since the algorithm stops adding new interactions once this allowed error is reached, the parameter serves as an indirect way to influence the number of connections in the model: The lower the allowed error, the more interactions the final model has. If the model error is set to low, however, NetGenerator may fail to create a model that satisfies the given criteria.

3.2.1.1 Parameters

The NetGenerator offers various parameters to optimise the modeling of the network. Three of the main parameters are:

allowedError: The allowed model error compared to the expression data.
weightingStruct: The global weight of the prior knowledge.
maxDynamicOrder: The dynamic model order.

The parameter *allowedError* regulates, how well NetGenerator tries to fit the model to the expression data. While it is of course desirable to fit the data as good as possible, NetGenerator is not always able to achieve the desired fit. This causes the algorithm to crash, giving no results. The model fit generally increases as more connections are included into the model. Enforcing a good fit can therefore cause a comparably large network, without the desired sparseness.

In comparison to the individual confidence weighting of the prior knowledge, the *weightingStruct* parameter influences the global effect of all prior knowledge. This means,

how much the algorithm insists on including a prior knowledge interaction in favor of another interaction, even if the other may yield a better fit. A *weightingStruct* value of 1 does not prefer any prior knowledge, while a value of 0 greatly insists on adding prior knowledge interactions.

The third parameter I estimated is *maxDynamicOrder*. It determines how many equations are used to explain the expression of a gene. This allows the sub-model to use a higher order integrator chain, which can be considered hidden states in order to model more complex behaviour. A common biological interpretation of this higher order is a delay in the expression.

3.2.2 Robustness tests

It is often difficult to tell how robust an interaction in an inferred network is. Here, robustness describes the likelihood of a certain interaction to appear, even if there are small changes in the data. Robustness tests ensure, small alterations in the data and prior knowledge do not lead to completely different models.

I performed two types of robustness tests on the final network. One testing for noise to the data, since one should always assume a little measurement error. To check how susceptible a network is to perturbations of the data, I added normal distributed noise with a mean = 0 and a standard deviation of 0.01 to the pre-processed data. After that, I repeated the inference and counted how many interactions of the original network were inferred again. I did this 1000 times and considered all interactions, that occur in an least 50% of the networks, as robust to noise.

I performed a similar test to check the reliability of an interaction on prior knowledge. I repeated the network inference but randomly omitted 10% of the prior knowledge. This test was also repeated 1000 times and an interaction of the original network that occurs in at least 50% of the inferences is considered robust to perturbation of the prior knowledge.

At the end, an interaction is considered (truly) robust, if it passed the test against both, noise in the gene expression data and missing prior knowledge.

3.2.3 RNA-Seq data

The RNA-Seq data used in this study comes from the *A. fumigatus* strain **CEA10**, also known as A1163. From this strain samples were taken for the wild type (wt) and for four different knock-out mutants: Δ*akuB*, Δ*mpkA* and Δ*sakA*, as can be seen in table 3.1. The conidiophores were grown for 16 h on a minimal media, and then treated with Caspofungin at a dose of 0.1 μg/ml. From the wt and Δ*akuB* strains, samples were taken at five different time points after treatment 0.5 h, 1 h, 4 h, 8 h and 24 h. From the Δ*sakA* and Δ*mpkA* strains, samples were taken at 4 h after treatment. As control, additional samples of all strains were taken before the administration of Caspofungin. All samples were taken in replicates from three different flasks. The measurements were taken using an *Illumina HiSeq2000* sequencing machine with single end technology.

Table 3.1: **RNA-Seq data used in this study.** For the wild type (wt) and the Δ*akuB* mutant, measurements were taken at six time points, whereas for the Δ*mpkA* and the Δ*sakA* mutant three time points were taken. All samples were taken in three replicates.

strain	wt	Δ*akuB*	Δ*mpkA*	Δ*sakA*
0 hour	3 ×	3 ×	3 ×	3 ×
0.5 hours	3 ×	3 ×		
1 hour	3 ×	3 ×		
4 hours	3 ×	3 ×		
8 hours	3 ×	3 ×		
24 hours	3 ×	3 ×		

3.2.4 Mapping of transcription data

The reads were provided in FASTA-files. TopHat v1.4.1 [121] was used with a butterfly search to map the reads to the genome. The reads were mapped to the genome sequence of the *A. fumigatus* A1163 strain provided by the CADRE [36] database. A CADRE gene transfer format (GTF) file of the same strain was used to annotate the mapped reads. The number of allowed multi-hits was set to 1 to prevent multiple mapping. After investigating the GTF file, the maximum intron length was set to 2700 bp. The reads that mapped successfully were sorted and indexed using SAMtools [68].

3.2.5 Differentially expressed genes

The indexed hits were read into *R* [115] using the easyRNA-Seq package [23]. The chromosome size was calculated from the CADRE genome sequence file and the read counts were calculated. BioMart [57] provided exon annotation. The *DESeq* package [6] was used to normalize the read counts for different library sizes. This is necessary since samples with a bigger library size (i. e. sequencing depth) can have more reads for a given gene then samples with a lower library size, even though the actual expression of the gene did not change [25]. The data contains read counts for 10.160 genes.

Differential expression analysis was performed *via* the "*nbinomTest*" function of the *DESeq* package. Genes were considered differentially expressed, with a FDR-adjusted p-value ≤ 0.05. For the wild type and Δ*akuB* strains, the time points 0.5 h, 1 h, 4 h, 8 h and 24 h were compared to the control (0 h). For the Δ*sakA* and Δ*mpkA* strains the time points 1 h, 4 h were compared to the control.

3.2.6 Collection of prior knowledge

In order to implement as much of the currently available information on the *A. fumigatus* pathway as possible, I followed a knowledge-driven approach. The prior knowledge is used to propose known interactions to the network and guide it. Those interactions are softly integrated, which means, the inference algorithm only implements them, if they fit to the provided expression data. I utilised two types of sources:

3.2.6.1 RNA-Seq data

I investigated the differential expression of the genes in the $\Delta mpkA$ and $\Delta sakA$ mutant strains. I compared the expression of the genes in the knock-out mutants with the wild type at time point 0 h. If a gene x is differentially expressed with an FDR-adjusted p-value ≤ 0.05 in the $\Delta sakA$ mutant, a regulatory interaction between x and $sakA$ was assumed. If x is down regulated in $\Delta sakA$, I assume an activating interaction $sakA \rightarrow x$, otherwise a repressing interaction $sakA \dashv x$. The same method was used to identify potential interactions in the $\Delta mpkA$ strain.

3.2.6.2 Literature

In addition to the RNA-Seq data, I also searched in the literature. A valuable source of prior knowledge is the publication of [95], a review written in 2009 by Rispail *et al.* It compares the MAP kinase and calcium-calcineurin pathway in different pathogenic fungi and plants. It includes orthologous genes for *sakA* (*hog1*) and *mpkA* (*mpk1*). While it already is an extensive study of the pathway, the information is not species specific and no cross-talk between *sakA* and *mpkA* was investigated. The names of the orthologous of *S. cerevisiae* were translated to the *A. fumigatus* naming system.

3.2.7 Biological validation

3.2.7.1 Western blot

The phosphorylation of a protein can be measured using western blot analysis. Here, protein samples are placed on a membrane and diffuse through it powered by gel electrophoresis. Special antibodies are given to the sample, that bind to the target protein and stain them for later evaluation. While diffusing through the gel, the proteins in the sample are separated by their molecular weight. The result is a quantitative analysis of the target proteins.

In this study, conidia of *A. fumigatus* were incubated for 18 h and then stressed with Caspofungin. Samples were taken before treatment as well as 0.5 h, 1 h, 2 h, 4 h and 8 h after treatment. Protein samples were extracted and the phospho-p38 MAPK antibody was used to detect SakA in its phosphorylated form, while phospho-p44/42 MAPK was used to detect phosphorylated MpkA.

3.2.7.2 Polymerase chain reaction

A polymerase chain reaction (PCR) is used to determine DNA concentrations in *vitro*. The PCR uses the DNA-polymerase to amplify the amount of DNA to detectable levels. One cycle of the amplification consists of three major steps:

1. **Denaturation:** At 95 °C, double stranded DNA-helix melts into two.

2. **Primer annealing:** The temperature is lowered to 50-65 °C and the primers, which bind to the part of DNA to be amplified, bind to the DNA-polymerase.

3. **Elongation:** The DNA-polymerase synthesises the missing strand with free nucleotides from the medium, starting form the primer. The primer remains on the DNA-strand, so it can be used again in the next cycle.

The cycle starts again, and is repeated until enough DNA material has been synthesised. This is usually after 20-50 cycles. The expression of the genes are then normalised to housekeeping genes, which are expected to not alter their expression.

An extension to this is the real-time or quantitative PCR (qPCR). Here, a fluorescent dye is given to the medium, which is activated during the elongation process. This allows to determine the amount of synthesised DNA in real-time.

To determine the activity of specific genes, measuring the RNA of the cell directly instead of DNA is advisable. In the reverse-transcriptase PCR (RT-PCR) directly uses the RNA-samples from the data, which are transformed into complementary DNA (cDNA) and then used for the classic PCR.

The results are often shown in Fold difference $2^{(-\Delta\Delta Ct)}$. $\Delta\Delta$Ct is the expression of the treated protein minus the experession of the untreated control.

The combination of quantitative and reverse transcription PCR is abbreviated with qRT-PCR.

3.2.7.3 Measurement of Rhodamine-123 uptake

Rhodamine-123 is a fluorescent dye that absorbs light in a spectrum of 505-560 nm. To measure the intracellular permeability of the fungal cell wall, Conidia (10^6/ml) were cultivated for 16 h in minimal media at 37°C, 200 rpm. A final concentration of 20 μM of Rhodamine-123 was added to the growing mycelia alone or in combination with Caspofungin. Samples were taken at the time of application as well as 0.5 h, 1 h, 2 h, 4 h and 8 h. The mycelium was ground with mortar and a pestle using liquid nitrogen. Then cytosolic fraction was extracted and measured using a TECAN microplate reader according to the manufacturer's instructions.

Caspofungin resistant strain Among the strains tested in the Rhodamine-123 study is also the Caspofungin resistant strain $EMFR^{S678P}$ [96]. The resistance is confered via a point mutation in *fks1*, at nucleotide position 2086 (thymin to cytosin).

3.3 Results

3.3.1 Differential expression & clustering

To test the quality of the data, I tested the correlation between the replicates. I used read counts normalised for the library size for the comparison. The mean Pearson correlations of all four strains are listed in table 3.2. The results show that the biological variance between the replicates is very small, since all scores are at 0.89 or above, except for the 24 h samples. Most correlations are actually around 0.99, indicating the high quality of the data.

Table 3.2: **Mean correlation between the biological replicates of the RNA-Seq data.** No samples were available for the Δ*sakA* and Δ*mpkA* strains at the time points 0.5 h, 8 h and 24 h. The except for the 24 h time point, the correlation is always very high.

	0 h	0.5 h	1 h	4 h	8 h	24 h
wild type	0.99	0.98	0.89	0.96	0.99	0.74
Δ*akuB*	0.99	0.98	0.89	0.99	0.99	0.76
Δ*mpkA*	0.99	-	0.96	0.96	-	-
Δ*sakA*	0.99	-	0.99	0.99	-	-

The first step in the investigation of the data consists of the determination of all genes affected by the Caspofungin treatment. So I identified all differentially expressed genes (DEGs) for all time points in the wild type (wt) strain (see 3.2.3 for details). For a p-value of 0.05, I found 6037 DEGs. This number is quite high, however, one should not forget, that the samples were taken over a comparably long time and the Caspofungin treatment has a strong effect on *A. fumigatus*.

As the next step in the investigation of the data, I clustered all DEGs in the wt-strain. I used the *fuzzy c-means* algorithm of the *R* package *e1071* [85]. I checked ten different cluster sizes, ranging from 2 to 12 groups and calculated the *CVI* [65]. I selected the cluster size with the highest value, which was a clustering into nine groups. Looking at the results of the clustering in figure 3.1, it is evident that one of the biggest expression change, compared to the untreated sample, happens at time point 24 h. By that time, the fungus should have adopted to the Caspofungin treatment and the expression is expected to return to "normal". The reason why this did not happen is most likely that *A. fumigatus* is facing another environmental stress: starvation. After 24 h, the nutrients in the flask start to diminish and the fungus has to change its metabolism to adopt to the new situation. This is most evident in the cluster 2 and 9, which contain genes that show no particular response to Caspofungin but react in the last time point. These two clusters are also the ones with the most members. The different expression

profile of time point 24 h can also be seen in the dendrogram for the wild type heat map in figure 3.3.

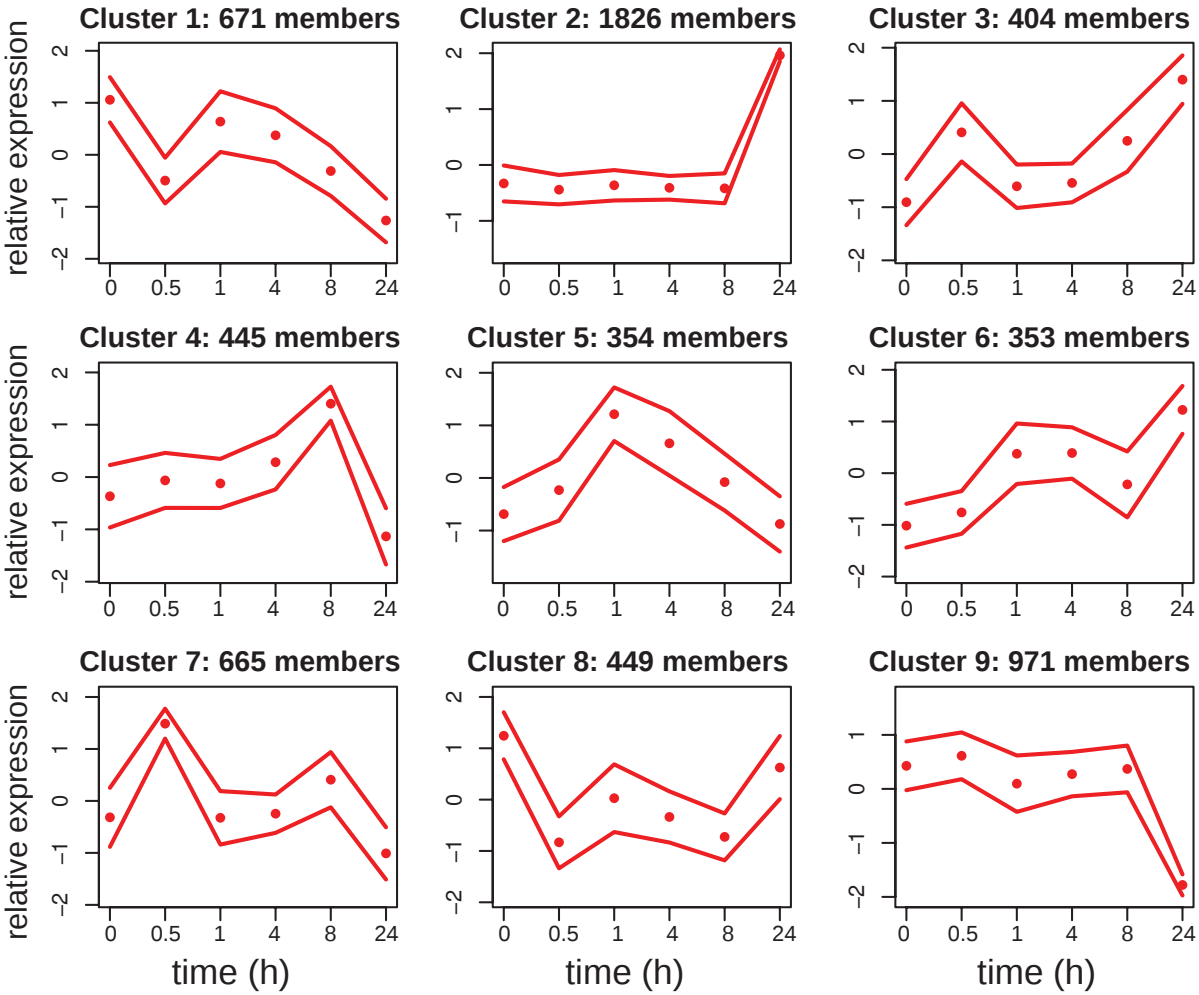

Figure 3.1: **Result of the clustering with all time points.** Cluster 2 and cluster 9 have the most members. In these clusters are mostly genes, that are mainly influenced by the starvation effect.

To investigate the claim that many genes are differentially expressed because of the starvation effect, I repeated the clustering for all genes that are differentially expressed in any but the last time point. The number of DEGs decreases to 4257. The result of the new clustering can be seen in figure 3.2. This time, the optimal *CVI* was reached at a cluster size of four. Of special interest here are the clusters one and four, as they show that at time point 0.5 h there is a distinct reaction to the drug stress, after which the regulation follows the general trend. This early response is most likely tied to drug specific response genes. Their regulation normalises later on, when the drug stress diminishes to prevent over-expression.

The goal of this study revolves around the study of the Caspofungin stress response of *A. fumigatus*, and not the hunger stress. I decided to omit the last time point (24 h)

from the data for the rest of the study, as it would distort the inference. The lower correlation of that data point would also weaken the statistical analysis.

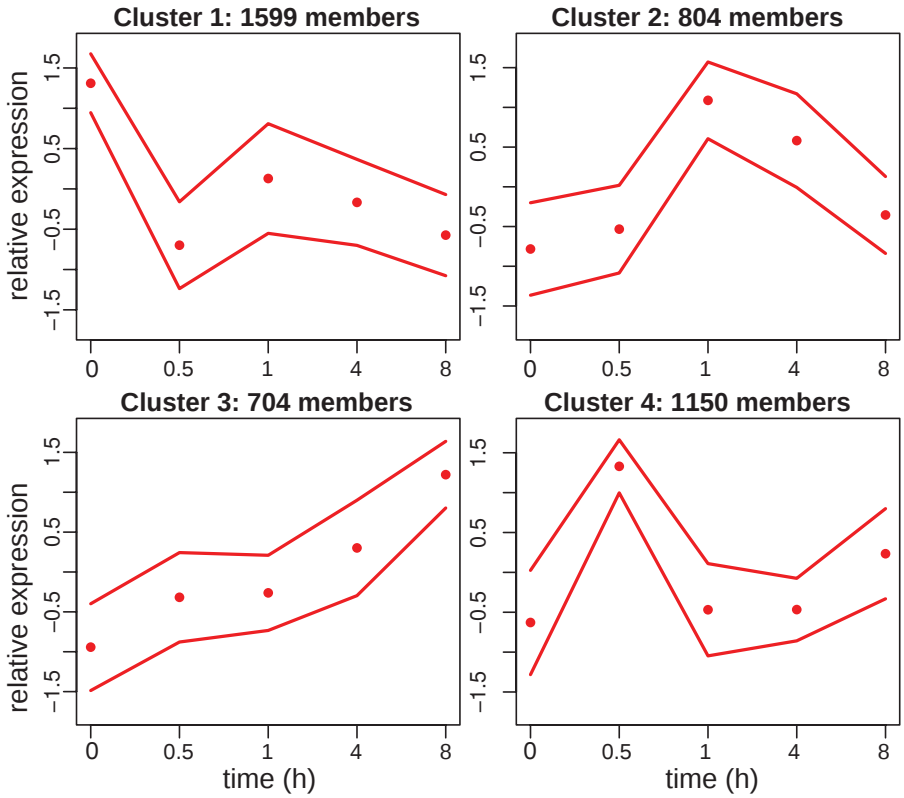

Figure 3.2: **Result of the clustering without time point 24 h.** Here, the last time point (24 h) of the expression data has been omitted. Of special interest is the time point 0.5 h, which shows an exception of the general trend in the clusters 1 and 4.

3.3.2 Gene selection

Since the 4257 DEGs of the first four time points are much more than the NetGenerator can handle, the selection of model genes has to be decreased further. The problem here is that with only five measurements, many genes will have very similar expression profiles. If two or more genes are potential regulators of another gene, and they all show the same expression changes over time, NetGenerator can impossibly decide which to chose. Judging from personal experience of colleagues and the maintainers of NetGenerator, I considered a number between 20 and 30 genes as manageable. To find genes connected to the response to Caspofungin treatment, I performed a *Gene Ontology* (GO) [7] analysis using the *FungiFun* tool [93]. The cut-off for the enrichment was set at a p-value of 0.05. For only 1876 of the genes an annotation was found. I selected genes associated with the terms *cell wall* and *membrane*, as well as *β-glucosidase* and *glucan synthase*, since the *A. fumigatus* cell wall is composed of sugars and Caspofungin targets these sugars and

these genes may play a role in its construction and maintenance. As *mpkA* and *sakA* are *protein-kinases* and known to be part of the response, genes with this term were also considered. Additionally, genes assigned to *signaling* and *transmembrane transport* were selected, since these categories are known to be connected to the drug response and response to cell wall perforation.

By far the largest group of DEGs belonged to *membrane* with 323 genes, whereas only eight belong to *cell wall*. 193 genes are assigned to *transmembrane transport* and 25 to *signal transduction*. 86 belong to the *protein-kinase* group and five *response to stress*. 15 belong to the term *glucan synthase* or *glucosidase*. This makes a total of 655 genes, which is still to much for the inference. We took the work of Rispail *et al.* [95], who worked on a similar subject, as guideline for further selection. DEGs mentioned in this publication are involved in the response regulation and were taken as hints for the gene selection. The final decision was done by my colleague Dr. Vito Valiante, who is also a co-author of Rispails publication. The genes that were considered for the modeling are listed in table 3.3 and figure 3.3 shows the expression profiles for all model genes.

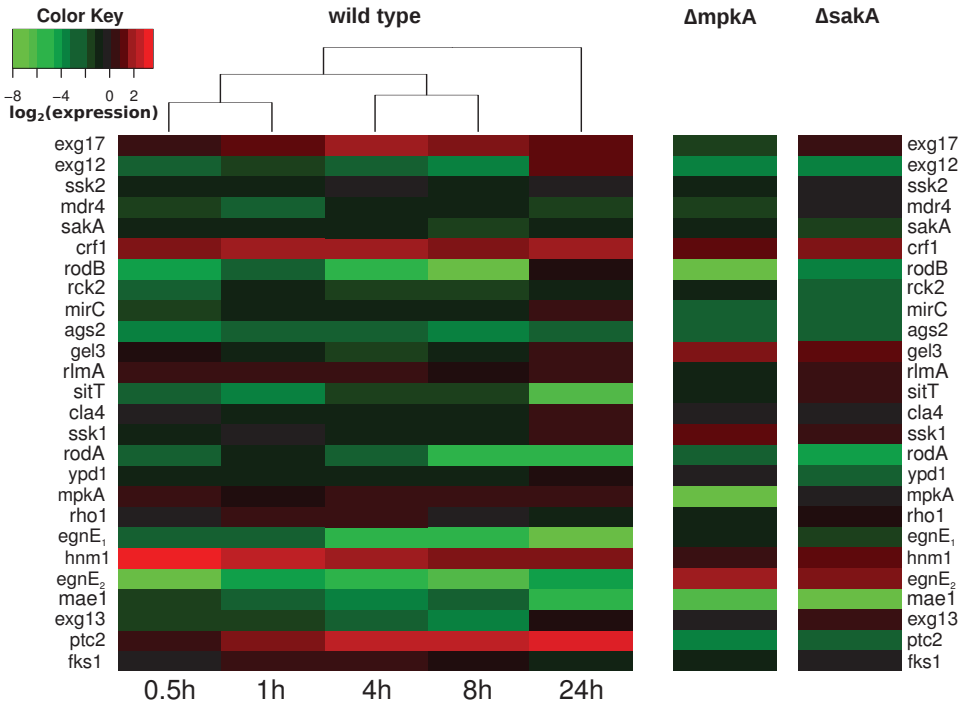

Figure 3.3: **Differential expression of genes selected for modeling.** The expression of all genes was compared to the untreated wild type strain. The expression values are $\log_2$ transformed. For the wild type strain, a dendrogram was created, which shows that the time point 24 h has the least similarity to the other time points considering the expression.

Table 3.3: **Genes of the model.** The table presents the names, description and GO-categories. The last column indicates, if the gene was mentioned in Rispail *et al.* [95]

Name	Systematic ID	Description	GO-category	Rispail
rodA	AFUB_057130	Conidial hydrophobin	cell wall	no
rodB	AFUB_016640	Conidial hydrophobin	cell wall	no
crf1	AFUB_015530	Extracellular cell wall glucanase /allergen	cell wall	no
rlmA	AFUB_040580	Cell wall organization, response to stress, Trans. Factor	cell wall	yes
gel3	AFUB_028470	1,3-β-glucanosyltransferase	membrane	no
rho1	AFUB_072830	GTPase; involved in radial growth and conidiation	membrane	yes
exg13	AFUB_091720	β-D-glucoside glucohydrolase	β-glucosidase	no
exg12	AFUB_006160	β-glucosidase	β-glucosidase	no
exg17	AFUB_000280	β-glucosidase	β-glucosidase	no
fks1	AFUB_078400	1,3-β-glucan syntase, target of Caspofungin	β-glucosidase	yes
egnE$_1$	AFUB_079180	Mutanase	glucan synthase	no
egnE$_2$	AFUB_081470	Mutanase	glucan synthase	no
ags2	AFUB_027030	α-1,3-glucan synthase	glucan synthase	no
mdr4	AFUB_012160	ABC multidrug transporter	transmembrane transport	no
hnm1	AFUB_081190	Amino acid permease	transmembrane transport	no
mae1	AFUB_082870	C4-dicarboxylate transporter/malic acid transport protein	transmembrane transport	no
mirC	AFUB_022760	Siderophore transporter	transmembrane transport	no
sitT	AFUB_044820	ABC multidrug transporter	transmembrane transport	no
ssk1	AFUB_055940	Response regulator, two-component signal transduction	transcription factor	yes
ssk2	AFUB_010360	Ortholog have role cellular response to osmotic stress	protein kinase activity	yes
mpkA	AFUB_070630	MAPK; expression increased by cell wall damage	protein kinase activity	yes
sakA	AFUB_012420	MAP kinase	protein kinase activity	yes
rck2	AFUB_020560	Calcium/calmodulin-dependent protein kinase	protein kinase activity	no
cla4	AFUB_053440	Ortholog(s) have protein serine/threonine kinase activity	protein kinase activity	yes
ypd1	AFUB_067390	Histidine-containing phosphotransfer intermediate protein	signal transduction	yes
ptc2	AFUB_101270	Predicted catalytic activity, phosphatase	catalytic activity	yes

3.3.3 Comparison wild type & Δ*akuB* strain

The gene *akuB* codes for a DNA helicase that is part of the non-homologous end-joining machinery in *A. fumigatus*. This makes the Δ*akuB* strain much more suitable for homologous deletion [21]. This knock-out is a necessary step to create the *sakA* and *mpkA* knock out strains. This raises the question of how much this deletion alter the expression of other genes, bringing a bias into the data.

To investigate the global effect of the deletion of *akuB* I compared the expression profiles of the wild type strain with the expression profile of the Δ*akuB* strain. I compared each of the six time points individually and calculated the Pearson and the Spearman correlation of the log_2 transformed read counts.

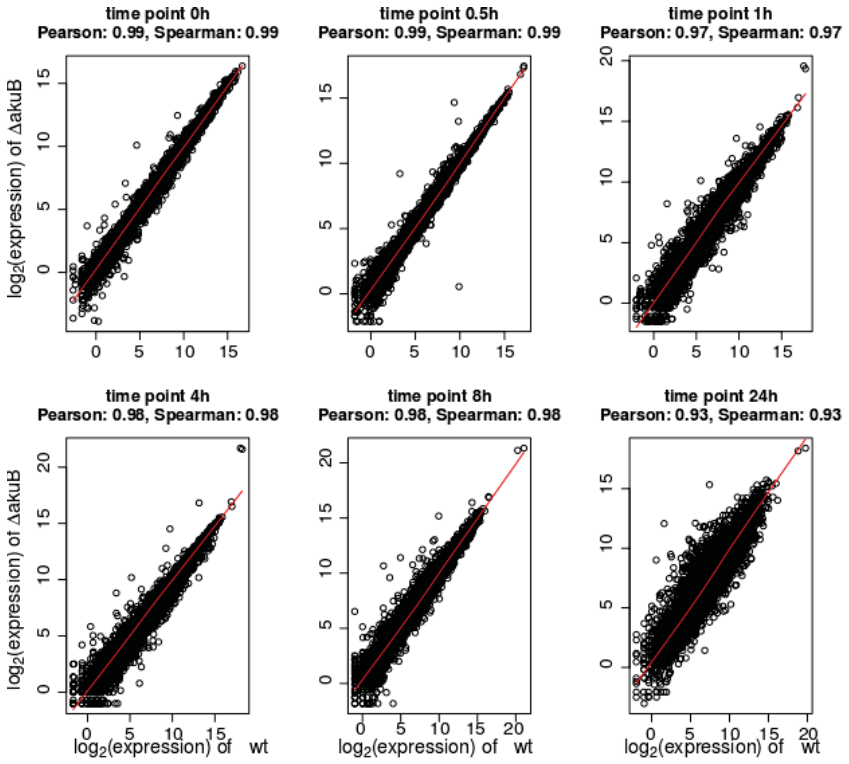

Figure 3.4: **Comparison of log_2 fold changes for the wild type (wt) and the Δ*akuB* mutant strain.** The high correlation indicates that the deletion of the *akuB* gene does not have significant effects on global Caspofungin response.

The results can be seen in figure 3.4. They show that during Caspofungin stress the wild type and Δ*akuB* strain have very similar expression patterns, with Spearman correlations ranging from 0.93 - 0.99. According to this test, the deletion of *akuB* does not have a significant impact on the global Caspofungin response, and the data from the Δ*sakA* and Δ*mpkA* mutants, which also contain the *akuB* deletion, is probably not biased.

To further investigate the influence of the knock-out mutants, I also calculated the

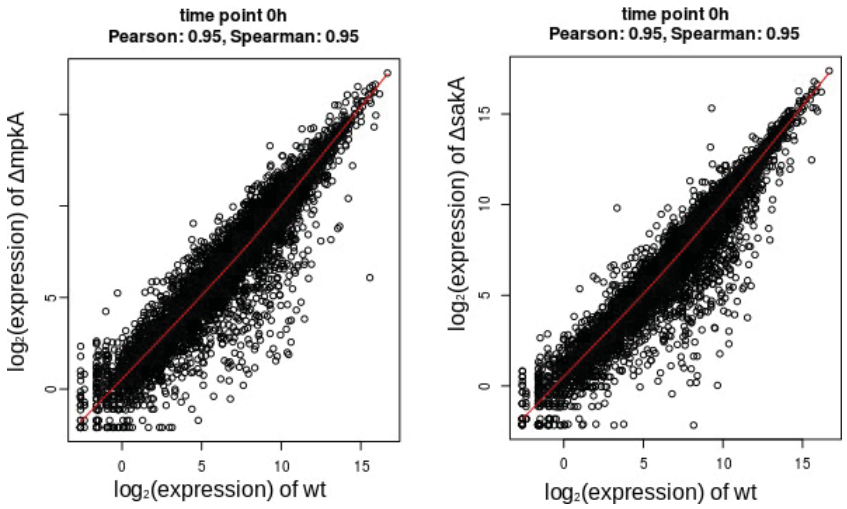

Figure 3.5: **Comparison of log_2 fold changes for the wild type (wt) and the Δ*mpkA* and Δ*sakA* mutant strains.** There is a high correlation between the wt and the knock-out mutants.

correlation of the wt with the other knock-out mutants, as seen in figure 3.5. The Spearman correlation is always quite high, ranging from 0.88 - 0.95 for the comparison with the Δ*mpkA* strain and 0.95 - 0.96 for the Δ*sakA* strain. Since the deletion of both genes should only influence the response pathway, I did not expect global changes in expression.

3.3.4 Harvest of prior knowledge

As described in chapter 3.2.3, I used the RNA-Seq data of the Δ*mpkA* and Δ*sakA* knock-out mutants to determine the prior knowledge for these genes. The idea is that genes, that are differentially expressed in the knock-out mutants versus wt, must be influenced by the knocked-out gene. The results can be seen in figure 3.6 and table 3.4. By this analysis, I identified 14 potential interactions for *mpkA*, of which 11 are activating, and three are repressing. For *sakA*, I found 15 of the model genes to be differentially expressed. Nine have an activating regulation from *sakA*, while five are repressed. This leads to a total of 29 RNA-Seq-based prior knowledge interactions.

Additionally, I searched in literature for known interactions between the model genes. The central source here was the publication by Rispail *et al.* [95]. From this source, I was able to harvest a total of seven additional interactions. Six of them are activating, while one is repressing. The literature offered prior knowledge with other sources than *sakA* and *mpkA*. Nevertheless, this prior knowledge adds another interaction for *sakA* and *mpkA* each. The interaction *mpkA* $\rightarrow$ *rlmA* is the only one that I found in both, literature and RNA-Seq data of the knock-out mutants. This prior knowledge from literature is very important, since it represents current knowledge about the gene regulation in *A. fumigatus*. It also adds "causal" information, like the connection between the

transcription factor *ssk1* and its target *ssk2*. This information can not be derived purely from the data. The amount of the current knowledge for these model genes, however, is very small. This is a general problem in network reconstruction.

After collecting all the prior knowledge, I had to decide if I want to weight all prior knowledge interactions equally. This decision is necessary so that the algorithm can deal with ambiguous situations or even contradicting prior knowledge. My set of prior knowledge does not contain contradicting information, yet I still differentiated between the literature knowledge and that extracted from the RNA-Seq data. The literature data was tested in the laboratory. This is why I decided to give the prior knowledge extracted from RNA-Seq data a medium value of 0.5 and a slightly higher value of 0.66 to the literature knowledge.

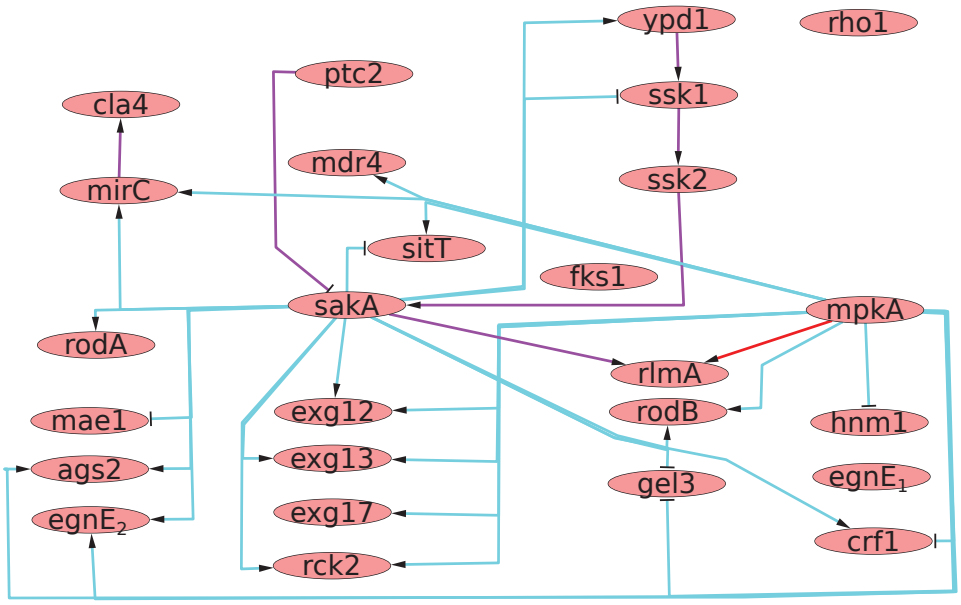

Figure 3.6: **Prior knowledge used in this study.** The prior knowledge was taken from two different sources, RNA-Seq data (turquoise) and literature (purple) [95]. The interaction *mpkA* → *rlmA* is the only one that is supported by both prior knowledge sources (red).

3.3.5 NetGenerator

3.3.5.1 Preprocessing

Before application of the NetGenerator, the data has to be prepared for analysis. As stated in chapter 3.3.1, the 24 h time point is omitted from the data, since the influence of the starvation effect is to big and creates a bias. I normalised the reads for the different library sizes with the *DESeq* package [6] in *R*.

The NetGenerator assumes that the organisms is in a steady state at the beginning and end of the study. Also, it is important to consider the relative expression changes

Table 3.4: **List of prior-knowledge used in this work.** The table contains the standard names for the regulators, targets, whether the interaction was activating/inhibiting as well as the confidence score that was assigned to it. The column "in model" shows if the interaction was found in the final model. Only the interaction *mpkA* → *rlmA* was found in prior knowledge based on RNA-Seq data of knock-out data and literature.

source	type	target	confidence	source	in model
sakA	→	*rlmA*	0.66	Rispail *et al.* [95]	**yes**
ypd1	→	*ssk1*	0.66	Rispail *et al.* [95]	no
ssk1	→	*ssk2*	0.66	Rispail *et al.* [95]	**yes**
ssk2	→	*sakA*	0.66	Rispail *et al.* [95]	no
ptc2	⊣	*sakA*	0.66	Rispail *et al.* [95]	**yes**
mirC	→	*cla4*	0.66	Rispail *et al.* [95]	**yes**
mpkA	→	*rlmA*	0.66	Rispail *et al.* / Δ*mpkA* mutant	**yes**
mpkA	→	*mdr4*	0.5	Diff. expr. in Δ*mpkA* mutant	**yes**
mpkA	→	*sitT*	0.5	Diff. expr. in Δ*mpkA* mutant	**yes**
mpkA	→	*ags2*	0.5	Diff. expr. in Δ*mpkA* mutant	no
mpkA	→	*exg13*	0.5	Diff. expr. in Δ*mpkA* mutant	no
mpkA	→	*exg12*	0.5	Diff. expr. in Δ*mpkA* mutant	no
mpkA	→	*exg17*	0.5	Diff. expr. in Δ*mpkA* mutant	no
mpkA	→	*rck2*	0.5	Diff. expr. in Δ*mpkA* mutant	no
mpkA	→	*rodB*	0.5	Diff. expr. in Δ*mpkA* mutant	**yes**
mpkA	→	$egnE_2$	0.5	Diff. expr. in Δ*mpkA* mutant	no
mpkA	→	*mirC*	0.5	Diff. expr. in Δ*mpkA* mutant	no
mpkA	⊣	*crf1*	0.5	Diff. expr. in Δ*mpkA* mutant	**yes**
mpkA	⊣	*hnm1*	0.5	Diff. expr. in Δ*mpkA* mutant	**yes**
mpkA	⊣	*gel3*	0.5	Diff. expr. in Δ*mpkA* mutant	**yes**
sakA	→	*ags2*	0.5	Diff. expr. in Δ*sakA* mutant	**yes**
sakA	→	*exg13*	0.5	Diff. expr. in Δ*sakA* mutant	**yes**
sakA	→	*exg12*	0.5	Diff. expr. in Δ*sakA* mutant	**yes**
sakA	→	*rck2*	0.5	Diff. expr. in Δ*sakA* mutant	**yes**
sakA	→	*rodA*	0.5	Diff. expr. in Δ*sakA* mutant	**yes**
sakA	→	*rodB*	0.5	Diff. expr. in Δ*sakA* mutant	**yes**
sakA	→	*crf1*	0.5	Diff. expr. in Δ*sakA* mutant	no
sakA	→	$egnE_2$	0.5	Diff. expr. in Δ*sakA* mutant	**yes**
sakA	→	*mirC*	0.5	Diff. expr. in Δ*sakA* mutant	**yes**
sakA	→	*ypd1*	0.5	Diff. expr. in Δ*sakA* mutant	no
sakA	⊣	*cla4*	0.5	Diff. expr. in Δ*sakA* mutant	**yes**
sakA	⊣	*gel3*	0.5	Diff. expr. in Δ*sakA* mutant	**yes**
sakA	⊣	*sitT*	0.5	Diff. expr. in Δ*sakA* mutant	no
sakA	⊣	*ssk1*	0.5	Diff. expr. in Δ*sakA* mutant	no
sakA	⊣	*mae1*	0.5	Diff. expr. in Δ*sakA* mutant	**yes**

during the treatment, and not the absolute ones. When it comes to regulation by gene expression, even small differences can cause big effects. The preprocessing was done for each gene and each series of replicates individually.

So I first centered the data by dividing the reads to the expression at time point 0 h. Next, I used log_2 transformation so the values revolve around 0. As a last step, the values are scaled to fall into the interval [-1, 1] by dividing them trough the maximum of the absolute values. The resulting relative expression values represent the change in expression starting from 0 (steady state) at time point 0 h and then gives the direction (activation or repression) of the expression change.

A feature of NetGenerator is the ability to work with independent experiment time series. In this study, every replicate is considered an individual experiment. The NetGenerator algorithm will try to optimise the fit over all time series.

To implement the prior knowledge into NetGenerator, two matrices were created, one for the direction (activating or repressing) and one for the confidence of the explicit interaction.

3.3.5.2 Parameter estimation

As described in chapter 3.2.1.1, I estimated three different parameters to find a model that implements many prior knowledge interactions, finds a good model fit while remain sparse in interactions. The parameters were checked for the following values:

maxDynamicOrder:
1, 2

weightingStruct:
0.000, 0.001, 0.010, 0.050, 0.100, 0.200, 0.300, 0.400, 0.500, 0.600, 0.700, 0.800, 0.900, 1.000

allowedError:
0.001, 0.010, 0.050, 0.100, 0.200, 0.300, 0.400, 0.500, 0.600, 0.700, 0.800, 0.900, 1.000

This leads to 364 different network models. While generating the different networks, the algorithm did not converge a total of 71 times. As mentioned before, some parameter combinations do not lead to a successful modeling. For example when the algorithm is not able to fit the coefficients enough to meet the *allowedError* restriction. This is especially difficult, if a high weighting of prior knowledge tries to force many prior knowledge interactions into the model. This also indicate that the inference may encounters situations, were the correlation between the expression profiles of potential regulators is to high, and the algorithm cannot decide which one to use. As an example, the expression profile of *ags2* and *egne₂* are very similar. Comparing the mean expression of all time points, these genes have a Pearson correlation of 0.99. NetGenerator does not take the mean expression of all time points but each replicate individually, but still, situations can occur, where this is still not enough.

3.3.5.3 Model selection

Given the number of parameter combinations, minus the number of failed attempts, NetGenerator produced 293 different network model, varying in size (number of connection) and topology. Selecting the correct model from this set of possibilities is an important step towards a reliable network prediction. I considered three different properties of the network for evaluation.

Included prior knowledge:
How many prior knowledge interactions were implemented into the model?

Model error:
How much does the simulation deviate from the expression data?

Degree of sparseness:
How many interactions were implemented in comparison to the number of nodes?

A small model error and high sparseness of the model are often contradicting goals. A lower model error is achieved by including additional interactions to the network, which decreases the sparseness. The number of implemented prior knowledge interactions determines, how close the network is to current knowledge.

To get a global view over the created networks, I calculated the minimal, maximal and average values for the model error, the number of implemented prior knowledge interactions and the total number of connections (see table 3.5). Even from these numbers the results vary greatly. The achieved model error ranges from 1.6 for the best fit 11.2 for the worst. The intermediate values are evenly distributed, and it is no surprise, that the values grow larger, the higher the *allowedError* is.

Table 3.5: **Global results for the inferred networks.** Showing the minimal, average and maximal values for the model error, number of implemented prior knowledge and network size (min / average / max).

model error	prior knowledge	# of connections
1.6 / 5.5 / 11.2	0 / 5 / 26	0 / 21 / 102

The best fit to data with a model error of 1.6 was reached by a model consisting of 102 interactions. It implemented 4 prior knowledge interactions and one model interaction contradicts the prior knowledge. This leads to a $\frac{\#\text{ of prior knowledge}}{\text{total }\#\text{ of interactions}}$ ratio of 0.06. Not only does this network represent very little of current knowledge, it even contradicts the exploratory analysis of the RNA-Seq data. In the investigation of the Δ*sakA*, I found a up-regulation of *gel3*. I therefore assume that *gel3* may be repressed by *sakA*. This network suggests the opposite, which is an activation of *gel3* by *sakA*. While I do not consider RNA-Seq-based prior knowledge as very confident and do not insist on implementing it into the network, contradicting this prior knowledge means to ignore the information we have gained from the knock-out data.

Table 3.6: **Comparison of different models**. The first column holds the results of the mode with the smallest model error and the second column the model with the highest number of implemented prior knowledge interactions. The last column shows the rerun of the second model without prior knowledge provided.

	lowest model error	most prior knowledge	started without prior knowledge
model error	1.6	2.3	1.6
prior knowledge	4	26	5
# of connections	102	95	100
maxDynamicOrder	1	2	2
weightingStruct	0.01	0.3	-
allowedError	0.001	0.01	0.01

The highest number of implemented prior knowledge interactions by any model is 26. This network has a model error of 2.3 and consists of 95 interactions, which leads to a $\frac{\text{\# of prior knowledge}}{\text{total \# of interactions}}$ ratio of 0.37. That means around one third of the model is supported by an additional data source. The model error is only slightly higher than the best value, while having less connections. In contrast to the model with the lowest error, this network uses second order dynamics, to fit the data. The second order has been used to simulate the expression of the genes *cla4, mirC, ssk1* and *gel3*. A number of 23 of the 26 considered genes have an autoregulation which is always negative. 19 genes are connected to the input, which represents the Caspofungin treatment. Seven of these connections are activating and 12 are repressing.

I selected this network as the final model, since it implements most of the current knowledge and the model error is only slightly higher than the best value.

3.3.5.4 Robustness assessment

To further increase the reliability of the network, I performed robustness tests concerning noise in the data and dependency to prior knowledge. The technical details can be seen in chapter 3.2.2. As a result, 18 interactions have been found to be not robust. 11 of these interactions passed the test for prior knowledge but failed the test for noisy data. The interaction *ssk2* $\dashv$ *rlmA* passed the test against noise in the data but was to dependent on prior knowledge. Additional six interactions failed both tests.

Among those which failed, is the autoregulation of *ssk1* which also looses another outgoing and incoming interaction. The autoregulation of the other genes seem to be very robust against noise. From the eight interactions of *ssk2* (four outgoing, three incoming

and one autoregulation), four fail the robustness test, including the prior knowledge supported activation of *sakA*.

ypd1 and *cla4* loose their only outgoing connection (except autoregulation), leaving them without target.

In total, four interactions supported by prior knowledge were found to be not robust against noise, while they were still robust against perturbation of the prior knowledge.

All interactions starting at the input proved to be robust. The robust network now contains 77 interactions, including autoregulation.

3.3.6 Model assessment

3.3.6.1 Simulated expressions

In the gene expression data, some genes show an early response reaction to the Caspofungin treatment (see figure 3.2). This early response, which can also be seen in *mpkA* and *sakA* in figure 3.7 is quite difficult to model for linear systems. Nevertheless, NetGenerator was able to fit these data points quite well. For some genes like *ssk2* the use of a second order was necessary to correctly model the expression.

Looking at the simulated gene expression results 3.7, there are several genes like *sakA*, *mpkA* or *roh1*, with small oscillations in their simulated expression. The ripples fade over the course of the expression, and it is difficult to pinpoint the cause of these phenomenas, since the effected genes form regulatory circles. These oscillations are most likely caused by the attempt of the NetGenerator algorithm to fit the expression in genes where the first time point differs greatly from the rest of the measurements. To fit this data, the algorithm starts the simulation with a strong change in expression and later tries to smooth the simulation back fit the other measurements.

3.3.7 Network topology

The results of the network modeling can be seen in figure 3.8 and table 3.7. *sakA* has 16 outgoing interactions (including autoregulation), which is the highest number for all genes in the network. Except for the autoregulation, all outgoing connections are supported by the prior knowledge and none failed the robustness test. Only one incoming connection did not pass the robustness test. It is the prior knowledge supported activation by *ssk2*. Indirectly, this activation is still included in the model, since *ssk2* represses *ptc2*, which is an inhibitor of *sakA*. *sakA* then inhibits *ssk1*, which is the transcription factor for *ssk2*, creating an indirect autoregulation.

Since the activation by *ssk2* failed the robustness test, only two incoming interactions reach *sakA*, one by *mpkA* and one by *ptc2*. Since the direct autoregulation of *sakA* is also repressing, this leaves the gene with only inhibiting interactions, which is not sufficient to describe the simulation of expression as seen in figure 3.7. There, little ripples can be seen in the simulated expression. This may indicate a slight over fitting, as these ripples occur between two time points, were no data can support such delicate modeling.

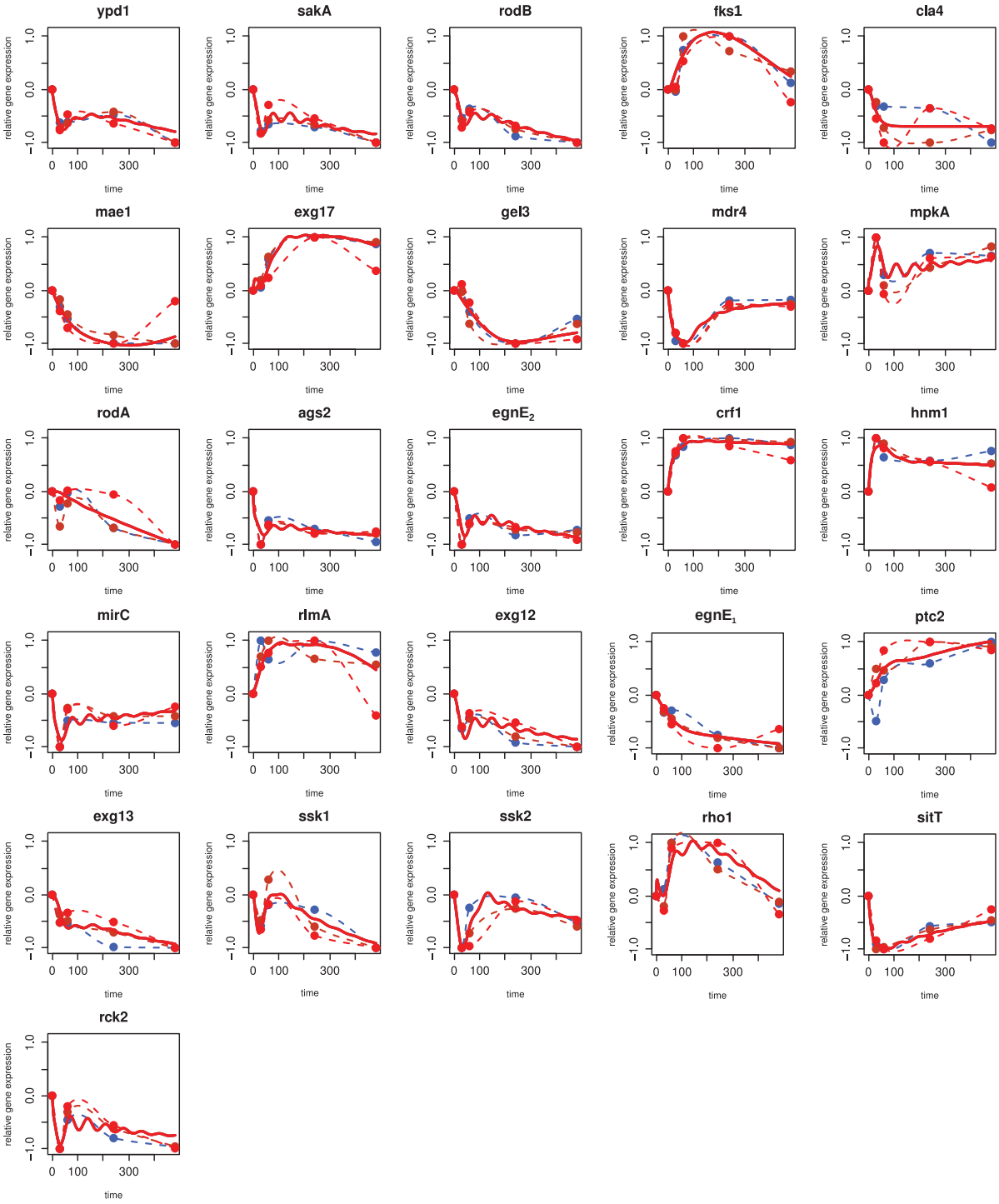

Figure 3.7: **Simulated gene expression under Caspofungin stress.** The three dashed lines (brown, red, blue) represent the interpolated measurements between the actual measured data at the time points (dots). The red solid line shows the simulated expression of the genes.

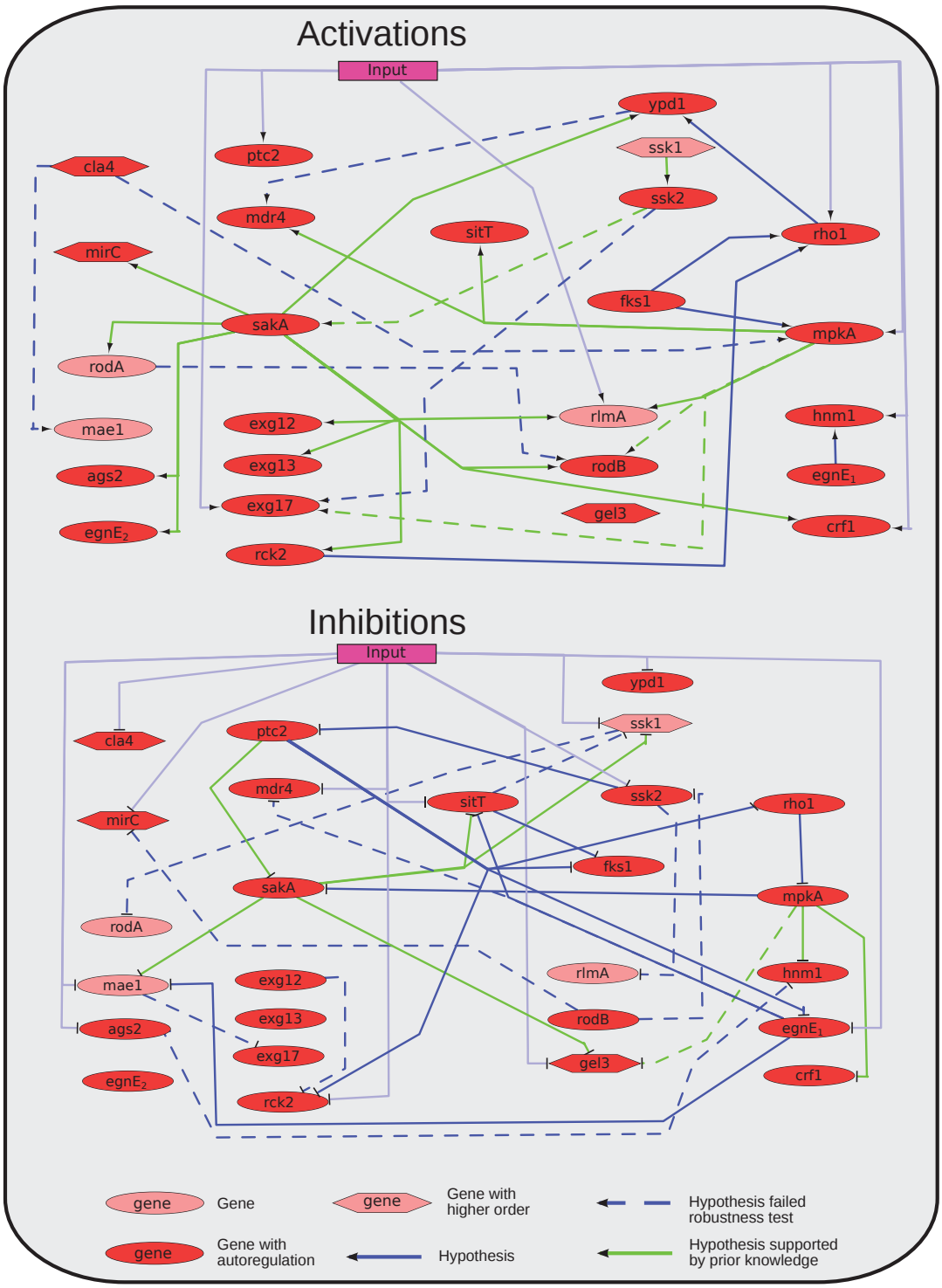

Figure 3.8: **Simulated network for Caspofungin stress.** The interactions were split into activating and inhibiting connections.

mpkA is the gene with the second highest number of outgoing regulations: six, excluding the autoregulation. Five of them are supported by prior knowledge. Before the robustness test, *mpkA* had three more outgoing interactions, all based on prior knowledge. The only regulation by *mpkA*, which is not part of the prior knowledge is a repression of *sakA*. This so far unreported cross-talk between two major response regulators is an interesting result of the inference.

mpkA is regulated by *fks1*, the drug target of Caspofungin in a rather contradicting manner. First, there is a direct activation, but also an indirect inhibition, since *fks1* also activates *rho1*, for which the model predicts a repressing effect on *mpkA*. This could hint on some fine tuning of the expression of *mpkA*. *fks1* itself is not connected to the input, despite the fact that the input represents Caspofungin treatment.

ptc2 is a phosphatase and alterations of phosphorylation states are a well recognised way of activity regulation in genes. It has inhibiting connections to *sakA* and *fks1*, to important regulators in the network, as well as $egnE_1$, a mutanase. Mutanases are sugar-degrading proteins. Since the cell wall of *A. fumigatus* contains a lot of sugar, its expression must be fine tuned in order to react to the cell wall damage.

The Input, which mainly represents the Caspofungin treatment, inhibits all four transporters: *mirC, mdr4, mae1* and *sitT*. This hinders *A. fumigatus*'s ability to decrease the osmotic stress and part of the antifungal effect of Caspofungin.

From the seven literature based prior knowledge, which I regarded as quite confident, four are part of the final model. The activations *mirC* $\rightarrow$ *cla3* and *ypd1* $\rightarrow$ *ssk1* were not considered by the inference, while *ssk2* $\rightarrow$ *sakA* failed the robustness test. The interaction *mpkA* $\rightarrow$ *rlmA*, which is supported by both prior knowledge sources, was implemented in the network and is robust to noise and perturbation.

Table 3.7: **Interactions of the simulated network.** Shows the source and target of every interaction as and if the regulation is activating ($\rightarrow$) or repressing ($\dashv$). If the interaction is supported by the prior knowledge can be seen in the column labeled "prior". The column "robust" shows if the interaction passed the robustness test.

source	type	target	prior	robust	source	type	target	prior	robust
ags2	$\dashv$	*ags2*	no	**yes**	*mpkA*	$\rightarrow$	*rlmA*	yes	**yes**
ags2	$\dashv$	*hnm1*	no	no	*mpkA*	$\rightarrow$	*mdr4*	yes	**yes**
cla4	$\dashv$	*cla4*	no	**yes**	*mpkA*	$\rightarrow$	*sitT*	yes	**yes**
cla4	$\rightarrow$	*mpkA*	no	no	*mpkA*	$\dashv$	*gel3*	yes	no
cla4	$\rightarrow$	*mae1*	no	no	*mpkA*	$\rightarrow$	*rodB*	yes	no
crf1	$\dashv$	*crf1*	no	**yes**	*mpkA*	$\rightarrow$	*exg17*	yes	no
$egnE_1$	$\dashv$	*mdr4*	no	**yes**	*ptc2*	$\dashv$	*sakA*	yes	**yes**
$egnE_1$	$\dashv$	*sitT*	no	**yes**	*ptc2*	$\dashv$	*fks1*	no	**yes**
$egnE_1$	$\dashv$	$egnE_1$	no	**yes**	*ptc2*	$\dashv$	$egnE_1$	no	**yes**
$egnE_1$	$\dashv$	*mae1*	no	**yes**	*ptc2*	$\dashv$	*ptc2*	no	**yes**
$egnE_1$	$\rightarrow$	*hnm1*	no	**yes**	*rck2*	$\dashv$	*rck2*	no	**yes**

Continued on next page

Table 3.7 – continued from previous page

source	type	target	prior	robust	source	type	target	prior	robust
$egnE_2$	⊣	$egnE_2$	no	**yes**	*rck2*	→	*rho1*	no	**yes**
exg12	⊣	*exg12*	no	**yes**	*rho1*	⊣	*mpkA*	no	**yes**
exg12	⊣	*rck2*	no	no	*rho1*	⊣	*rho1*	no	**yes**
exg13	⊣	*exg13*	no	**yes**	*rho1*	→	*ypd1*	no	**yes**
exg17	⊣	*exg17*	no	**yes**	*rodA*	→	*rodB*	no	no
fks1	⊣	*fks1*	no	**yes**	*rodB*	⊣	*rodB*	no	**yes**
fks1	→	*mpkA*	no	**yes**	*rodB*	⊣	*mirC*	no	no
fks1	→	*rho1*	no	**yes**	*rodB*	⊣	*ssk2*	no	no
gel3	⊣	*gel3*	no	**yes**	*sakA*	⊣	*sakA*	no	**yes**
hnm1	⊣	*hnm1*	no	**yes**	*sakA*	⊣	*sitT*	yes	**yes**
Input	⊣	*ags2*	no	**yes**	*sakA*	⊣	*ssk1*	yes	**yes**
Input	⊣	*sitT*	no	**yes**	*sakA*	⊣	*mae1*	yes	**yes**
Input	⊣	*mirC*	no	**yes**	*sakA*	⊣	*gel3*	yes	**yes**
Input	⊣	*mdr4*	no	**yes**	*sakA*	→	*rck2*	yes	**yes**
Input	⊣	*rck2*	no	**yes**	*sakA*	→	*rodB*	yes	**yes**
Input	⊣	*ypd1*	no	**yes**	*sakA*	→	$egnE_2$	yes	**yes**
Input	⊣	*ssk2*	no	**yes**	*sakA*	→	*mirC*	yes	**yes**
Input	⊣	*ssk1*	no	**yes**	*sakA*	→	*exg12*	yes	**yes**
Input	⊣	*mae1*	no	**yes**	*sakA*	→	*ags2*	yes	**yes**
Input	⊣	$egnE_1$	no	**yes**	*sakA*	→	*rlmA*	yes	**yes**
Input	⊣	*cla4*	no	**yes**	*sakA*	→	*exg13*	yes	**yes**
Input	⊣	*gel3*	no	**yes**	*sakA*	→	*ypd1*	yes	**yes**
Input	→	*mpkA*	no	**yes**	*sakA*	→	*crf1*	yes	**yes**
Input	→	*rho1*	no	**yes**	*sakA*	→	*rodA*	yes	**yes**
Input	→	*exg17*	no	**yes**	*sitT*	⊣	*sitT*	no	**yes**
Input	→	*crf1*	no	**yes**	*sitT*	⊣	*fks1*	no	**yes**
Input	→	*hnm1*	no	**yes**	*sitT*	⊣	*ssk1*	no	no
Input	→	*rlmA*	no	**yes**	*ssk1*	→	*ssk2*	yes	**yes**
Input	→	*ptc2*	no	**yes**	*ssk1*	⊣	*ssk1*	no	no
mae1	⊣	*exg17*	no	no	*ssk1*	⊣	*rodA*	no	no
mdr4	⊣	*mdr4*	no	**yes**	*ssk2*	⊣	*ssk2*	no	**yes**
mirC	⊣	*mirC*	no	**yes**	*ssk2*	⊣	*ptc2*	no	**yes**
mpkA	⊣	*mpkA*	no	**yes**	*ssk2*	⊣	*rlmA*	no	no
mpkA	⊣	*sakA*	no	**yes**	*ssk2*	→	*exg17*	no	no
mpkA	⊣	*hnm1*	yes	**yes**	*ssk2*	→	*sakA*	yes	no
mpkA	⊣	*crf1*	yes	**yes**	*ypd1*	⊣	*ypd1*	no	**yes**
ypd1	→	*mdr4*	no	no					

3.3.7.1 Simulation without prior knowledge

To investigate closer, how many influence the prior knowledge has on the edge selection process, I also studied the network that was not given any advantage by the prior knowledge. I was especially interested, which prior knowledge supported interactions were still selected. I simulated the network for the same parameters as for the final model, except I did not include any prior knowledge. I also did not perform a robustness analysis.

The network inferred without prior knowledge (*NoPK*) consists of 100 interactions, including autoregulations, and has a model error of 1.6, similar to the best fitted network (See table 3.6 for details). This model only identified three interactions that are also part of the prior knowledge, which are $sakA \rightarrow exg12$, $sakA \rightarrow rodB$ and $sakA \dashv ssk1$. All of them are also part of the network inferred with prior knowledge (table 3.7).

Another significant difference between the model with prior knowledge and NoPK is *mpkA* and *sakA* are not hubs without prior knowledge. In NoPK, both *mpkA* and *sakA* have degrees of eight, while the average degree in the network is 6.8. The maximum degree has *sitT* with 10 and the minimal degree is five, which was reached by several genes. The inhibition of *sakA via mpkA* is also not part of the NoPK model. Both models shared 50 interactions, mostly autoregulations and interactions starting from the Input.

3.3.8 Biological validation

Mathematical model prediction is a valuable tool in the investigation of gene regulatory networks. However, despite careful modeling and robustness testing, final validation (or falsification) of an interaction can only be done in the laboratory. Because of this, selected interactions in the network model were tested with RT-PCR and western blot. The tests were done by my colleagues Dr. Vito Valiante and Clara Baldin.

3.3.8.1 Western blot

The western blot measures the phosphorylation of proteins, which are interpreted as gene expression levels. The phosphorylation levels of MpkA and SakA were taken in the strain of Δ*akuB*, Δ*ptc2*, Δ*mpkA* and Δ*sakA* knock-outs. We saw that the deletion of *akuB* does not globally alter the gene expression (figure 3.4). This way, the western blot of the Δ*akuB* mutant can be used as comparison for the other mutants. The results can be seen in figure 3.9. It clearly shows an increase in activation of SakA in the Δ*mpkA* mutant, hinting that there is indeed an inhibition of *sakA* by *mpkA*, validating our hypothesis.

The phosphorylation of SakA is decreased in the Δ*ptc2* mutant. This contradicts the model prediction, where *ptc2* inhibits the expression of *sakA*.

Looking at the results of the MpkA measurements, it is interesting to see a decrease of phosphorylation 2 h after treatment. This effect is especially prominent in the Δ*sakA* mutant. It can be the result of an autoregulatory mechanism to stop the Caspofungin

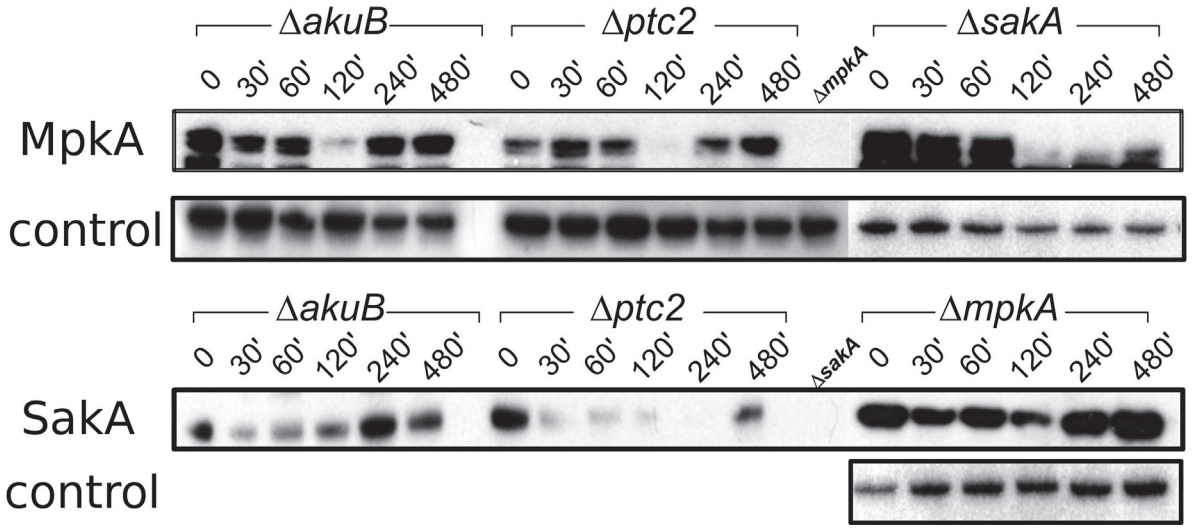

Figure 3.9: **Results of the western blotting.** Phosphorylation levels were measured for MpkA and SakA in the Δ*akuB*, Δ*ptc2*, Δ*sakA* and Δ*mpkA* strains. The control is the untreated sample of *A. fumigatus*.

response, after the adaption process is completed. The phosphorylation level in the Δ*ptc2* do not differ from the ones in the Δ*akuB* mutants. According to the model, *ptc2* inhibits *fks1*, an activator of *mpkA*, so a decrease of activation would be expected in the Δ*ptc2* mutant. However, this effect is shown in figure 3.9 only slightly.

3.3.8.2 qRT-PCR

To further investigate the predicted interactions in the network, a qRT-PCR study was conducted. The results can be seen figure 3.10. As the western blots, the PCR also show a distinct up-regulation of *sakA* in the Δ*mpkA* mutant. This up-regulation can only be seen 4 h after treatment, not in the 0 h measurement, indicating that it is indeed a response to the Caspofungin stress. This validates the prediction of the network. Rispail *et al.* [95] states and slao the model confirms that *ptc2* inhibits *sakA* under Caspofungin stress. The qRT-PCR shows that in the Δ*ptc2* mutant, *sakA* is up-regulated before treatment, but then down-regulated after the treatment. This could also be a delayed down-regulation by *mpkA*, which is still expressed in this mutant.

A similar pattern can be seen for *mpkA*. It is up-regulated in the Δ*ptc2* mutant at time point 0 h, but down-regulated at time point 4 h. Since the model predicted that *ptc2* inhibits *fskA*, the activator of *mpkA*, the initial up-regulation supports the hypothesis. The down-regulation later on should require further study.

According to the model, *rodB* is activated by *mpkA* and *sakA*, with the interaction from *mpkA* failing the robustness test. In the qRT-PCR, *rodB* is strongly down-regulated in both mutants, Δ*sakA* and Δ*mpkA*. This is in accordance with the model, even if the activation by *mpkA* was not part of the final model.

In the model, *rlmA* is activated by the Input, *mpkA* and *sakA*, but up-regulated in

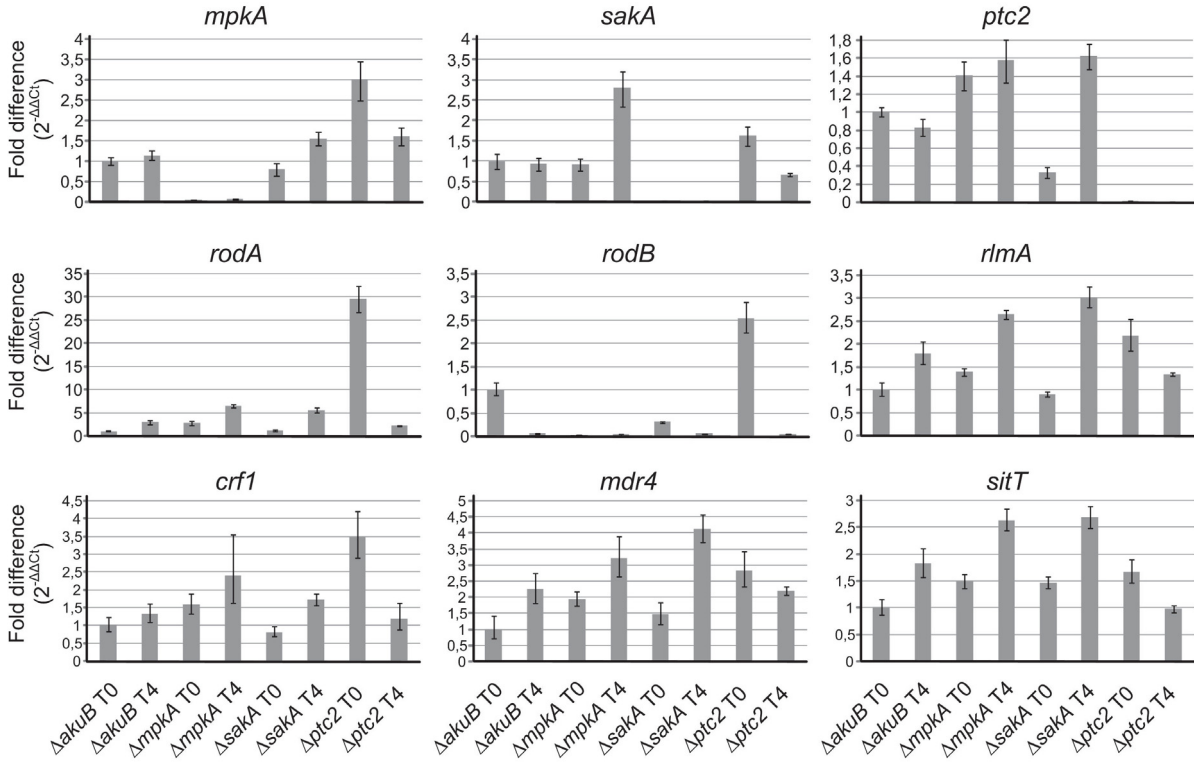

Figure 3.10: **Results of the qRT-PCR study.** RNA samples were taken before Caspofungin treatment and 4 h after in Δ*akuB*, Δ*mpkA*, Δ*sakA* and Δ*ptc2* strains. The protein expression at time point 0 h of the Δ*akuB* mutant is the control, according to which all other expressions are quantified.

the PCR of the knock-outs. Looking at the simulated values in figure 3.7, where *rlmA* is activated early on, it is reasonable to assume a dominant influence of Caspofungin on the expression. Especially since the activation of *rlmA* by *mpkA* and *sakA* are both supported by literature.

crf1 is up-regulated by *sakA* and down-regulated by *mpkA*, as far as the model is concerned. In the qRT-PCR, *crf1* is indeed up-regulated at time point 0 h and 4 h in the Δ*mpkA* strain, supporting the network. In the Δ*sakA* strain, it is down-regulated at first, but up-regulated 4 h after treatment. The network predicts an up-regulation by Caspofungin, so this may be seen as a confirmation of the model.

The model predicts an inhibition of *mdr4* by Caspofungin, but in the qRT-PCR, a up-regulation after the treatment is visible in the Δ*ptc2* mutant. Looking at the simulation in figure 3.7, it is visible, that the down-regulation happens within the first hour, and then the expression increases again. It is possible, that the PCR measurements misses this early reaction and captures the up-regulation later on. While the model also predicts an activation by *mpkA*, the PCR shows an up-regulation in the *mpkA* knock-out mutant. This contradicts my prediction.

sitT is supposed to be activated by *mpkA* and repressed by *sakA*. In the PCR, *sitT* is

up-regulated in both knock-out mutants. This leaves an ambiguous picture, from which no specific conclusions can be drawn.

We found hints for the correct responses to some of our predictions, especially the inhibition of *sakA* by *mpkA*, while some prediction could not be verified and some results were inconclusive. It is important not to forget, that in the predicted model, genes are often regulated by several other genes, while the qRT-PCR was done with single knock-out mutants. This means that the missing stimulus from one knocked-out genes may be compensated by another, without giving us an opportunity to study this.

3.3.8.3 Membrane permeability

The network model suggests, that Caspofungin influences the transport proteins in the cell wall, *mirC, mdr4, mae1* and *sitT*. We reasoned that this hinders *A. fumigatus* do react to the osmotic stress, as the fungus is incapable to remove outside agents from the cell, without the transporters. In order to investigate the osmotic stress during the treatment, a Rhodamine-123 was used.

The results can be seen in figure 3.11. They show that after about 4 h, the Rhodamine-123 uptake in the wt is more than two times higher in the samples with a dose of 0.1 μg/ml^{-1} Caspofungin. Similar effects can be seen for the other strains, except for Δ*ptc2*. This shows that the cell wall integrity is disrupted from the drug exposure. The Δ*sakA* showed an higher absolute Rhodamine-123 uptake than the wt, and the uptake of Δ*mpkA* strain was even higher. As for Δ*ptc2*, the Rhodamine-123 uptake of the treated sample is higher during the first four hours, but then normalises at equal level with the untreated sample. The Caspofungin resistant strain EMFR^{S678P} also showed an increased uptake, even though it can not have been influenced by a Caspofungin induced inhibition of 1,3-β-glucan synthase.

The wt was also tested with higher dosages of Caspofungin, which lead to a decreased Rhodamine-123 uptake. This shows the so called paradoxic effect.

3.4 Discussion

3.4.1 Gene Selection

One of the first problems when starting a small-scale network inference is the selection of genes, that are to be investigated. But in contrast to large-scale models, the expected prediction accuracy for single interaction is much higher. In our study, we wanted to investigate the gene regulatory effect of *A. fumigatus* on Caspofungin treatment. The first reasonable step was to select all genes that are differentially expressed during the Caspofungin treatment.

The differentially expressed genes (DEGs) are determined using the fold change (often logarithmic), the p-value, which tests for the null hypothesis that there is no expression change, or both. It is difficult to observe small changes in expression with the naked eye in semi-quantative methods like western blotting or other biological experiments later on, which is why often fold changes ≥ 2 are preferred. However, I considered genes to be

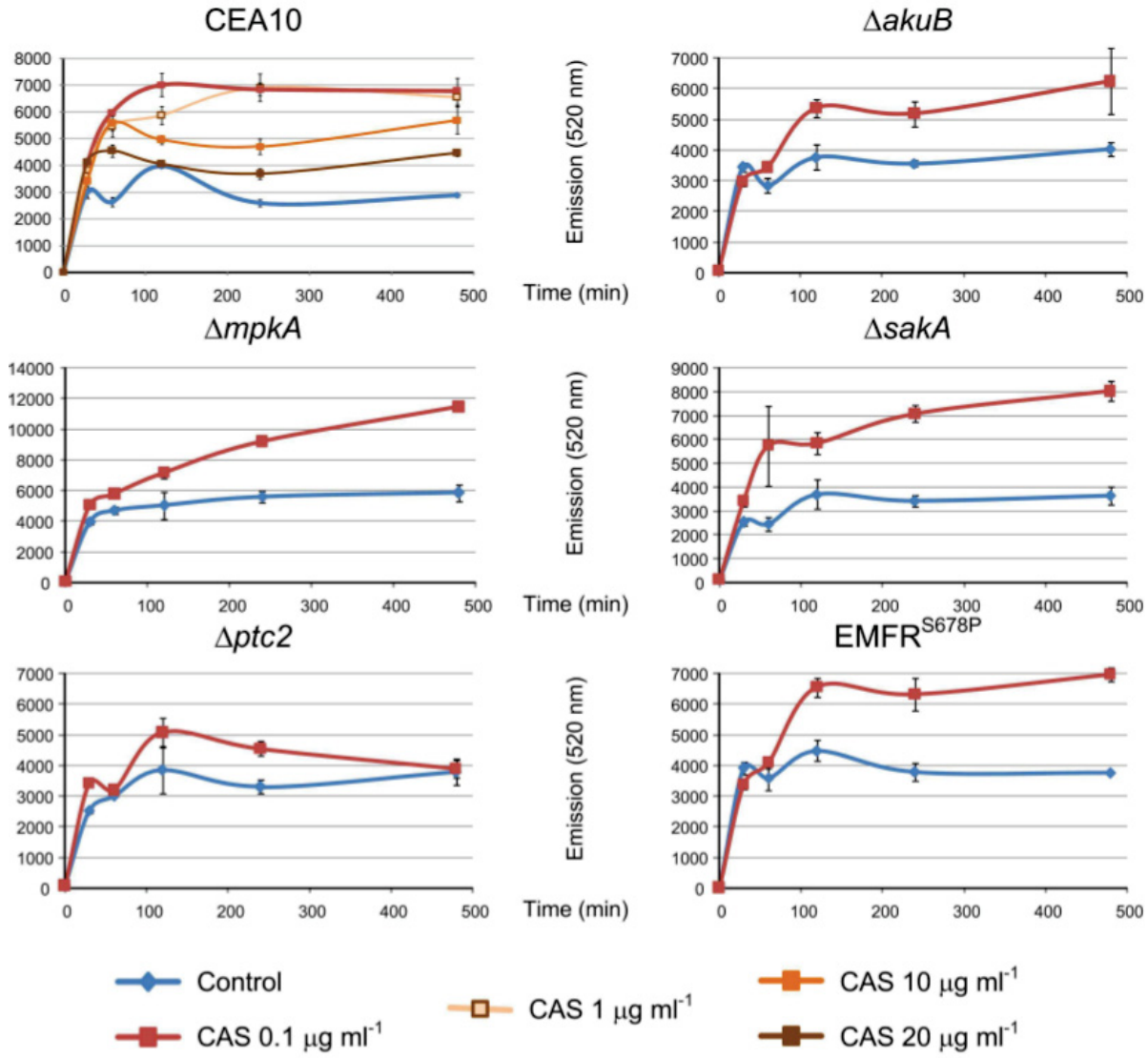

Figure 3.11: **Results of the Rhodamine study.** Measurements of the Rhodamine-123 intake during the with (red line) and without (blue line) Caspofungin treatment for different strains. The knock-out mutant were treated with a sub-lethal dose of Caspofungin (CAS; 0.1 μg ml^{-1}), while the wild type (CEA10) was treated with different dosages. The figures show that with Caspofungin, the Rhodamine-123 uptake is significantly increased after 4 h, proving that Caspofungin increases the permeability of the cell wall.

differential expressed if the FDR-adjusted p-value is ≤ 0.05. The reason for this decision comes from the good quality of the RNA-Seq data, which can be seen in table 3.2.

Most of the Pearson correlations between the replicates are above 0.96. At time point 1 h, the score is "only" 0.89, probably because *A. fumigatus* is in the middle of the Caspofungin adaption, which causes great changes in the gene regulatory network. During that phase, gene regulation is timely regulated, and within a few minutes small alterations in regulation can occur. The correlation of the samples taken after 24 h is considerably lower. The gene-regulation at this time point is mostly influenced by the beginning starvation.

When the correlation between the three replicates is very high, significant expression changes at different time points can be found, even if the absolute expression change is rather low. This was one of the reasons I omitted the 24 h time point from the analysis. The mean correlation of 0.74 and 0.76 would weaken the statistical analysis of the data, for example the calculation of the p-value to identify DEGs. It is also more difficult for the NetGenerator to find a fitting model. Additionally, I figured that the 24 h samples are largely influenced by starvation than by the Caspofungin treatment. Omitting data from the analysis is always a difficult decision, since valuable information can be lost, especially considering the general lack of data we have to face during network inferences. However, during the collection of data, one should always consider the aim of the study. In focused network studies like this one, it is of uttermost importance not to introduce biased or distorted data, as this can quickly lead to false predictions.

Of special interest is also the time point 0.5 h. Several genes showed a strong reaction, a peak or valley, at that time point, which indicates an early response mechanism. The *A. fumigatus* wild type strain is able to completely adapt to Caspofungin treatment, and these early responders are a vital part of this adaptation. Important drug stress response genes like *mpkA*, *sakA* or *ssk2* exhibit that strong deviation from the initial state. The first 30 minutes are of crucial importance when an organism tries to respond to cell surface stress, and the adoption measures vary from the "long term" responses, like modifying the glucan / chitin ratio in the cell wall. Short term responses include initialised cell wall repair mechanism and transporter activity, in order to limit the damage done to the cell.

The second step was functional categorisation of the genes I found. Since Caspofungin is an outside agent that disrupts the cell wall, I knew that investigation of cell wall related genes would be fruitful.

The number of annotated genes for non-model organisms remains low and the little annotation that can be found often originates from sequence similarity to other model species. Despite the increasing clinical importance of *A. fumigatus*, there is still little species specific information available. This makes it especially difficult to reliably identify candidate genes. Despite the very good quality of the data, the quantity limits the inference process to 20-30 genes only. This limitation comes not from the used inference algorithm, but caused by a more general issue. When the inference algorithm tries to find a regulator for a certain gene, and two candidates have the same expression pattern, the algorithm can not decide, which regulator to choose. So even after functional categorisation, the final decision on which genes to choose requires expert knowledge,

since the number of candidates was still to high.

Transcription data alone is often insufficient to make reliable predictions about genetic regulation. The amount of data is not enough to find unique solutions for the equation systems. Most systems are underdetemined, leaving only heuristic approaches. Additionally, transcription data alone gives no information about the semantic information that has been discovered over the years. Prior knowledge is the way to include literature knowledge and therefore the insight gained by other studies. It is also possible to include additional data sources and data types in the inference. At present, network inference with data from multiple sources is still cumbersome, but as the amount of available data increases and standards become more unified, more and more information can be combined.

3.4.2 NetGenerator

The NetGenerator algorithm used in this study is not only able to include prior knowledge, but also multi stimuli data. This might have been a way to handle the influence of hunger stress at the 24 h time point and include the data, by considering nutrition a depleting stimulus. But again, this is not the aim of the study and the genes considered in this network inference are not supposed to directly react to hunger stress.

A general assumption of the NetGenerator algorithm is that the investigated genes are in a steady state at the beginning of the experiment, when an external perturbation causes a change in expression. After the perturbation subsides, the genes are expected to return to the steady state. This is another reason, why I excluded the hunger stress in the simulation, as it prevents the fungus to return to "normal" conditions. While figure 3.12 shows that the relative expression of genes like *fks1* and *roh1* returns to zero after 8 h, genes like *ptc2* or *sakA* do not. However, that the expression of the genes returns to the "normal" state is preferred, but not necessary to get reliable network predictions.

The NetGenerator uses many direct autoregulations in the model. Out of 26 genes, 23 had an autoregulation, from which one (*ssk1*) is not robust. Mathematically, it prevents the simulation from running out of bounds, i.e. to increase indefinitely. From a modeling point of view, negative autoregulation is similar to RNA-denaturation. It stops the uncontrolled accumulation of gene products in a cell.

Genes like *mpkA*, *sakA* or *ssk1* show a strong reaction in the first time point after treatment, showing a early response behaviour, depicted in figure 3.7. Those heavy perturbation at the 0,5 h time point are difficult to model in ordinary differential equations. On my first attempt, I tried modeling the data using only first-order equations. The result was, that some times, the strong response in the first hour threw the model "of balance", e.g. causing huge oscillations, as the model tried to implement the other time points. These oscillations can still be seen in figure 3.12, but not as dominant.

3.4.3 Robustness

The robustness tests showed, that 18 of the inferred interactions are not robust. From the seven interactions, that are susceptible to missing prior knowledge, none is actually part of any prior knowledge. This can happen, when a prior knowledge supported interactions is not inferred due to omitted prior knowledge. This can lead to alternative connectivity in other parts of the network and even cause interactions that are not supported by prior knowledge to disappear.

Modeled networks are often depicted with robust interactions only. I think this representation is misleading, since these edges are still necessary to explain the expression of the gene in the simulation. The information, that the interaction is rather fragile is nevertheless important in order to evaluate the reliability of the network. Especially when selecting candidates for biological validation, one should refrain from choosing interactions, which are not robust.

An alternative way of depicting the results of robustness tests is showing the alternative interactions, which came up during the test. During every cycle of the test, the input data is distorted, and a network is inferred, that may has a different topology than the original network, lacking certain interactions. These interactions are usually substituted by alternative regulation. It may be interesting to count these alternative interactions and show them in the network, revealing alternative pathways. On the other hand, finding a good visualisation gene regulatory networks including the confidence and alternatives of the infered edges is complicated even for small networks. Showing all alternative pathways would increase the complexity of the figure and the evaluation. One might also argue, that the alternative networks are done with intentionally distorted data, and should therefore not be considered in the first place.

3.4.4 Cross-talk of *mpkA* & *sakA* pathways

A goal of this study was the investigation of cross-talk between the high osmolarity glycerol (HOG) pathway, with *sakA* as a key regulator and the cell wall integrity (CWI) pathway, with *mpkA* as a key regulator. By investigating the transcriptional activity of the genes using the NetGenerator, I found a repressive influence of *mpkA* on *sakA* as shown in figures 3.8 and 3.12. This hypothesis was confirmed by western blot, seen in figure 3.9, where the phosphorylation level of SakA is generally increased in the Δ*mpkA* mutant. Also, the expression of *mpkA* decreases 2 h after the Caspofungin treatment. This decrease is especially dominant in the Δ*sakA* mutant, probably because *sakA* is needed as a target of the CWI pathway. The decrease may also be the end of the stress response of *A. fumigatus*, as it needs to return to normal gene expression levels, after the drug adaption is completed.

However, here, the western blot contradicts the findings of the PCR (figure 3.10). There, the expression of MpkA is increased 4 h after treatment. A possible explanation could be, that *A. fumigatus* expresses *mpkA*, but, as the target protein is missing, does not phosphorelate it, rendering it inactive. The qRT-PCR also shows that the expression of *sakA* is up to three times higher in the Δ*mpkA* mutant, but only after the Caspofungin

treatment, confirming the hypothesis of the model. The western blot (figure 3.9) shows that the phosphorylation of SakA is on equal levels at 0 h and 4 h after treatment, and increased in the next time point at 8 h. This shows that the connection between expression and phosphorylation levels is not always linear and can sometimes lead to contradicting results. This can partly be explained by the seperation of expression and activation (*via* phosphorylation) of a protein.

The model also predicts feed-back loops of *mpkA* and *sakA* as seen in figure 3.12. The indirect regulation loop of *sakA* is even confirmed by prior knowledge: $sakA \dashv ssk1 \rightarrow ssk2 \rightarrow sakA$, although the connection $ssk2 \rightarrow sakA$ is not robust and alternatively replaced *via* an inhibition of and by *ptc2*. The other feed-back loop is $mpkA \rightarrow sitT \dashv fsk1 \rightarrow mpkA$. Alternatively, *fks1* can also activate *roh1*, which then inhibits *mpkA*. From the data at hand, I can not tell, under which conditions *fks1* acts activating or inhibiting. Here, only the connection of *mpkA* and *sitT* is predicted by the prior knowledge. The prior knowledge also supports the inhibition of *sitT* by *sakA*, which is another way into the feed-back loop of *mpkA*. Since *mpkA* also inhibits *sakA*, it is also a feed-back loop mof *sakA* itself. And there is one big feed-back loop of all genes except *sitT* in figure 3.12.

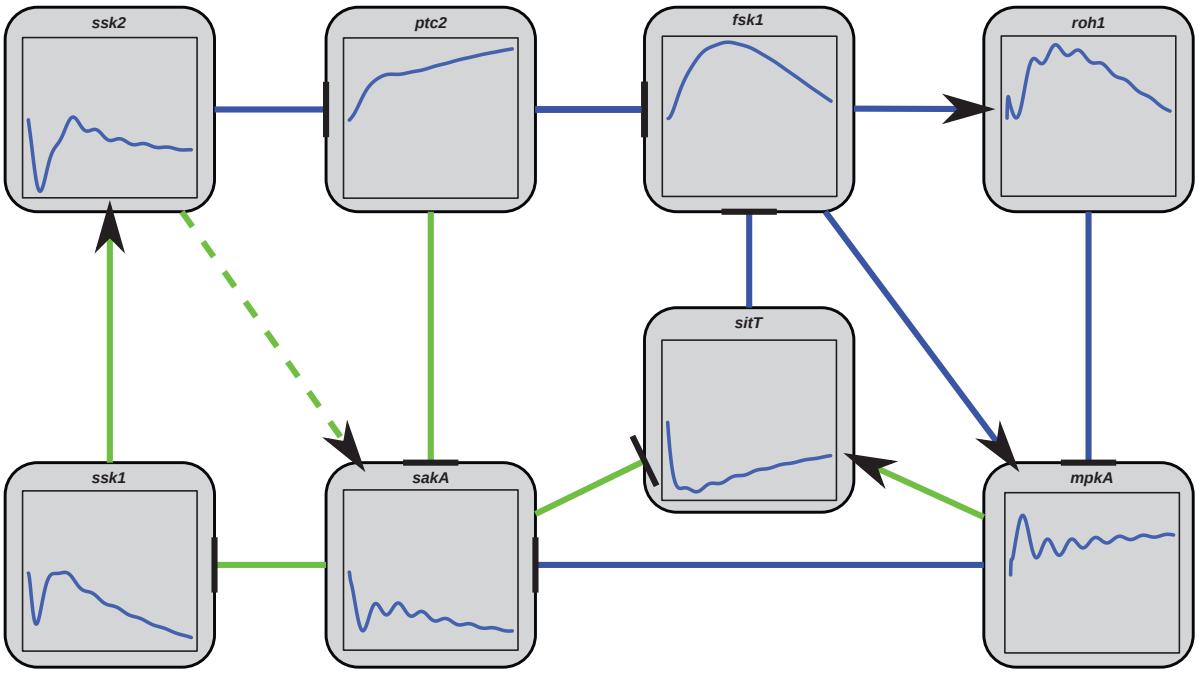

Figure 3.12: **Focused view on the regulatory center of the network.**

If the feed-back loop is mediated by post-transcriptional phosphorylation, *ptc2* is a suitable candidate as regulator, as it was the only putative phosphatase to be found differentially expressed. Prior knowledge already suggested that *ptc2* may be responsible for the repression of *sakA*, albeit in another species. The model also predicts this repression, but only if this prior knowledge is provided. The western blot on the other hand did not reflect this hypothesis, as *sakA* is not up-regulated in the $\Delta ptc2$ mutant.

The qRT-PCR shows an altered protein concentration of *sakA* and *mpkA* in the Δ*ptc2* mutant, but as an up-regulation. These results show that *ptc2* may not be the way *mpkA* regulates *sakA*, but that there is an additional, hidden regulator.

This also demonstrates the limitations of the western blot and qRT-PCR, as we only have data from single-knock-out mutants, while the model mostly predicts regulation by multiple sources. If only one of the potential regulators is knocked-out and no change in regulation is detected, it can be because the other regulator compensates for the missing one. Multi-knock-out mutants would be necessary to detect this, but this is extremely costly.

As mentioned before, *A. fumigatus* is able to adjust the ratio of chitin and glucan in its cell wall. By increasing the chitin content of the cell wall, it adopts to the Caspofungin treatment, since the drug targets the glucan synthesis. The Rhodamine-123 experiment of figure 3.11 shows that the Caspofungin resistant strain also shows an increased Rhodamine-123 uptake, showing that Caspofungin induces cell wall disruption even without an effect of *fks1*. So Caspofungin attacks *via* two way: once by inducing cell wall stress, and second by targeting *fks1*, preventing the glucan synthase and reconstruction of the cell wall. *sakA* is a key regulator of the HOG pathway and activated when osmotic stress occurs.

As shown in [32,81], the mitogen-activated kinase is responsible for the chitin balance in *C. albicans*, and most likely also conserved in *A. fumigatus*. After 2 h, *A. fumigatus* also represses the expression of *sakA via mpkA*, and silences the HOG pathway. The increase in chitin in the cell wall, confers the Caspofungin resistance.

A. fumigatus shows a paradoxic effect, which is an increased resistance at a higher drug dose. The cross-talk between *mpkA* and *sakA* can be used to explain this effect, since the treatment with Caspofungin activates *mpkA*, which then inhibits *sakA*. The higher the Caspofungin dose, the stronger the repression of *sakA*, and the more chitin is introduced to the cell wall, countering the drug effect. At the dose used in this study, the paradoxic effect did not occur. To further investigate this effect comprehensively, multiple RNA-Seq studies at different Caspofungin concentrations are necessary.

3.4.5 New prior knowledge

The importance of the prior knowledge was again emphasised by the inference done without prior knowledge (see end of chapter 3.3.7.1). *mpkA* and *sakA* were not identified as hubs and the cross-talk was not found. Transcription data is an important source of information, but still only a small piece of the big picture. Additional data from the outside is indispensable to get closer to the big picture.

The investigation of the cross-talk does not end with this study. The verified knowledge gained in these simulations can be used as prior knowledge in future studies. This includes the inhibition of *sakA* by *mpkA*, which is predicted by the simulation and validated by western blot and qRT-PCR, making it a very reliable piece of information.

There are also several rather ambiguous results of the biological validation, like the predicted activation of *rodB* by *mpkA* and *sakA* or the regulation of *crf1* by *mpkA* and

sakA. This leads to rather unreliable prior knowledge and if this information is to be used in future studies, it should be done with a small weighting only.

There is knowledge that is hard to acquire and very valuable: negative data. According biological validations (or rather falsifications), the inhibition of *sakA* and *fks1* by *ptc2* could not be confirmed, at least not directly. The same is true for the activation of *rlmA* by *sakA* and *mpkA*. This information can be included as negative prior knowledge, making these interactions less likely to occur in the network, so the algorithm can search for alternative regulations. It is unfortunate, that negative data is so hard to come by, even though it is most certainly produced in large quantities.

3.4.6 Workflow of the RNA-Seq study

The workflow presented here is depicted in figure 3.13. It can serve as a template for RNA-Seq-based small scale network inference. The raw data is mapped and the genes are selected *via* differential expression, functional categorisation and expert knowledge. One part of the prior knowledge is gained from the same mapped data and an additional part is harvested from literature information. The transcription data of the selected candidate genes is pre-processed, before the mathematical modeling begins and different networks are created. The final model is selected and its interactions are assessed for robustness and new hypotheses are crested. These new hypotheses are tested in the laboratory, where they are verified or falsified. In the end, new knowledge is gained, which can be used as prior knowledge for future investigations.

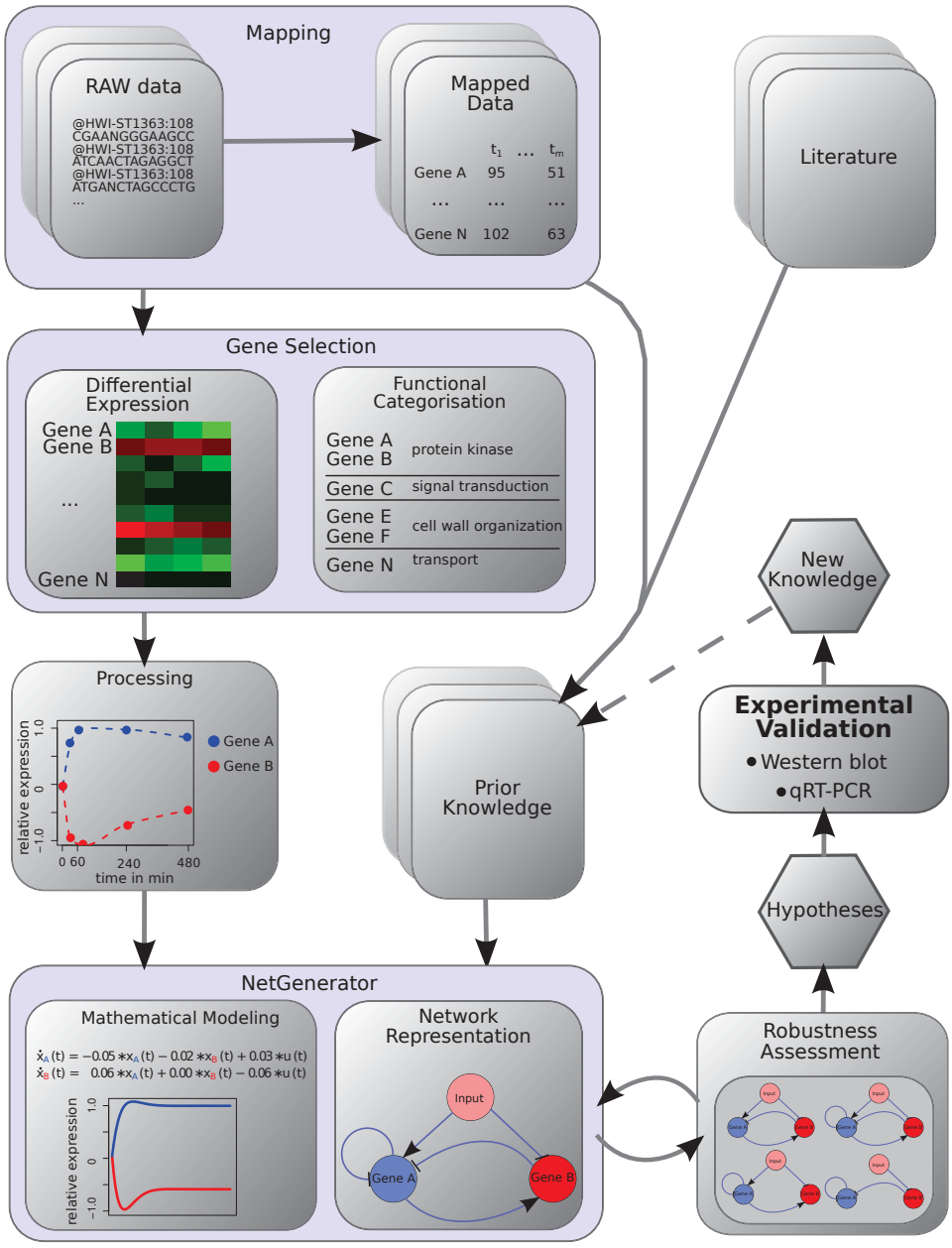

Figure 3.13: **Workflow of this study.** After the RAW data was mapped to the *A. fumigatus* genome, the model genes were selected according to their differential expression and functional categories. The data was processed and fed to the NetGenerator in combination with the prior knowledge. The algorithm models the expression values. The final network is test for robustness and new hypothesis are made. These hypothesis are tested in the laboratory, and new knowledge is gained.

4 Comparison with the adaptive LASSO

In order to compare the NetGenerator and the adaptive LASSO algorithm, I also created for the dataset and experimental setup of chapter 3 a network using the methods described in the chapters 1.3.1 and 2.2.2. Again, I first estimated the best values for λ, which represents the influence of the prior knowledge and c, which determines the network size. I performed a grid search over both parameters, checking 13 different values ranging from 0.001 to 1. This lead to 169 different parameter configuration, which took about 5 seconds to calculate on a standard workplace laptop.

Results show that giving more influence on the prior knowledge, by decreasing the value of λ, the more prior knowledge became part of the network. Since the implementation of prior knowledge is a main goal of this study, I selected $\lambda = 0.001$. The estimation of c, which is a factor for the results of the ridge regression was more difficult. In contrast to the study on Candida of chapter 2, I did not have a gold standard to compare my results too. In order to get a comparison between the results of the final network and the NetGenerator approach, I selected the network that is the closest in size. The final network of the NetGenerator had 53 gene-to-gene interactions, excluding autoregulations and Input-to-gene interactions.

Figure 4.1 shows the result of the network inference using different values of c. The smallest network at $c = 0.001$ has 26 gene-to-gene interactions, as do the networks with value for c of 0.01 and 0.5. This is the smallest number of connections, the LASSO *via* ridge regression can return for a network with 26 genes, since it has to find at least one regulator for each gene, otherwise, the network reconstruction subroutine will throw an error and the algorithm stops the calculation. Out of the 26 interactions, 21 were supported by prior knowledge, but this may be misleading. The implementation of the algorithm does not allow to differentiate between activating or repressing prior knowledge. It only considers the information about source and target of an interaction.

The network inferred with $c = 1$ consists of 52 interactions, and is therefore very close in size to network inferred by the NetGenerator, promising a good comparison. While it shares 29 interactions with the prior knowledge, 12 of them have a different sign, meaning they are repressing when the prior knowledge suggests activation and *vice versa*. With the help of the prior knowledge, *sakA* and *mpkA* became hubs again. With 14 interactions for *mpkA* (12 outgoing and 2 incoming) and 16 interactions for *sakA* (14 outgoing and 2 incoming). An interesting observation in the network are small regulatory circles of two genes. One can be seen by the mutual inhibition between *exg17* and *gel3*, another between $egnE_1$ and *ptc2*.

When comparing both networks, one can see that they have 10 interactions in com-

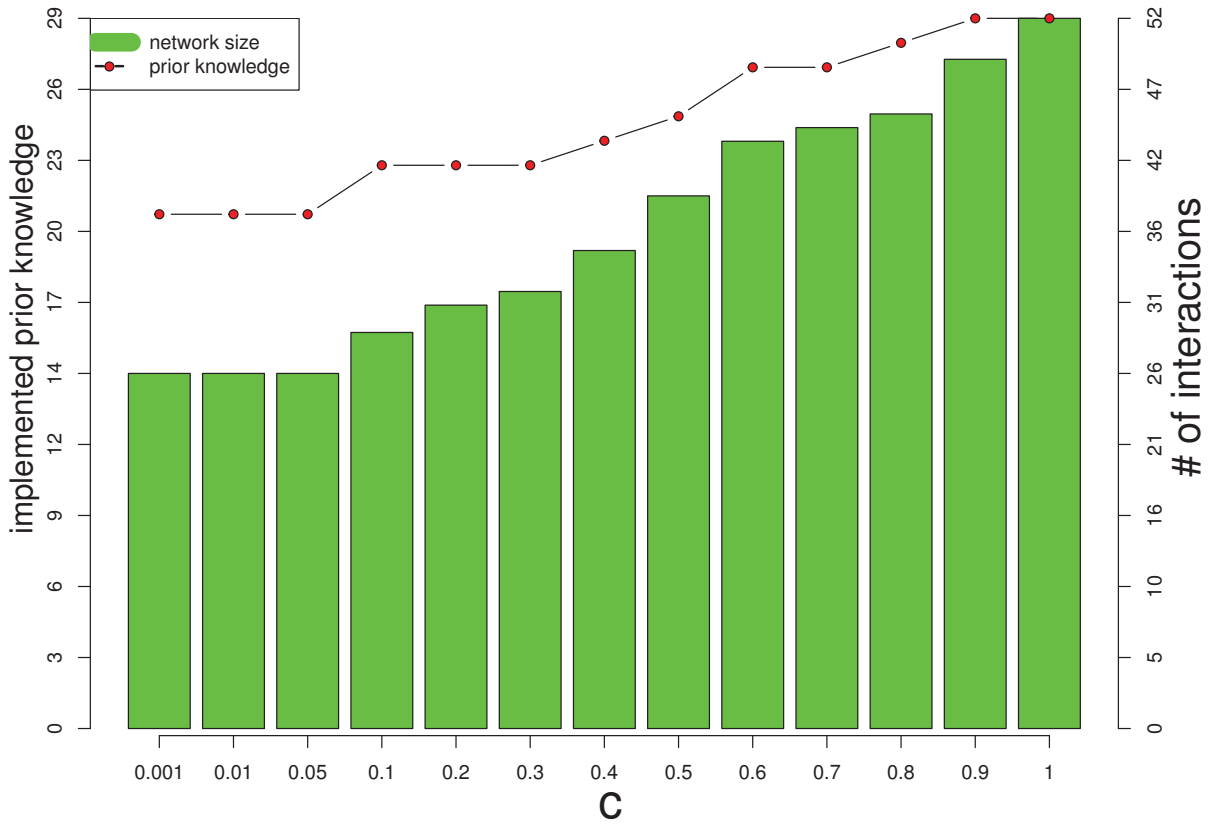

Figure 4.1: **Results of the network inference for different values of** c**.** It is important to notice, that the counting of the prior knowledge only considered the correct source and target. Not, if the interaction is repressing or activating, as the implementation of the adaptive LASSO did not allow this distinction.

mon, counting only correct signed interactions of the robust network inferred by the NetGenerator. Figure 4.3 illustrates this with connections in blue. It is little surprising that most of them are part of prior knowledge and come or go to *sakA* and *mpkA*. The activation of *roh1* by *fks1* is not supported by prior knowledge and was found in both network predictions, as was the inhibition of $egnE_1$ by *ptc2*.

In addition to the 10 interactions that are fully in accordance, there are 14 interactions, where the source and target of the interaction is correct, but the sign is wrong. And three interactions belong to NetGenerator-network but failed the robustness test: $mpkA \rightarrow exg17$, $mpkA \dashv gel3$ and $ssk2 \rightarrow sakA$. This may be a reason to reconsider deleting these edges.

The inhibition of *sakA* by *mpkA* was not found in the LASSO-based network inference.

Not being able to determine the sign of a prior knowledge interaction in the LASSO-based gives the algorithm more freedom to decide for it's own. The point of prior knowledge is, however, to give guidance by more or less verified interactions. It is still interesting, that the LASSO algorithm implemented many prior knowledge interactions

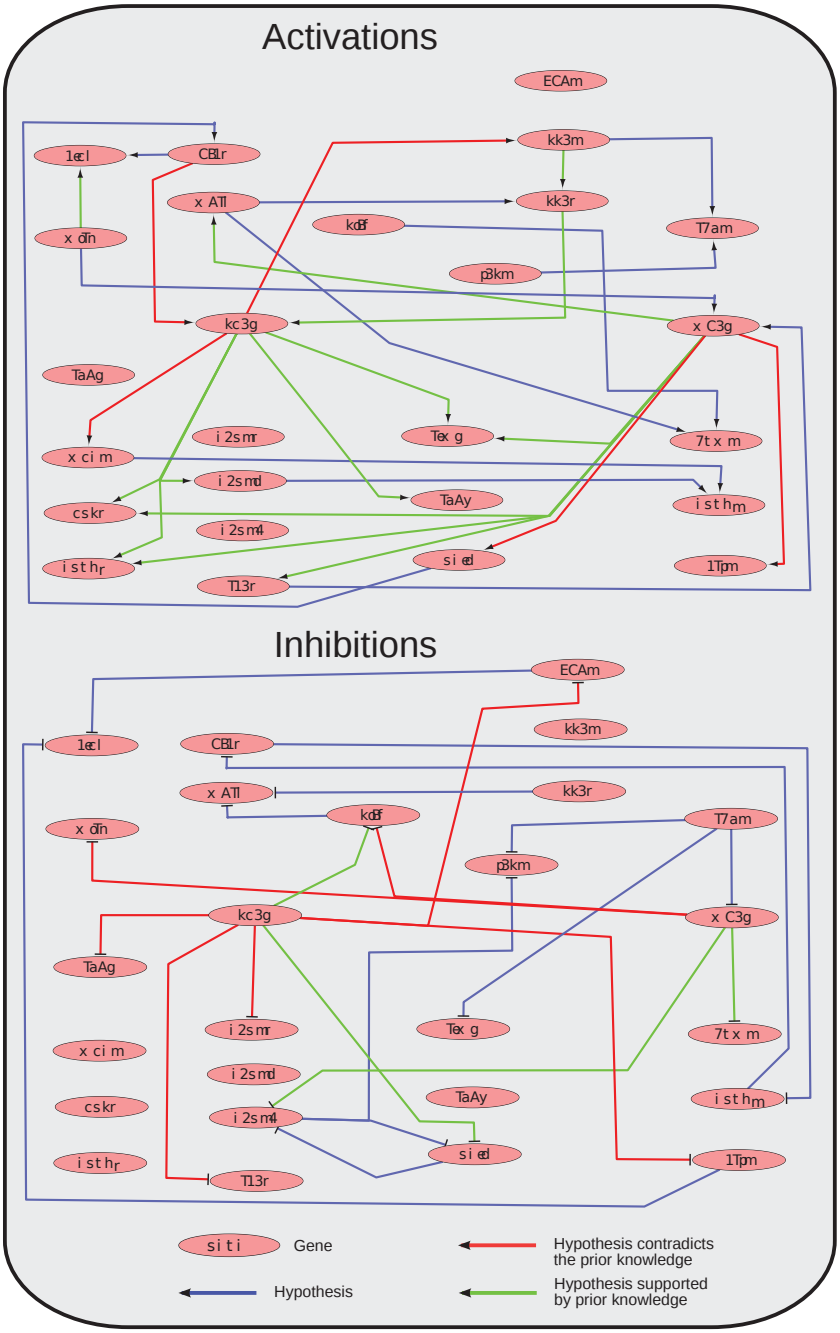

Figure 4.2: **Result of the network inference using adaptive LASSO *via* ridge regression.** The model consists of 52 interactions, from which 17 are in accordance with the prior knowledge. 12 interactions have the same source and target as the prior knowledge, but a different sign.

with the opposite sign.

The adaptive LASSO *via* ridge regression and the NetGenerator share many features, like the implementation of prior knowledge and indirectly setting upper limits to the

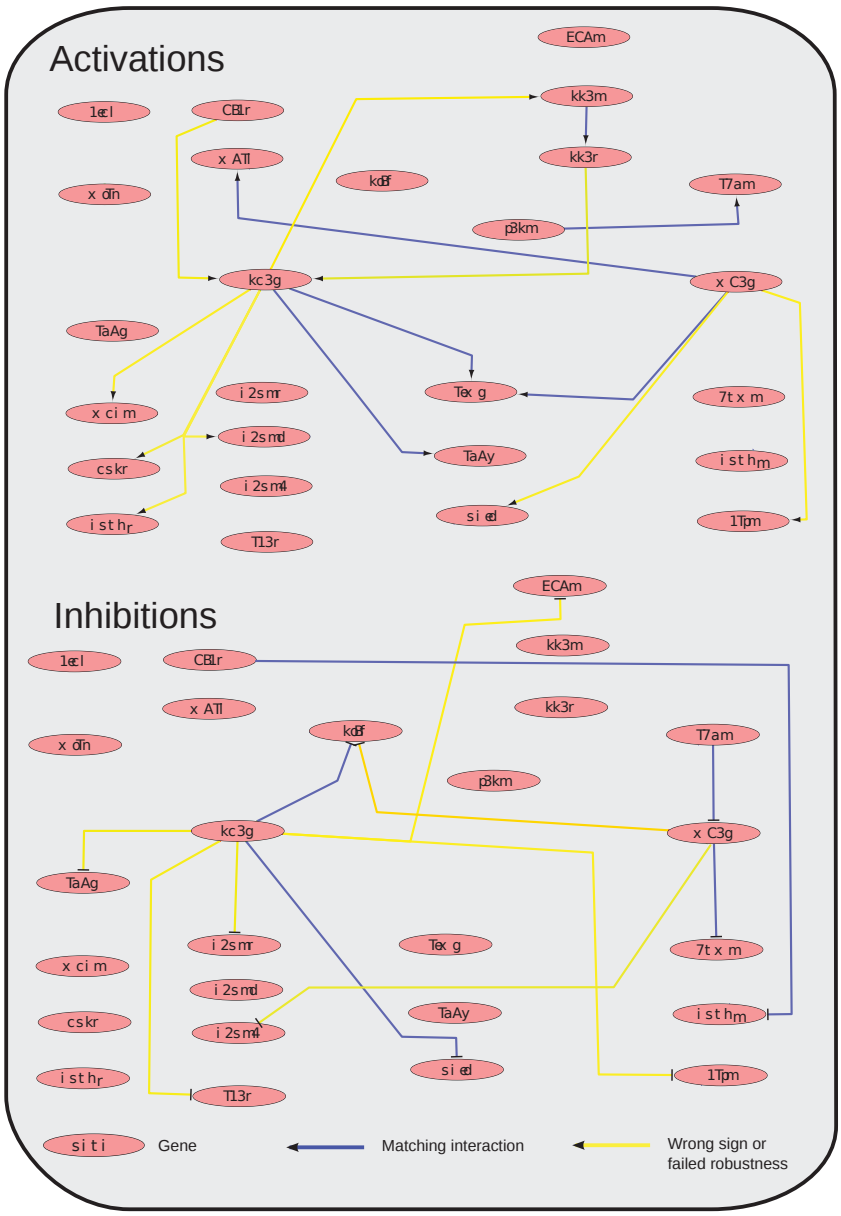

Figure 4.3: **Consensus network between the LASSO-based and NetGenerator-based approach.** 10 interactions are correct in sign and passed the robustness test during the NetGenerator-approach. 14 are in accordance by direction, but not by sign and additional three failed the robustness test of the NetGenerator-approach.

network size. But there are also distinct differences as the NetGenerator has some features that the adaptive LASSO does not have, like the consideration of external stimuli or directed prior knowledge. The inference process is also much more sophisticated, by creating individual sub models, optimises and assembles them. It is also able to use the replicates for the modeling, while the LASSO took the averaged values as input,

discarding a lot of information. The adaptive LASSO processes the expression data of each gene individually without considering the results for other genes. The price for the elaborate way used by NetGenerator creates the networks inference is the increased run time. The LASSO-based network inference completes the task within seconds, while the NetGenerator needs days. It might therefore be interesting to get first estimates using the adaptive LASSO and create the final model with the NetGenerator.

5 Humidimycin

5.1 Introduction

The occurrence of drug resistance is an emerging problem in antifungal therapies. In addition that, even clinically applied drugs like Caspofungin (see chapter 3.1.1 for details) sometime show unpredicted results, like the so called "paradoxic effect" [66, 96]. It describes the observation, that an increase of the drug dosage leads to a decrease in effectivity. Scientists are continuously on the search for novel antifungal drugs and a complementary approach is the search for compounds with synergistic or enhancing effects. These are compounds that often have little to no effect on their own, but enhance the effectivity of other agents. In an attempt to identify potentially enhancing compounds, 20,000 microbial natural products were extracted from actinomycetes and fungi and tested in combination with a sub-lethal dosage of Caspofungin in an high-throughput screening approach. Of these extracts, 0.94% showed the capability to enhance the effect of Caspofungin. The extracts were tested against two *A. fumigatus* strains, ATCC46645 and CEA17$^{\text{ku80}}$, also known as Δ*akuB*.

From these positively tested extracts, 36.5% showed good dose-response curves and were chosen for further investigation. Using re-fermentation and fractionation, the compounds were obtained and identified. Among them was the peptide Humidimycin, which was harvested from fermentation broths of *Streptomyces humidus* using chromatography. According to mass spectrometry, the molecular formula is $C_{98}H_{132}N_{22}O_{27}S_4$, which gives it a close similarity with siamycin II.

Further studies about the antifungal effect of Humidimycin were done using a concentration of 8 μg/ml with and without a sub-lethal dose of Caspofungin. Tests were performed against *Aspergillus fumigatus*, *Candida albicans*, *Candida glabrata*, *Fusarium oxysporum*, *Cryptococcus neoformans* and *Saccharomyces cerevisiae*. The sub-lethal dose of Caspofungin was defined as 1/8 of the minimal effective/inhibitory concentration (MEC/MIC) for that organism, except for *F. oxysporum* and *C. neoformans*, where a concentration of 1 μg/ml was chosen, as they are not susceptible to Caspofungin.

In *A. fumigatus*, a dose of 1-2 μg/ml Humidimycin caused the MEC to drop from 0.12 to 0.015 μg/ml and in *C. albicans*, the MIC dropped from 0.25 to 0.03 μg/ml. This demonstrated the enhancing effect of Humidimycin in *A. fumigatus* and *C. albicans* for Caspofungin, while combined treatment showed no effect in *C. glabrata*, *F. oxysporum*, *C. neoformans* and *S. cerevisiae*.

5.2 Data & Methods

5.2.1 RNASeq data

In order to investigate the mechanism of the enhancement on a genetic level, RNASeq data was gathered from *A. fumigatus* conidia stressed with Caspofungin, Humidimycin, both compounds in combination and an unstressed control (see table 5.1). The treated samples were taken 4 h after the treatment in order to give *A. fumigatus* time for a response.

Table 5.1: **RNASeq data used in this study.** The used strain is the Δ*akuB* mutant and measurements were taken at three different conditions: stressed with Caspofungin, Humidimycin or both in combination. The samples taken at 0 hour is the untreated control. All samples were taken in three replicates.

	Caspofungin	Humidimycin	Caspofungin Humidimycin
0 hour	3 ×	-	-
4 hours	3 ×	3 ×	3 ×

The RNASeq data was mapped as described in chapter 3.2.3.

5.2.2 Differential expression & Clustering

I considered genes as differentially expressed, if the absolute $\log_2$(fold change) is > 2 and the adjusted p-value < 0.01. Genes of each treatmend were tested against the untreated control (0 h). Genes were categorised according to GO [7] and KEGG [55] using the enrichment analysis tool FungiFun [93].

I performed the subsequent clustering using the R package e1071 [85], with parameters set to try a maximum number of iteration of 500 and 25 tries to find the best clustering. I searched for an optimal number of clusters between a cluster size of 2 and 8. Instead of individual replicates, the total number of reads over all replicates was considered. To compare the relative gene expression change compared to the control and not the absolute values, all gene expressions have been scaled down to fall between [-1, 1].

5.3 Results

In the data that includes the Caspofungin treatment only, 668 gene were differentially expressed, 140 in response to the Humidimycin treatment and 571 in response to the combined approach. In total 833 genes are differentially expressed in any of the samples and the overlap of DEGs between the samples is quite big, as can be seen in figure 5.1.

The Caspofungin treatment and the combined approach share the most DEGs, while the Humidimycin treated DEGs have a rather small overlap. Only 60 DEGs are common in all samples.

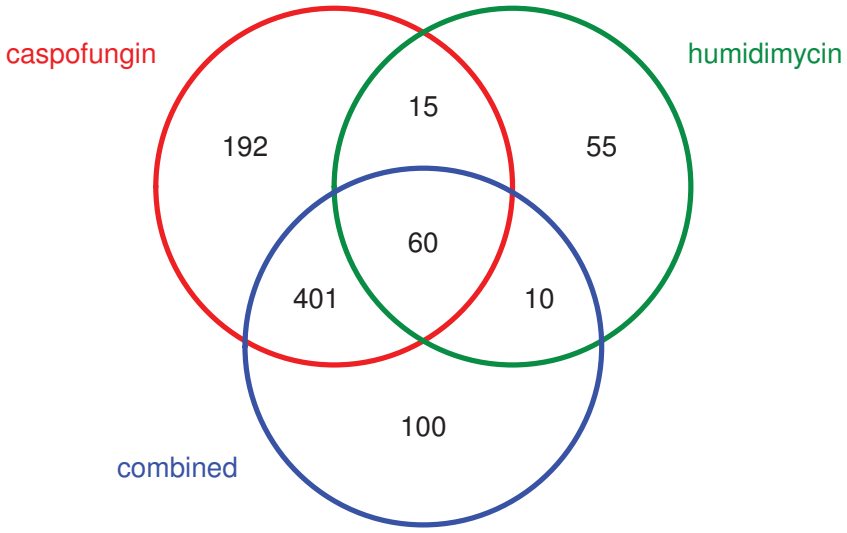

Figure 5.1: **Venn diagram of the differentially expressed genes.**

To further compare the different influence of the treatment, I clustered the differentially expressed genes according to their expression patterns. The result can be seen in figure 5.2. The genes are rather equally distributed between the clusters. Cluster 2 shows genes, which are effected by all treatments. GO-analysis reveals, they mostly belong to the class of "metabolic processes" and "membrane components". The cluster 3 and 5 show genes that are no influenced by the Humidimycin, but by Caspofungin and the combined approach. Cluster 3 is enriched with genes that regulate the metabolic processes of carbonhydrates, while cluster 5 contains more transcription factors than one would expect by random selection. Of special interest is cluster 4, since it contains genes that are only regulated if Humidimycin and Caspofungin are combined. With only 57 members, it is the smallest of all cluster, and 32 of the genes are not annotated in GO yet. The remaining genes come from various categories, which do not draw a clear picture, which process is influenced here. I also checked KEGG, with similar results. Here, 35 genes are without annotation and the remaining genes are distributed among different categories, mostly belong to the main categories "metabolism" and "transport". Cluster 6 on the other hand, combines genes that only show a differential expression under the influence of Humidimycin, not under Caspofungin or a combined approach. Again, many genes (45%) are do not have a GO annotation and the category "metabolic process" has the most members. Among the other categories "oleate hydratase activity" consists of two genes in *A. fumigatus*, AFUB_044650 and AFUB_057580, which are both part of this cluster.

Since Caspofungin especially targets the cell wall, I took a closer look at genes associated with it. I selected 531 genes, belonging to the biosynthesis of the cell wall, disregarding the results of the differential expression analysis. The cluster analysis,

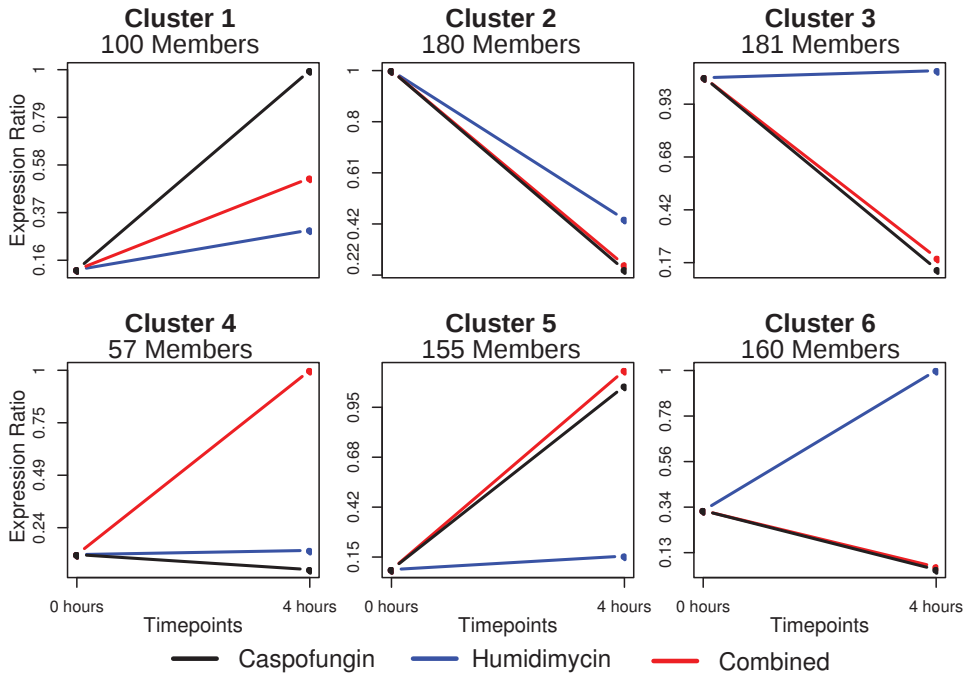

Figure 5.2: **Cluster analysis for all differentially expressed genes.** The 0 h sample represents the untreated control, while the genes show different expression changes depending on the treatment: Caspofungin or Humidimycin alone, or a combination of both.

shown in figure 5.3, returned six cluster similar to those seen in the analysis before. The first two cluster combine genes that have a similar expression profiles in all conditions, first generally down-regulated, then up-regulated due to the treatment. In the third cluster are mainly genes, that react with an up-regulation, when Humidimycin and Caspofungin are combined. The GO-category with the most genes here is “hydrolase activity”, “polysaccharide catabolic process” and “extracellular region”. The cluster 4 consists of genes that seem to react to Humidimycin, alone or in combination with Caspofungin, and is the smallest of all cluster. The cluster contains genes responsible for dephosphorylation and chitin metabolism. The fifth cluster shows genes that only respond to genes that react to Humidimycin alone by an up-regulation and with a down-regulation on Caspofungin, alone or in combination. The GO-category most enriched here is “carbohydrate metabolic process”, “polysaccharide catabolic process” and “hydrolase activity”, which could represent the effort to repair and salvage damaged cell wall molecules under Humidimycin treatment. These genes are not activated when Caspofungin is present, probably because the situation is more dangerous for the fungi and salvaging does not have priority yet. The last cluster collects all genes that are up-regulated during Caspofungin treatment but not by Humidimycin. This group contains a lot of genes coding for transferases, phosphatases and chitin biosynthesis. It is easy to explain that the phosphatases and transferases activate the response to the Caspofungin treatment and the activation of the chitin biosynthesis is the adaption to

the inhibited glucose synthesis caused by Caspofungin.

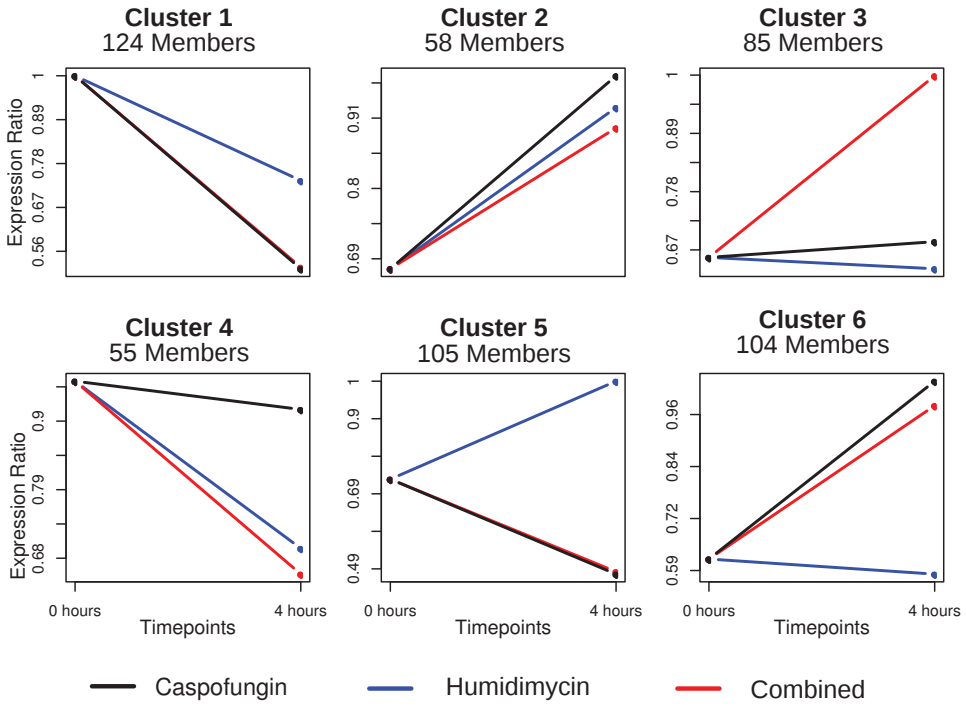

Figure 5.3: **Cluster analysis for all cell wall-related genes.** The 0 h sample represents the untreated control, while the genes show different expression changes depending on the treatment: Caspofungin or Humidimycin alone, or a combination of both.

5.4 Discussion

Investigating the influence different drugs have on each other is not only important to counter the increasing occurrence of resistance in fungi, but also since they allow the decrease of the amount of a certain drug to be administered, which can possibly lower the degree of side effects. Using different treatments in parallel can also cause new side effects and different drugs combined may also block each other out.

Humidimycin seems to enhance the effectiveness of Caspofungin, lowering the necessary dosage needed. Looking at figure 5.1, it can be seen, that the number of DEGs decreases from 668 from the Caspofungin treatment to 571 DEGs in the combined approach. This shows, that Humidimycin, while not damaging the fungus directly, still lowers the drug response. Further analysis shows that the genes cluster in distinct groups according to their reaction to different treatments. When comparing Caspofungin and the combined approach, few genes show different expression profiles, and cluster 4 in figure 5.2 contains genes, that only react to the combined treatment. With 57 members, it is the smallest cluster of the test and 32 of them are not annotated in GO yet, which

makes investigation difficult. The remaining genes belong to different categories, that do draw a clear picture, what sets these genes aside from all others.

The cluster analysis for the cell wall related genes gave a comparable results, even though there are more genes that react to Humidimycin than in the previous clustering. Again, cluster 4 is the smallest of all and contains genes that seem to react to Humidimycin alone and in combination with Caspofungin, but not with Caspofungin alone. Interestingly, this cluster contains genes related to dephosphorylation and the chitin metabolism, which is considered as an important part of the Caspofungin drug adaption. In this cluster, these genes are down-regulated, which may be connected to the enhancing effect of the combined drug approach. Cluster 6 on the other hand contains genes that react to Caspofungin by up-regulation. These are phosphatases and transferases and again chitin biosynthesis, which show the drug response in *A. fumigatus*. Maybe cluster 4 shows the dampening of the drug response signal in *A. fumigatus*, caused by Humidimycin, while cluster 6 shows the genes that are nevertheless activated by the drug Caspofungin.

In conclusion, the only thing that can be safely said is, that Humidimycin lowers the number of DEGs when combined with Caspofungin. It also causes the down-regulation of some parts of the chitin metabolism which may also be part of the enhancing effect towards Caspofungin.

6 Conclusion

6.1 Large-scale network prediction

The main goal aim of this study was to investigate the application of network inference methods for the study of gene regulation. I started a full-genomic approach to create scale-free gene regulatory networks. The presented approach uses the combination of different linear regression methods to investigate multi-experiment microarray data, but can also be used to investigate other high-throughput datasets. I could show that the additional implementation of multiple sources of prior knowledge is able to increase the predictive power of the network model. The use of an automatically harvested gold standard to compare the results of the evaluation was of great help when I had to select the final network. Despite careful programming and a constant increase in computation power, inferring and evaluating large scale networks remain a time consuming task, as seen in the study of chapter 3, where the complete calculation of the final, robust network of 26 genes took about a week on a computer cluster with 32 cores..

An important property the networks showed is sparseness, which increases reliability and interpretability of the networks, by focusing on the most influential network regulations and is also a property assumed in biological networks. The scale-freeness allows the investigation of topological properties, especially the identification of hubs. The model suggested several possible hub genes, which can be drug targets in the future. The experimental testing of drug targets is costly and time consuming, which makes careful computational study beforehand very important, as it can give valuable hints for promising candidates.

The biggest obstacle in my work is the general lack of a sufficient amount of data. Despite the steady increase of the size of high-throughput data, the available amount is still insufficient for reliable large scale predictions, especially for non-model organisms. Of special concern is the gold standard that I used to evaluate the final network. Despite the use of automated text-mining software and scanning of over 6000 full-text research papers, the amount of current knowledge is still very low. Additional effort is necessary to increase the mining output in face of a constantly growing number of publications. Right now, the step from large-scale gene-regulatory networks to the testing of drug targets in the laboratory should be accompanied by an intermediate step: small-scale network inference.

6.2 Small-scale network prediction

In order to get a focused view on the gene regulation of genes, small-scale networks are a valuable tool. The critical first decision here is the selection of candidates for the inference. Large-scale models are one way to identify potential hubs, and therefore genes of interest.

I investigated the adaption of *A. fumigatus* on the clinical drug Caspofungin on a gene regulatory level in chapter 3. I used high quality RNA-Seq data and, as no large-scale network models were available, used the combination of differential expression and functional categorisation coupled with expert knowledge, to select candidate genes for the inference. Again, I used prior knowledge to improve the quality of the prediction. The search in literature revealed only little known information about how *A. fumigatus* regulates the drug adaption. I gained additional knowledge from analysing data of knock-out experiments of known regulators. This additional data is necessary to complement the little expression data, that is available.

The NetGenerator creates network models of different sizes and the lack of a gold standard makes it difficult to evaluate their quality. I decided to judge the networks by how close they are to the current knowledge and their modeling error. Further robustness test helped to separate interactions that are robust to noise from those that are susceptible to noise in the data or changes in the prior knowledge.

From here I derived new hypothesis about the interactions between the regulatory genes. The most important one is the predicted inhibition of *sakA* by *mpkA*, the key regulators of the two main drug response signalling pathways. Tests in the laboratory verified this important prediction *via* western blotting and qRT-PCR. This new interaction can explain some critical expects of the *A. fumigatus* drug adaption, as the genes are known to influence the chitin / glucan balance of the cell wall. Since Caspofungin is known to inhibit the glucan biosynthesis, a shift to chitin synthesis can explain the loss of effectivity. This new knowledge can now used as new input for future studies to unravel even more knowledge about the emergence of resistances in pathogenic fungi.

6.3 Analysis of expression

Sometimes, the amount of available data is not sufficient to create gene regulatory networks, even in small-scale. In chapter 5, I worked with only four data points, each under different conditions. My work was a small part of an extensive study of how Humidimycin influences the antifungal properties of Caspofungin, without having antifungal properties itself. The methods I applied here were also part of the other studies, used in a different context. The identification of differentially expressed genes was used in the small-scale study to identify possible candidate genes. Here, it was used in a broader sense of identifying the global influence a treatment has on the genetic activity. The results showed that the enhancing effect comes with a decrease in the number of differential expressed genes, which can explain, why *A. fumigatus* is not hindered in its ability to adopt to Caspofungin.

Further investigation included clustering to sort the differentially expressed genes according to their reaction to different treatments. This allows further hypothesis of how Humidimycin influences the expression. Detailed examination remained inconclusive, as it is often the case due to a lack of annotated genes. Nevertheless, these study can give valuable hints which genes might be affected by Humidimycin and Caspofungin.

6.4 Comparison of LASSO & NetGenerator

The tools available for network inferences are as various as the scientific questions they try to answer. In my work, I applied mutual information networks, different form of linear regression models and the NetGenerator tool. The direct comparison of my implementation of the adaptive LASSO *via* ridge regression and the NetGenerator in chapter 4 showed the possibilities and limitations of both approaches.

The LASSO-based approach has one indispensable property, when it comes to large-scale network inference: it is fast. To be precise, this is rather thanks to the LARS implementation by Efron *et al.* It made it possible to calculate all LASSO solutions in one run, making the testing for different model sizes tractable. Small-scale modeling that took days with the NetGenerator was done within seconds using the LASSO. Combining the ridge regression for parameter evaluation and the LASSO for parameter selection seems to be a reasonable way to create network models. Not being to discriminate between activating and repressing prior knowledge is certainly a disadvantage, but this feature can be added in future implementations. As it is, the LASSO-based approach is a mature way to model large-scale expression data and allows for a quick overview in small-scale data.

The NetGenerator splits the modeling in several smaller sub-models and optimises them individually, before combining them to a bigger network. While this leads to more accurate results and more implemented prior knowledge, it is also much more time consuming. This limits the number of genes expression profiles, that can be modeled, before the computation time becomes limiting. The NetGenerator also offers a lot more possibilities to influence the modeling results. I focused on the optimisation of three parameters, but there are much more that can positively influence the results of the inference. Without a proper gold standard, however, it is difficult to evaluate the model and optimise the results.

6.5 Final remarks

Eventually, my study revealed the following conclusions: The biggest obstacle to tackle gene regulatory network inference is the lack of high quality data. Often researchers can not be picky when choosing the data for the modeling. However, one must not forget the aim of the study, which has to be well defined. When using different sources of information, it is very easy to include data that introduces a bias in the modeling, since data are always taken under certain conditions, which may not always match your

question of study. Large-scale studies require the highest amount of data and often compendia of different studies from various laboratories are used. These datasets are often very homogeneous, which must not necessarily be a disadvantage. The aim of large-scale studies of often to get an overview of the genetic network. Therefore, having data taken under different conditions may be an advantage. One should keep in mind that not every interaction is active at all times and occur only at certain conditions. Large-scale models usually do not capture these subtleties.

If the study is about specific reactions to fixed conditions, a small-scale analysis is advisable. While in large-scale models the need for quantity of data exceeds the need for quality (up to a certain extend), small-scale models rely on the quality of the data. Here, the research question is often more specific and includes reactions to fixed treatments. It is often necessary to collect data individually for each study, since it is unlikely to find a suitable dataset available public.

While each study aims at discovering new knowledge, it is very important to consider current knowledge and not only include it into the modeling, but also observe, how much the model is able to reflect verified knowledge. A gold standard is therefore an invaluable help, since it allows the evaluation of different networks. These gold standards are often incomprehensive or completely unavailable, especially for non-model organisms. The modeling does not return one single network, but an ensemble of networks in different sizes and structures. Selecting one model as the final one is difficult due to a lack of independent measures. Taking the model error only limits the selection process to the expression data only, ignoring the current knowledge.

The regulation of genes in an organism is a dynamic processes that change over time. The network model is a still frame of the most prominent interactions during the time studied, and by no means does it capture the full complexity of genetic regulations. This is also a reason why different prior knowledge sources and gold standard sometimes have contradicting information or do not match the regulatory model. The information origins from studies under different conditions, that may lead to different predictions on regulations.

In all, scientific research is always a balancing between current and new knowledge. Putting to much known information into a study and forcing it into the model blocks the way to new discoveries and returns only the current knowledge. Not offering any additional information to the model and not guiding it along current knowledge leads to highly unreliable results and disregards the work of other scientists. Both extremes block the path to insight and it is the challenge and responsibility of every scientist to look at the past and the future.

Bibliography

[1] A. Abad, J. V. Fernández-Molina, J. Bikandi, A. Ramírez, J. Margareto, J. Sendino, F. L. Hernando, J. Pontón, J. Garaizar, and A. Rementeria. What makes *Aspergillus fumigatus* a successful pathogen? genes and molecules involved in invasive aspergillosis. *Rev Iberoam Micol*, 27(4):155–182, 2010.

[2] Albert, Jeong, and Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, July 2000.

[3] D. Albrecht, R. Guthke, A. A. Brakhage, and O. Kniemeyer. Integrative analysis of the heat shock response in *Aspergillus fumigatus*. *BMC genomics*, 11(1):32, 2010.

[4] R. Altwasser, J. Linde, and R. Guthke. Genome-wide scale-free network inference for *Candida albicans*. *Frontiers in Microbial Immunology*, 3:51, 2012.

[5] R. Amitani, G. Taylor, E.-N. Elezis, C. Llewellyn-Jones, J. Mitchell, F. Kuze, P. J. Cole, and R. Wilson. Purification and characterization of factors produced by *Aspergillus fumigatus* which affect human ciliated respiratory epithelium. *Infection and immunity*, 63(9):3266–3271, 1995.

[6] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[7] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, and others. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[8] L. R. Asmundsdóttir, H. Erlendsdóttir, and M. Gottfredsson. Increasing incidence of candidemia: results from a 20-year nationwide study in iceland. *J Clin Microbiol*, 40(9):3489–3492, Sept. 2002.

[9] G. D. Bader. BIND: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250, Jan. 2003.

[10] S. Balaji, M. M. Babu, L. M. Iyer, N. M. Luscombe, and L. Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *Journal of Molecular Biology*, 360(1):213–227, June 2006.

[11] V. Balloy and M. Chignard. The innate immune response to *Aspergillus fumigatus*. *Microbes and Infection*, 11(12):919–927, Oct. 2009.

[12] A. Barabási. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct. 1999.

[13] J. C. Bowman, P. S. Hicks, M. B. Kurtz, H. Rosen, D. M. Schmatz, P. A. Liberator, and C. M. Douglas. The antifungal echinocandin caspofungin acetate kills growing cells of *Aspergillus fumigatus* in vitro. *Antimicrobial Agents and Chemotherapy*, 46(9):3001–3012, Sept. 2002.

[14] A. Brakhage. Systemic fungal infections caused by aspergillus species: Epidemiology, infection process and virulence determinants. *Current Drug Targets*, 6(8):875–886, Dec. 2005.

[15] A. A. Brakhage, S. Bruns, A. Thywissen, P. F. Zipfel, and J. Behnsen. Interaction of phagocytes with filamentous fungi. *Curr Opin Microbiol*, 13(4):409–415, Aug. 2010.

[16] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 418–429, 2000.

[17] E. Buyko, E. Faessler, J. Wermter, and U. Hahn. Syntactic simplification and semantic enrichment - trimming dependency graphs for event extraction. *Computational Intelligence*, 27(4):610–644, 2011.

[18] D. Cappelletty and K. Eiselstein-McKitrick. The echinocandins. *Pharmacotherapy*, 27(3):369–388, Mar. 2007.

[19] S. Christley, Q. Nie, and X. Xie. Incorporating existing network information into gene network inference. *PLoS One*, 4(8):e6799, 2009.

[20] V. Czaika, P. Nenoff, A. Glöckner, W. Fegeler, K. Becker, and A. F. Schmalreck. Epidemiology and changes in patient-related factors from 1997 to 2009 in clinical yeast isolates related to dermatology, gynaecology, and paediatrics. *Int J Microbiol*, 2013:703905, 2013.

[21] M. E. da Silva Ferreira, M. R. V. Z. Kress, M. Savoldi, M. H. S. Goldman, A. Hartl, T. Heinekamp, A. A. Brakhage, and G. H. Goldman. The akuBKU80 mutant deficient for nonhomologous end joining is a powerful tool for analyzing pathogenicity in *Aspergillus fumigatus*. *Eukaryotic Cell*, 5(1):207–211, Jan. 2006.

[22] T. R. T. Dagenais and N. P. Keller. Pathogenesis of *Aspergillus fumigatus* in invasive aspergillosis. *Clinical Microbiology Reviews*, 22(3):447–465, July 2009.

[23] N. Delhomme, I. Padioleau, E. E. Furlong, and L. M. Steinmetz. easyRNASeq: a bioconductor package for processing RNA-seq data. *Bioinformatics*, 28(19):2532–2533, Oct. 2012.

[24] C. D'Enfert and B. Hube. *Candida: comparative and functional genomics*. Caister Academic Press. Norfolk,U.K., 2007.

[25] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. L. Gall, B. Schaëffer, S. L. Crom, M. Guedj, and Jaffrézic. A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*, Sept. 2012.

[26] C. Du, J. Sarfati, J.-P. Latge, and R. Calderone. The role of the sakA (hog1) and tcsB (sln1) genes in the oxidant adaptation of *Aspergillus fumigatus*. *Medical Mycology*, 44(3):211–218, Jan. 2006.

[27] B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. D. Montigny, C. Marck, C. Neuvéglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J.-M. Beckerich, E. Beyne, C. Bleykasten, A. Boisramé, J. Boyer, L. Cattolico, F. Confanioleri, A. D. Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J.-M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G.-F. Richard, M.-L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wésolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker, and J.-L. Souciet. Genome evolution in yeasts. *Nature*, 430(6995):35–44, July 2004.

[28] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, Jan. 2000.

[29] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, Aug. 1998.

[30] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1):e8, 2007.

[31] C. V. Forst. Host-pathogen systems biology. *Drug discovery today*, 11(5-6):220–227, Mar. 2006.

[32] J. R. Fortwendel, P. R. Juvvadi, B. Z. Perfect, L. E. Rogg, J. R. Perfect, and W. J. Steinbach. Transcriptional regulation of chitin synthases by calcineurin controls paradoxical growth of *Aspergillus fumigatus* in response to caspofungin. *Antimicrobial Agents and Chemotherapy*, 54(4):1555–1563, Apr. 2010.

[33] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[34] C. Garcia-Vidal, D. Viasus, and J. Carratalà. Pathogenesis of invasive fungal infections. *Curr Opin Infect Dis*, 26(3):270–276, June 2013.

[35] A. Gastebois, C. Clavaud, V. Aimanianda, and J.-P. Latgé. *Aspergillus fumigatus*: cell wall polysaccharides, their biosynthesis and organization. *Future Microbiology*, 4(5):583–595, June 2009.

[36] J. M. Gilsenan, J. Cooley, and P. Bowyer. CADRE: the Central Aspergillus Data REpository 2012. *Nucleic Acids Res*, 40(Database issue):D660–D666, Jan. 2012.

[37] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):546–567, Oct. 1996.

[38] N. Grahl, K. M. Shepardson, D. Chung, and R. A. Cramer. Hypoxia and fungal pathogenesis: to air or not to air? *Eukaryot Cell*, 11(5):560–570, May 2012.

[39] M. Gustafsson. Gene networks from high-throughput data: Reverse engineering and analysis. 2010.

[40] M. Gustafsson, M. Hörnquist, and A. Lombardi. Large-scale reverse engineering by the lasso. *http://arxiv.org/abs/q-bio/0403012v1*, Mar. 2004.

[41] M. Gustafsson, M. Hörnquist, and A. Lombardi. Constructing and analyzing a large-scale gene-to-gene regulatory network lasso-constrained inference and biological validation. *IEEE/ACM transactions on computational biology and bioinformatics*, 2(3):254–261, 2005.

[42] R. Guthke, O. Kniemeyer, D. Albrecht, A. A. Brakhage, and U. Möller. Discovery of gene regulatory networks in *Aspergillus fumigatus*. In *Knowledge discovery and emergent complexity in bioinformatics*, page 22–41. Springer, 2007.

[43] R. Guthke, U. Möller, M. Hoffmann, F. Thies, and S. Töpfer. Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics*, 21(8):1626–1634, Apr. 2005.

[44] U. Güldener, M. Münsterkötter, M. Oesterheld, P. Pagel, A. Ruepp, H.-W. Mewes, and V. Stümpflen. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Research*, 34(Database issue):D436–441, Jan. 2006.

[45] L.-P. Hamel, M.-C. Nicole, S. Duplessis, and B. E. Ellis. Mitogen-activated protein kinase signaling in plant-interacting fungi: distinct messages from conserved messengers. *Plant Cell*, 24(4):1327–1351, Apr. 2012.

[46] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88–93, July 2004.

[47] X. He and J. Zhang. Why do hubs tend to be essential in protein networks? *PLoS Genet*, 2(6):e88, June 2006.

[48] M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke. Gene regulatory network inference: Data integration in dynamic models—a review. *Biosystems*, 96(1):86–103, Apr. 2009.

[49] F. Horn, V. Valiante, R. Guthke, and A. A. Brakhage. Data-driven systems biology of fungal infections. In *Human Pathogenic Fungi: Molecular Biology and Pathogenic Mechanisms*. Caister Academic Press, June 2014.

[50] B. Hube. Infection-associated genes of *Candida albicans*. *Future Microbiol*, 1(2):209–218, Aug. 2006.

[51] J. Ihmels, S. Bergmann, J. Berman, and N. Barkai. Comparative gene expression analysis by a differential clustering approach: Application to the *Candida albicans* transcription program. *PLoS Genet*, 1(3):e39, 2005.

[52] W. R. Jarvis. Epidemiology of nosocomial fungal infections, with emphasis on candida species. *Clin Infect Dis*, 20(6):1526–1530, June 1995.

[53] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001.

[54] M. Johnston. A model fungal gene regulatory mechanism: the GAL genes of saccharomyces cerevisiae. *Microbiol Rev*, 51(4):458–476, Dec. 1987.

[55] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–D114, Jan. 2012.

[56] M. Karthaus. Prophylaxis and treatment of invasive aspergillosis with voriconazole, posaconazole and caspofungin: review of the literature. *European journal of medical research*, 16(4):145–152, 2011.

[57] A. Kasprzyk. BioMart: driving a paradigm change in biological data management. *Database*, 2011(0):bar049–bar049, Nov. 2011.

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

[59] M. A. Kaufman, G. J. Duke, F. McGain, C. French, C. Aboltins, G. Lane, and G. A. Gutteridge. Life-threatening respiratory failure from h1n1 influenza 09 (human swine influenza). *Medical Journal of Australia*, 191(3), 2009.

[60] J.-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. Extracting bio-molecular events from literature—the bionlp'09 shared task. *Computational Intelligence*, 27(4):513–540, 2011.

[61] J. Kohler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Ruegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–1390, June 2006.

[62] S. Kozhenkov, M. Sedova, Y. Dubinina, A. Gupta, A. Ray, J. Ponomarenko, and M. Baitaluk. BiologicalNetworks-tools enabling the integration of multi-scale data for the host-pathogen studies. *BMC systems biology*, 5(1):7, 2011.

[63] S. Krappmann, C. Sasse, and G. H. Braus. Gene targeting in *Aspergillus fumigatus* by homologous recombination is facilitated in a nonhomologous end- joining-deficient genetic background. *Eukaryotic Cell*, 5(1):212–215, Jan. 2006.

[64] V. Krcmery and A. J. Barnes. Non-albicans candida spp. causing fungaemia: pathogenicity and antifungal resistance. *The Journal of Hospital Infection*, 50(4):243–260, Apr. 2002.

[65] K. Kryszczuk and P. Hurley. Estimation of the number of clusters using multiple clustering validity indices. In *Multiple Classifier Systems*, page 114–123. Springer, 2010.

[66] F. Lamoth, P. R. Juvvadi, E. J. Soderblom, M. A. Moseley, Y. G. Asfaw, and W. J. Steinbach. Identification of a key lysine residue in heat shock protein 90 required for azole and echinocandin resistance in *Aspergillus fumigatus*. *Antimicrobial Agents and Chemotherapy*, 58(4):1889–1896, Apr. 2014.

[67] J. P. Latgé. *Aspergillus fumigatus* and aspergillosis. *Clin Microbiol Rev*, 12(2):310–350, Apr. 1999.

[68] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, Aug. 2009.

[69] J. Linde, E. Buyko, R. Altwasser, U. Hahn, and R. Guthke. Full-genomic network inference for non-model organisms: A case study for the fungal pathogen *Candida albicans*. In *Proc. International Conf. on Bioinformatics, Computational Biology and Biomedical Engineering, ICBCBBE2011*, Paris, France, Aug. 2011. WASET.

[70] J. Linde, P. Hortschansky, E. Fazius, A. A. Brakhage, R. Guthke, and H. Haas. Regulatory interactions for iron homeostasis in *Aspergillus fumigatus* inferred by a systems biology approach. *BMC Syst Biol*, 6:6, 2012.

[71] J. Linde, D. Wilson, B. Hube, and R. Guthke. Regulatory network modelling of iron acquisition by a fungal pathogen in contact with epithelial cells. *BMC Syst Biol*, 4:148, 2010.

[72] T. T. Liu, R. E. B. Lee, K. S. Barker, R. E. Lee, L. Wei, R. Homayouni, and P. D. Rogers. Genome-wide expression profiling of the response to azole, polyene, echinocandin, and pyrimidine antifungal agents in *Candida albicans*. *Antimicrobial Agents and Chemotherapy*, 49(6):2226–2236, June 2005.

[73] D. Lohr, P. Venkov, and J. Zlatanova. Transcriptional regulation in the yeast GAL gene family: a complex genetic network. *FASEB J*, 9(9):777–787, June 1995.

[74] O. Marchetti, J. Bille, U. Fluckiger, P. Eggimann, C. Ruef, J. Garbino, T. Calandra, M.-P. Glauser, M. G. Täuber, D. Pittet, and Fungal Infection Network of Switzerland. Epidemiology of candidemia in swiss tertiary care hospitals: secular trends, 1991-2000. *Clin Infect Dis*, 38(3):311–320, Feb. 2004.

[75] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

[76] G. S. Martin, D. M. Mannino, S. Eaton, and M. Moss. The epidemiology of sepsis in the united states from 1979 through 2000. *N Engl J Med*, 348(16):1546–1554, Apr. 2003.

[77] E. Meyer, C. Geffers, P. Gastmeier, and F. Schwab. No increase in primary nosocomial candidemia in 682 german intensive care units during 2006 to 2011. *Euro Surveill*, 18(24), 2013.

[78] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007:1–9, 2007.

[79] P. E. Meyer, F. Lafitte, and G. Bontempi. minet: A R/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9(1):461, 2008.

[80] V. Meyer, R. A. Damveld, M. Arentshorst, U. Stahl, C. A. M. J. J. van den Hondel, and A. F. J. Ram. Survival in the presence of antifungals: Genome-wide expression profiling of *Aspergillus niger* in response to sublethal concentrations of caspofungin and fenpropimorph. *Journal of Biological Chemistry*, 282(45):32935–32948, Nov. 2007.

[81] C. A. Munro, S. Selvaggini, I. de Bruijn, L. Walker, M. D. Lenardon, B. Gerssen, S. Milne, A. J. P. Brown, and N. A. R. Gow. The PKC, HOG and ca^{2+} signalling pathways co-ordinately regulate chitin synthesis in *Candida albicans*. *Molecular Microbiology*, 63(5):1399–1413, Mar. 2007.

[82] A. M. A. Murad, C. d'Enfert, C. Gaillardin, H. Tournu, F. Tekaia, D. Talibi, D. Marechal, V. Marchais, J. Cottin, and A. J. Brown. Transcript profiling in *Candida albicans* reveals new cellular functions for the transcriptional repressors CaTup1, CaMig1 and CaNrg1. *Molecular microbiology*, 42(4):981–993, 2001.

[83] S. Müller, C. Baldin, M. Groth, R. Guthke, O. Kniemeyer, A. A. Brakhage, and V. Valiante. Comparison of transcriptome technologies in the pathogenic fungus *Aspergillus fumigatus* reveals novel insights into the genome and MpkA dependent gene expression. *BMC Genomics*, 13(1):519, Oct. 2012.

[84] W. C. Nierman, A. Pain, M. J. Anderson, J. R. Wortman, H. S. Kim, J. Arroyo, M. Berriman, K. Abe, D. B. Archer, C. Bermejo, J. Bennett, P. Bowyer, D. Chen, M. Collins, R. Coulsen, R. Davies, P. S. Dyer, M. Farman, N. Fedorova, N. Fedorova, T. V. Feldblyum, R. Fischer, N. Fosker, A. Fraser, J. L. García, M. J. García, A. Goble, G. H. Goldman, K. Gomi, S. Griffith-Jones, R. Gwilliam, B. Haas, H. Haas, D. Harris, H. Horiuchi, J. Huang, S. Humphray, J. Jiménez, N. Keller, H. Khouri, K. Kitamoto, T. Kobayashi, S. Konzack, R. Kulkarni, T. Kumagai, A. Lafon, A. Lafton, J.-P. Latgé, W. Li, A. Lord, C. Lu, W. H. Majoros, G. S. May, B. L. Miller, Y. Mohamoud, M. Molina, M. Monod, I. Mouyna, S. Mulligan, L. Murphy, S. O'Neil, I. Paulsen, M. A. Peñalva, M. Pertea, C. Price, B. L. Pritchard, M. A. Quail, E. Rabbinowitsch, N. Rawlins, M.-A. Rajandream, U. Reichard, H. Renauld, G. D. Robson, S. Rodriguez de Córdoba, J. M. Rodríguez-Peña, C. M. Ronning, S. Rutter, S. L. Salzberg, M. Sanchez, J. C. Sánchez-Ferrero, D. Saunders, K. Seeger, R. Squares, S. Squares, M. Takeuchi, F. Tekaia, G. Turner, C. R. Vazquez de Aldana, J. Weidman, O. White, J. Woodward, J.-H. Yu, C. Fraser, J. E. Galagan, K. Asai, M. Machida, N. Hall, B. Barrell, and D. W. Denning. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature*, 438(7071):1151–1156, Dec. 2005.

[85] N. R. Pal, J. C. Bezdek, and R. J. Hathaway. Sequential competitive learning and the fuzzy c-means clustering algorithms. *Neural Networks*, 9(5):787–796, July 1996.

[86] K. Pang, H. Sheng, and X. Ma. Understanding gene essentiality by finely characterizing hubs in the yeast protein interaction network. *Biochem Biophys Res Commun*, 401(1):112–116, Oct. 2010.

[87] E. Paramythiotou, F. Frantzeskaki, A. Flevari, A. Armaganidis, and G. Dimopoulos. Invasive fungal infections in the ICU: How to approach, how to treat. *Molecules*, 19(1):1085–1119, 2014.

[88] S. Paris, E. Boisvieux-Ulrich, B. Crestani, O. Houcine, D. Taramelli, L. Lombardi, and J.-P. Latge. Internalization of *Aspergillus fumigatus* conidia by epithelial and endothelial cells. *Infection and immunity*, 65(4):1510–1514, 1997.

[89] G. R. Pasha and M. A. Shah. Application of ridge regression to multicollinear data. *Journal of research science*, 15:97–106, 2004.

[90] J. Perlroth, B. Choi, and B. Spellberg. Nosocomial fungal infections: epidemiology, diagnosis, and treatment. *Med Mycol*, 45(4):321–346, June 2007.

[91] M. A. Pfaller and D. J. Diekema. Epidemiology of invasive candidiasis: a persistent public health problem. *Clin Microbiol Rev*, 20(1):133–163, Jan. 2007.

[92] E. Poikonen, O. Lyytikäinen, V.-J. Anttila, and P. Ruutu. Candidemia in finland, 1995-1999. *Emerg Infect Dis*, 9(8):985–990, Aug. 2003.

[93] S. Priebe, J. Linde, D. Albrecht, R. Guthke, and A. A. Brakhage. FungiFun: a web-based application for functional categorization of fungal genes and proteins. *Fungal Genet Biol*, 48(4):353–358, Apr. 2011.

[94] C. J. V. Rijsbergen. *Information Retrieval (2nd ed.)*, volume 30. London: Butterworths, Nov. 1979.

[95] N. Rispail, D. M. Soanes, C. Ant, R. Czajkowski, A. Grünler, R. Huguet, E. Perez-Nadales, A. Poli, E. Sartorel, V. Valiante, M. Yang, R. Beffa, A. A. Brakhage, N. A. R. Gow, R. Kahmann, M.-H. Lebrun, H. Lenasi, J. Perez-Martin, N. J. Talbot, J. Wendland, and A. Di Pietro. Comparative genomics of MAP kinase and calcium-calcineurin signalling components in plant and human pathogenic fungi. *Fungal Genet Biol*, 46(4):287–298, Apr. 2009.

[96] E. M. F. Rocha, G. Garcia-Effron, S. Park, and D. S. Perlin. A ser678pro substitution in fks1p confers resistance to echinocandin drugs in *Aspergillus fumigatus*. *Antimicrobial Agents and Chemotherapy*, 51(11):4174–4176, Nov. 2007.

[97] A. Rokas and C. T. Hittinger. Transcriptional rewiring: the proof is in the eating. *Curr Biol*, 17(16):R626–R628, Aug. 2007.

[98] M. Ruhnke. Epidemiology of *Candida albicans* infections and role of non-*Candida-albicans* yeasts. *Curr Drug Targets*, 7(4):495–504, Apr. 2006.

[99] R. G. R. S. Yan and D. D. Roberts. Hemoglobin-induced binding of candida albicans to the cell-binding domain of fibronectin is independent of the arg-gly-asp sequence. *Infect. Immun.*, 66(5):1904–1909, May 1998.

[100] P. Sandven, L. Bevanger, A. Digranes, P. Gaustad, H. H. Haukland, and M. Steinbakk. Constant low rate of fungemia in norway, 1991 to 1996. The Norwegian Yeast Study Group. *J Clin Microbiol*, 36(12):3455–3459, Dec. 1998.

[101] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences*, 93(20):10614–10619, 1996.

[102] E. Shelest and E. Fazius. Database for fungal transcription factors. Jena, Sept. 2011.

[103] M. S. Skrzypek, M. B. Arnaud, M. C. Costanzo, D. O. Inglis, P. Shah, G. Binkley, S. R. Miyasato, and G. Sherlock. *Candida genome database*. Published: Website: http://www.candidagenome.org.

[104] M. S. Skrzypek, M. B. Arnaud, M. C. Costanzo, D. O. Inglis, P. Shah, G. Binkley, S. R. Miyasato, and G. Sherlock. New tools at the candida genome database: biochemical pathways and full-text literature search. *Nucleic Acids Research*, 38(Database):D428–D432, Jan. 2010.

[105] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, Feb. 2011.

[106] W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig. pcaMethods a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, May 2007.

[107] J. E. Stajich, T. Harris, B. P. Brunk, J. Brestelli, S. Fischer, O. S. Harb, J. C. Kissinger, W. Li, V. Nayak, D. F. Pinney, C. J. Stoeckert, and D. S. Roos. FungiDB: an integrated functional genomics database for fungi. *Nucleic Acids Research*, 40(D1):D675–D681, Jan. 2012.

[108] D. A. Stevens, M. Ichinomiya, Y. Koshi, and H. Horiuchi. Escape of candida from caspofungin inhibition at concentrations above the MIC (paradoxical effect) accomplished by increased cell wall chitin; evidence for beta-1,6-glucan synthesis inhibition by caspofungin. *Antimicrob Agents Chemother*, 50(9):3160–3161, Sept. 2006.

[109] G. Stolovitzky, D. Monroe, and A. Califano. Dialogue on reverse-engineering assessment and methods: The DREAM of high-throughput pathway inference. *Annals of the New York Academy of Sciences*, 1115(1):1–22, Oct. 2007.

[110] J. M. Stuart. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, Oct. 2003.

[111] P. E. Sudbery. Growth of *Candida albicans* hyphae. *Nat Rev Microbiol*, 9(10):737–748, Oct. 2011.

[112] R. Sutak, E. Lesuisse, J. Tachezy, and D. R. Richardson. Crusade for iron: iron uptake in unicellular eukaryotes and its significance for virulence. *Trends Microbiol*, 16(6):261–268, June 2008.

[113] Z. Szallasi. Genetic network analysis in light of massively parallel biological data acquisition. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 5–16, 1999.

[114] J. W. Taylor and M. L. Berbee. Dating divergences in the fungal tree of life: review and new analyses. *Mycologia*, 98(6):838–849, 2006.

[115] R. C. Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[116] R. Tibshirani. Regression shrinkage and selection via the Lasso. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 58:267—288, 1994.

[117] R. Tibshirani, I. Johnstone, T. Hastie, and B. Efron. Least angle regression. *The Annals of Statistics*, 32(2):407–499, Apr. 2004.

[118] L. Tierney, J. Linde, S. Müller, S. Brunke, J. C. Molina, B. Hube, U. Schöck, R. Guthke, and K. Kuchler. An interspecies regulatory network inferred from simultaneous RNA-seq of *Candida albicans* invading innate immune cells. *Frontiers in Microbiology*, 3, 2012.

[119] S. Toepfer, R. Guthke, D. Driesch, D. Woetzel, and M. Pfaff. The NetGenerator algorithm: Reconstruction of gene regulatory networks. In K. Tuyls, R. Westra, Y. Saeys, and A. Nowé, editors, *Knowledge Discovery and Emergent Complexity in Bioinformatics*, volume 4366 of *Lecture Notes in Computer Science*, pages 119–130. Springer Berlin Heidelberg, 2007.

[120] A. Tragiannidis, C. Tsoulas, K. Kerl, and A. H. Groll. Invasive candidiasis: update on current pharmacotherapy options and future perspectives. *Expert Opin Pharmacother*, 14(11):1515–1528, Aug. 2013.

[121] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7(3):562–578, Mar. 2012.

[122] S. Urbanek. *multicore: Parallel processing of R code on machines with multiple cores or CPUs*. 2011. R package version 0.1-7.

[123] V. Valiante, R. Jain, T. Heinekamp, and A. A. Brakhage. The MpkA MAP kinase module regulates cell wall integrity signaling and pyomelanin formation in *Aspergillus fumigatus*. *Fungal Genetics and Biology*, 46(12):909–918, Dec. 2009.

[124] P. E. B. Verwer, M. L. van Duijn, M. Tavakol, I. A. J. M. Bakker-Woudenberg, and W. W. J. van de Sande. Reshuffling of *Aspergillus fumigatus* cell wall components chitin and -glucan under the influence of caspofungin or nikkomycin z alone or in combination. *Antimicrobial Agents and Chemotherapy*, 56(3):1595–1598, Mar. 2012.

[125] J.-L. Vincent, J. Rello, J. Marshall, E. Silva, A. Anzueto, C. D. Martin, R. Moreno, J. Lipman, C. Gomersall, Y. Sakr, K. Reinhart, and E. P. I. C. I. G. o. Investigators. International study of the prevalence and outcomes of infection in intensive care units. *JAMA*, 302(21):2323–2329, Dec. 2009.

[126] A. Voss, J. A. Kluytmans, J. G. Koeleman, L. Spanjaard, C. M. Vandenbroucke-Grauls, H. A. Verbrugh, M. C. Vos, A. Y. Weersink, J. A. Hoogkamp-Korstanje, and J. F. Meis. Occurrence of yeast bloodstream infections between 1987 and 1995 in five dutch university hospitals. *Eur J Clin Microbiol Infect Dis*, 15(12):909–912, Dec. 1996.

[127] L. A. Walker, C. A. Munro, I. de Bruijn, M. D. Lenardon, A. McKinnon, and N. A. R. Gow. Stimulation of chitin synthesis rescues *Candida albicans* from echinocandins. *PLoS Pathogens*, 4(4):e1000040, Apr. 2008.

[128] Z. Wang, M. Gerstein, and M. Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, Jan. 2009.

[129] D. W. Warnock. Trends in the epidemiology of invasive fungal infections. *Nihon Ishinkin Gakkai Zasshi*, 48(1):1–12, 2007.

[130] J. A. Wasylnka. *Aspergillus fumigatus* conidia survive and germinate in acidic organelles of a549 epithelial cells. *Journal of Cell Science*, 116(8):1579–1587, Apr. 2003.

[131] M. Weber, S. G. Henkel, S. Vlaic, R. Guthke, E. J. v. Zoelen, and D. Driesch. Inference of dynamical gene-regulatory networks based on time-resolved multi-stimuli multi-experiment data applying NetGenerator v2.0. *BMC Syst Biol*, 7:1, 2013.

[132] R. P. Wenzel. Nosocomial candidemia: risk factors and attributable mortality. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 20(6):1531–1534, June 1995.

[133] A. V. Werhli and D. Husmeier. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*, 6(1), Jan. 2007.

[134] E. Wingender, P. Dietze, H. Karas, and R. Knüppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic acids research*, 24(1):238–241, 1996.

[135] M. S. Yeung, J. Tegnér, and J. J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences*, 99(9):6163–6168, 2002.

[136] H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. d. Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110, Oct. 2008.

[137] P. Zoppoli, S. Morganella, and M. Ceccarelli. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics*, 11(1):154, 2010.

[138] E. Zotenko, J. Mestre, D. P. O'Leary, and T. M. Przytycka. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol*, 4(8):e1000140, 2008.

[139] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, Dec. 2006.

7 Appendix

The attached CD contains the data used in this thesis.
The directory "candida" contains the following files:

imputedMatrix.csv: The imputed microarray data, that was used to infer the network in chapter 2.

goldStandard.csv: The result of the automated text mining done by the Buyko *et al.* This data was used as gold standard to evaluate the network.

pkBind.csv The BIND prior knowledge used in this work.

pkFac.csv The FAC prior knowledge used in this work.

pkPpi.csv The PPI prior knowledge used in this work.

pkTrans.csv The TRANS prior knowledge used in this work.

The directory "aspergillus" contains the following files:

rawCounts.csv: Contains the raw, unnormalised RNASeq-counts of *A. fumigatus*. Used in the chapters 3, 4 and 5.

prior.knowledge.csv: List with the prior knowledge used in the chapters 3 and 4.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel, persönlichen Mitteilungen und Quellen angefertigt habe. Mir ist die geltende Promotionsordnung bekannt. Ich habe weder die Hilfe eines Promotionsberaters in Anspruch genommen, noch haben Dritte unmittelbare oder mittelbare geldwerte Leistungen für Arbeit erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die vorgelegte Dissertation wurde bisher nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Weiterhin habe ich mich mit der vorliegenden Arbeit an keiner anderen Hochschule um einen akademischen Grad beworben und weder früher noch gegenwärtig die Eröffnung eines Verfahrens zum Erwerb eines akademischen Grades an einer anderen Hochschule beantragt.

Jena, der 2. Februar 2015

Robert Altwasser