Functional Analysis of High-Throughput Data for Dynamic Modeling in Eukaryotic Systems

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

im Fach Biophysik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät I

der Humboldt-Universität zu Berlin

von

Max Flöttmann

Präsident der Humboldt-Universität zu Berlin:

Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:

Prof. Stefan Hecht, PhD

Gutachter/innen:
1. Prof. Dr. Dr. h. c. Edda Klipp
2. Prof. Dr. Andreas Herrmann
3. Prof. Dr. Ralf Mrowka

Tag der mündlichen Prüfung: 05.06.2013

Thesis advisor: Prof. Dr. Dr. Edda Klipp Max Flöttmann

Functional Analysis of High-Throughput Data for Dynamic Modeling in Eukaryotic Systems

Abstract

The behavior of all biological systems is governed by numerous regulatory mechanisms, acting on different levels of time and space. The study of these regulations has greatly benefited from the immense amount of data that has become available from high-throughput experiments in recent years. To interpret this mass of data and gain new knowledge about studied systems, mathematical modeling has proven to be an invaluable method. Nevertheless, before data can be integrated into a model it needs to be aggregated, analyzed, and the most important aspects need to be extracted.

We present four Systems Biology studies on different cellular organizational levels and in different organisms. Additionally, we describe two software applications that enable easy comparison of data and model results. We use these in two of our studies on the mitogen-activated-protein (MAP) kinase signaling in *Saccharomyces cerevisiae* to generate model alternatives and adapt our representation of the system to biological data. In the two remaining studies we apply Bioinformatic methods to analyze two high-throughput time series on proteins and mRNA expression in mammalian cells. We combine the results with network data and use annotations to identify modules and pathways that change in expression over time to be able to interpret the datasets. In case of the human somatic cell reprogramming (SCR) system this analysis leads to the generation of a probabilistic Boolean model which we use to generate new hypotheses about the system. In the last system we examined, the infection of mammalian (*Canis familiaris*) cells by the influenza A virus, we find new interconnections between host and virus and are able to integrate our data with existing networks.

In summary, many of our findings show the importance of data integration into mathematical models and the high degree of connectivity between different levels of regulation.

Betreuer der Arbeit: Prof. Dr. Dr. Edda Klipp Max Flöttmann

Functional Analysis of High-Throughput Data for Dynamic Modeling in Eukaryotic Systems

Zusammenfassung

Das Verhalten Biologischer Systeme wird durch eine Vielzahl regulatorischer Prozesse beeinflusst, die sich auf verschiedenen Ebenen abspielen. Die Forschung an diesen Regulationen hat stark von den großen Mengen von Hochdurchsatzdaten profitiert, die in den letzten Jahren verfügbar wurden. Um diese Daten zu interpretieren und neue Erkenntnisse aus ihnen zu gewinnen, hat sich die mathematische Modellierung als hilfreich erwiesen. Allerdings müssen die Daten vor der Integration in Modelle aggregiert und analysiert werden.

Wir präsentieren vier Studien auf unterschiedlichen zellulären Ebenen und in verschiedenen Organismen. Zusätzlich beschreiben wir zwei Computerprogramme die den Vergleich zwischen Modell und Experimentellen Daten erleichtern. Wir wenden diese Programme in zwei Studien über die MAP Kinase (MAP, engl. mitogen-acticated-protein) Signalwege in *Saccharomyces cerevisiae* an, um Modellalternativen zu generieren und unsere Vorstellung des Systems an Daten anzupassen. In den zwei verbleibenden Studien nutzen wir bioinformatische Methoden, um Hochdurchsatz-Zeitreihendaten von Protein und mRNA Expression zu analysieren. Um die Daten interpretieren zu können kombinieren wir sie mit Netzwerken und nutzen Annotationen um Module identifizieren, die ihre Expression im Lauf der Zeit ändern. Im Fall der humanen somatischen Zell Reprogrammierung führte diese Analyse zu einem probabilistischen Boolschen Modell des Systems, welches wir nutzen konnten um neue Hypothesen über seine Funktionsweise aufzustellen. Bei der Infektion von Säugerzellen (*Canis familiaris*) mit dem Influenza A Virus konnten wir neue Verbindungen zwischen dem Virus und seinem Wirt herausfinden und unsere Zeitreihendaten in bestehende Netzwerke einbinden.

Zusammenfassend zeigen viele unserer Ergebnisse die Wichtigkeit von Datenintegration in mathematische Modelle, sowie den hohen Grad der Verschaltung zwischen verschiedenen Regulationssystemen.

Contents

1 Introduction 1

1.1 Outline . . . 1

1.2 Systems Biology . . . 3

1.3 Mathematical Modeling of Biological Systems . . . 4

1.3.1 Scope of a Model . . . 4

1.3.2 Data Driven Modeling . . . 6

2 Methods 9

2.1 Biological Methods . . . 9

2.1.1 Mass Spectrometry . . . 9

2.1.2 Expression Profiling by Microarrays . . . 12

2.2 Functional Analysis . . . 14

2.2.1 Clustering of Timecourse Data . . . 14

2.2.2 Enrichment in Functional Databases . . . 16

2.2.3 Proteomic Phenotyping . . . 17

2.2.4 Network Component Analysis . . . 18

2.3 Dynamic Modeling . . . 18

2.3.1 Boolean Modeling . . . 18

2.3.2 Ordinary Differential Equations . . . 23

3 Software Frameworks for Accelerated Modeling 27

3.1 Introduction . . . 28

3.1.1 Available Software . . . 28

3.1.2 MAP Kinase Pathways in Yeast . . . 28

3.2 Model Management and Generation for SBML Models 29
3.2.1 Model Formats and Workflow 30
3.2.2 Analyzing Feedback Mechanisms in the Sho1 Branch of the HOG Pathway 31
3.2.3 Model Discrimination 32
3.2.4 Model Predictions and Validation 33
3.3 Rxncon for Boolean Models 34
3.3.1 The Rxncon Format 36
3.3.2 Bipartite Boolean Models 36
3.3.3 Iterative Model Building and Validation 42
3.4 Discussion 45

4 Gene Regulatory Networks in Pluripotency **49**
4.1 Introduction 50
4.1.1 Stem cells and somatic reprogramming 50
4.2 Results 56
4.2.1 Gene Regulatory Networks that Govern Pluripotency 56
4.2.2 A Probabilistic Boolean Model of Somatic Reprogramming and Differentiation 66
4.3 Discussion 85

5 Proteomes of Influenza Virus and its Host in the Course of the Infection Cycle **89**
5.1 Introduction 89
5.1.1 Influenza A Infection and Host Virus Interaction 89
5.2 Results 94
5.2.1 Virus Proteins Show Strong Differences in Dynamics of Translation 96
5.2.2 Influenza Infection Influences Protein Production in the Host . 97
5.2.3 A Literature Derived Network Helps Interpreting Proteome Data 101
5.2.4 Virus Host Interactions 102
5.3 Discussion 103

6 Discussion and Outlook **107**
6.1 From Data to Models 108

6.2 Modeling Software Catalyzes Progress 108
6.3 Biological Advances 109
6.3.1 Somatic Cell Reprogramming 109
6.3.2 Influenza Infection 110
6.3.3 Different Levels - Different Perspective 110
6.4 All Levels are Equal 111
6.5 Outlook and Further Work 112
6.6 Concluding Remarks 113

References 114

A Software 127
A.1 Implementation 127
A.2 *ModelMaGe* 127

B Pluripotency 137

C Influenza A 151

Listing of figures

1.3.1 Scheme of Systems Biology research 5

1.3.2 Modeling frameworks 6

2.1.1 Mass spectrometry experiments 10

2.1.2 Stable isotopic labeling of amino acids in cell culture 12

2.1.3 Microarray experiments 13

2.2.1 Principle of the NCA analysis. 19

2.3.1 Basins of attraction 21

3.2.1 *ModelMaGe* workflow 30

3.2.2 Structure of the master model in SBGN 32

3.2.3 Tree of model relationships 32

3.2.4 Osmotic shock experiments 34

3.3.1 HOG example network 37

3.3.2 rxncon screenshot 41

3.3.3 Complete MAP kinase model 43

3.3.4 Iterative model building 44

3.3.5 Simulation of Boolean MAP kinase model 46

4.1.1 Development of germ layers 51

4.1.2 Epigenetic landscape 52

4.2.1 Reprogramming experiment 57

4.2.2 Clusters of timecourse 58

4.2.3 SPIA analysis 60

4.2.4 Focal adhesion pathway 62

4.2.5 Networks of reprogramming 64

4.2.6 NCA transcription factor activation 65
4.2.7 TFA network 67
4.2.8 NCA connectivity strengths 68
4.2.9 Reprogramming model structure 69
4.2.10Epigenetic mechanisms 71
4.2.11Landscape of differentiation 77
4.2.12Landscape of reprogramming 78
4.2.13Statespace of reprogramming 81
4.2.14Reprogramming efficiency 84
5.1.1 Influenza infection 93
5.2.1 Experiments and phenotyping 95
5.2.2 Viral proteins after infection 97
5.2.3 Clustered proteins 98
5.2.4 Glycolysis proteins 100
5.2.5 Influenza A infection KEGG pathways 101
5.2.6 Infection cycle proteins 104
A.2.1 Candidates 1 129
A.2.2 Candidates 2 130
A.2.3 Complete MAPK network 135
B.0.1 Fold changes in reprogramming 137
B.0.2 NCA TFA heatmap 138
C.0.1 ATP synthetase subunits 152
C.0.2 TCA cycle proteins 153
C.0.3 Host virus interaction network 154

Listing of tables

3.2.1 Candidate models ranked by the AIC value 33
4.2.1 Top pathways from SPIA analysis 61
5.1.1 Abundance of viral proteins 92
A.2.1 Changes in the MAPK model 131
A.2.2 Additional Reactions 132
A.2.3 Minimal HOG model formulas 133
B.0.1 SPIA results 24h 139
B.0.2 SPIA results 48h 140
B.0.3 SPIA results 72h 141
B.0.4 Enriched GO terms cluster 1 142
B.0.5 Enriched GO terms cluster 2 143
B.0.6 Enriched GO terms cluster 3 143
B.0.7 Enriched GO terms cluster 4 144
B.0.8 Enriched GO terms cluster 5 144
B.0.9 Enriched GO terms cluster 6 144
B.0.10Enriched GO terms cluster 7 145
B.0.11Enriched GO terms cluster 8 146
B.0.12Enriched GO terms cluster 9 146
B.0.13Model description 147
B.0.14Experimental findings from literature compared to simulation results. . . 148
B.0.14Experimental findings from literature compared to simulation results. . . 149
C.0.1 Enriched GO terms cluster 1 155
C.0.2 Enriched GO terms cluster 2 155

C.0.2 Enriched GO terms cluster 2 156
C.0.3 Enriched GO terms cluster 3 157
C.0.4 Enriched GO terms cluster 4 157
C.0.5 Enriched GO terms cluster 5 158
C.0.6 Enriched GO terms cluster 6 158

Acknowledgments

First of all, I would like to thank my supervisor **Edda Klipp** for the great support, the freedom to do independent research, and the opportunity to meet many researchers around the world. Without her this work would not have been possible.

I would also like to thank all my collaborators who laid the basis for this work. Especially all the people whose experimental work enabled me to feed my algorithms and models some real numbers. **Ying Wang** performed the reprogramming experiments and produced the timecourse data and **Nancy Mah** provided me with a basic analysis of the microarray data.

A special thanks goes to **Susann Kummer** and **Björn Schwanhäusser**, who performed the infection experiments and the MS measurements and invited me to join an interesting excursion to virology and proteomics.

On the theoretical side, I am especially grateful to **Jörg Schaber**, who brought me to the field of Systems Biology and shared his vision of reducing models to the max with *ModelMaGe*. I am also thankful for the efficient way **Marcus Krantz** pulled me aboard the *rxncon* ship and to **Falko Krause** for showing me all the Python and Javascript magic to sail it. I would also like to thank **Till Scharp** who joined me in in the development of the reprogramming model and the struggle to distinguish between mice and men.

I am grateful to the geeks 505, **Katharina Albers**, **Timo Lubitz**, and **Marvin Schulz**, for proofreading hideous drafts of this work and all the fun we had. I am also deeply indebted to **Thomas Spießer** (also for proofreading) and **Christian Diener** for the countless discussions in our group meeting after hours, in coffee breaks, and on japanese beaches. Many thanks to the rest of the **TBP Group**, I really enjoyed working with all of you!

Last but not least, I want to thank my wife **Ricarda** for she certainly was the greatest support I could ever wish for while working on this book.

For my boys.

Science is what we understand well enough to explain to a computer. Art is everything else we do.

Donald Knuth

1 Introduction

1.1 Outline

The large variety of regulatory processes in modern Cell Biology and their high degree of connectivity poses a lot of interesting and complex questions. Regulation happens on numerous levels each of which includes plenty of mechanisms and acts on different scales. Nevertheless all these processes are interconnected and only in combination lead the desired behavior of the system as a whole. The physical entities that play a role in these regulations are proteins, RNA, DNA, or other types of molecules. All of these can play a part in any of the different regulatory layers, be it gene expression, protein translation, or epigenetic modifications.

Mass production of data on various biological systems has changed the way in which biology as a science is done over the last decade. Stringent analysis, combined with mathematical modeling is required to utilize the full potential of this data. This thesis demonstrates the integration of large scale data into the research process from a Bioinformatics and Systems Biology perspective. We present multiple studies of systems reacting and adapting to environmental stimuli. The reaction was measured with high-throughput methods on the protein and mRNA level. By utilizing Bioinformatics analysis and mathematical modeling, we try to isolate the most important parts of the different mechanisms and propose interactions between them. In the course of this endeavor we developed tools that assist us in the construction of mathematical models and facilitate an efficient approach to Computational Systems Biology.

Regulation of the behavior and dynamics of a system is generally classified into a number of different levels. These are grouped by scale (in time and space) or by chemical properties (stable modifications vs. non-covalent-interactions). To find similarities, differences, and connections between the different levels, we span a large bandwidth of systems and regulatory layers. In detail, we will cover four different levels and three species:

- **Species interactions:** Influenza A virus / domestic dog (*Canis familiaris*)
- **Cell signaling:** baker's yeast (*Saccharomyces cerevisiae*)
- **Gene regulatory networks:** human (*Homo sapiens*) stem cells
- **Epigenetic regulation:** human (*Homo sapiens*) stem cells

These different systems can all be analyzed using similar theoretical methods and their behavior is governed by universal basic biological mechanisms and complex systems. This exemplifies the generality of theoretical methods — and modeling in particular — and their potential to unify different areas of biological research by highlighting common principles and standardizing a common language.

Systems Biology is a strongly interdisciplinary field. It requires detailed knowledge of a wide array of topics, spanning from mathematical theory to biological methods and background. As a result we will introduce many different topics to the reader. The mathematical analysis and experimental methods used in the different studies, are explained in chapter 2. We show the results of the different research topics in chapters 3,4 and 5.

In chapter 3 we present the results of software development efforts and techniques for the generation, management, and discrimination of mathematical models. We demonstrate the use of two resulting software applications on different models of mitogen-activated protein (MAP) kinase pathways in baker's yeast.

Chapter 4 is the main part of the thesis, and the focus of my work. It examines the reprogramming of somatic body cells to pluripotent stem cells by viral transduction of exogenous factors. This chapter consists of section 4.2.1, which is an in depth analysis of mRNA expression changes in a reprogramming experiment and section 4.2.2 which presents a mathematical model that describes an hypothesis about the mechanisms regulating reprogramming of human somatic cells to a pluripotent state.

Chapter 5 deals with another infection process, but focuses on the interaction between the Influenza A virus and its host. As we will see, the analysis methods used in this chapter partly overlap with chapter 4, although they deal with protein rather than mRNA data.

Because of the large biological variety, we chose to give a biological introduction at the beginning of each chapter, while the unifying theoretical background is given in the remainder of this chapter. As last and concluding part, chapter 6 gives an overview of the results of the different projects and discusses the conclusions that can be drawn from these. We present new approaches for software assisted model construction, insights on the somatic reprogramming process, and new findings on the virus-host-interactions during influenza infection. The chapter will also give an outlook to further research efforts that might result from the presented work.

1.2 Systems Biology

In the 20th century, biology has seen an explosion of knowledge of unprecedented proportions. From medical advances like antibiotics, over the discovery of the structure of the DNA (Watson and Crick, 1953), to the full sequence of the human genome (Venter et al., 2001), it was a success story throughout. Many important discoveries in the last decades were based on the detailed study of the properties and functions of certain genes and proteins. Researchers took a "reductionist" approach to understand life, by reducing it to single pieces and their properties. These studies produced an enormous amount of knowledge about the details of cellular components and processes. The fields of genetics and genomics were rapidly developing and genes were seen as the most important building blocks that determine the fate of an organism. These billions of parts are an invaluable resource for modern biology, and the task is now to put the pieces back together again.

Systems Biology is a new school of thought for the Biology of the 21st century. It shifts the focus from the reductionist view of the biological system towards a wider view that is more holistic. From this wide angle one tries to gain an overall abstract understanding to be able to identify the important details of the system to include into further studies. Systems Biology ideally connects all levels of regulation of the system to understand its behavior. These connections between the levels are also not unidirectional, from the direction of genes to the complete organism, which is implied by the classic dogma of molecular biology (from DNA to mRNA to protein) (Crick, 1970). But can also work in the opposite direction, in which proteins influence genes and their expression. The reprogramming of body cells by transfer edansfer, in which the nucleus from a somatic cell is transfered into an oocyte, is a good example of these mechanisms. In this process proteins that are present in the moment of nuclear transfer decide which genes are activated and which are not. This results in the reprogramming of the whole cell to a stem cell, illustrating that proteins influence the expression of genes and also chromatin structure.

The Systems Biology approach utilizes a combination of mathematical, computational and experimental methods to get close to the goal of understanding why the whole system is built as it is, and how its different levels are interconnected. This approach is not an entirely new idea, and has been practiced by pioneers of the field already in the 1960s (e.g. Kell, 1979; Noble, 1960). Nevertheless the approach has become more popular only in the last decade (Kitano, 2002). This development is mainly due to the following reasons: (i) The idea that the function of biological systems can be understood by mere intuition after looking at data has become increasingly unrealistic given the complexity of contemporary biology. (ii) Biology has become a much more data driven science than it used to be, by the advances in bioinformatics algorithms, data analysis, and the establishment of large databases storing the huge amounts of data that are gathered by modern "omics" techniques like deep sequencing and proteomics. (iii) Computers are still becoming more powerful very quickly and more specialized tools are becoming available for systems analysis by software developers from biological labs. Combined with the easy exchange of information over the internet, this makes computational approaches a powerful tool for biology to help unravel the complex systems that need to be investigated.

1.3 Mathematical Modeling of Biological Systems

Mathematical models of the studied system are an integral part of studies in Systems Biology. A model formulates the biological phenomena in a system using mathematical language and techniques. If used well, models can be a great help in structuring information and predict experimental outcomes for a system and propose new experiments. Models can come in a variety of flavors, e.g. they can describe the flow of metabolites through an organism or a cell, or predict which genes are active in an organism under certain circumstances. A model in general is an abstraction of reality that highlights certain aspects of a system and cannot aim to explain it completely. The superior aim in model building should be to elucidate the key features of the processes under observation. A well known and fitting quote in this context was phrased by Box, 1976:

> "Essentially, all models are wrong, but some are useful."

Models are often used in an iterative cycle with experiments to state precise hypotheses about a system and refine the understanding of the processes that govern it (Figure 1.3.1). Ideally the process starts off with an idea or hypothesis about a specific biological question that leads to a model formulation based on the current knowledge. The model can then be simulated to predict outcomes of new experiments. This model output can be compared to experimental data that is used to readjust and improve the model, which can then start the cycle anew. The data is ideally gathered in a way that it can clearly validate or reject the model predictions.

To researchers new to the field, the merits of theoretical models are not always obvious at first sight. Nevertheless, working with models is often necessary and can prove useful in practice. The most evident use is probably prediction of system behavior under altered conditions, as described above. Another advantage of building mathematical models of the studied system is inherent in the process of their creation. Modeling forces researchers to formulate precise verbal hypotheses to enable the dialog between the experimentalists and theoreticians already during the early phases of a study and thereby helps to specify clear aims and questions. This often sheds light on crucial gaps in understanding of the system and in many cases inspires new experiments to close these gaps.

1.3.1 Scope of a Model

A model represents the current knowledge about a system in an abstract and usable format (Kitano, 2002). Therefore, to be useful, the model has to be formulated in a way that is adapted to the level of detail of said knowledge.

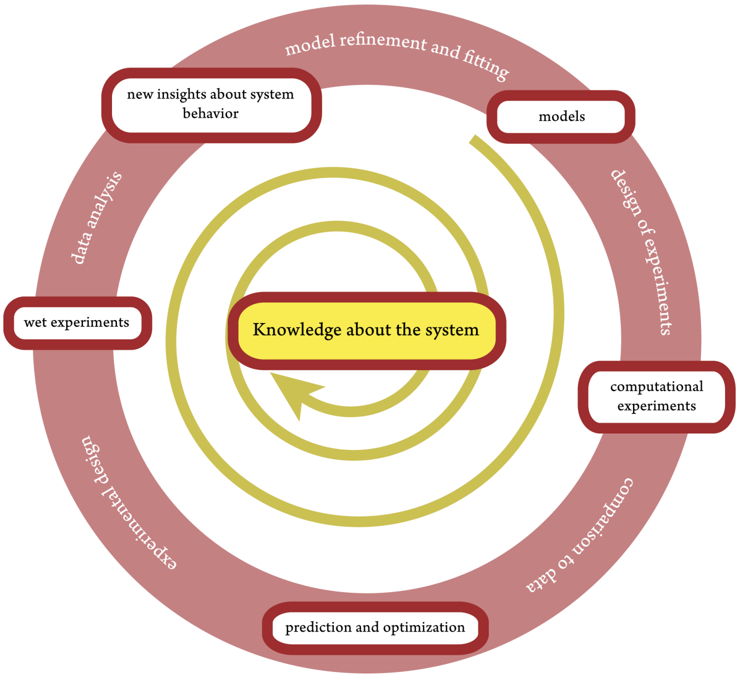

model refinement and fitting
new insights about system behavior
models
data analysis
design of experiments
wet experiments
Knowledge about the system
computational experiments
experimental design
comparison to data
prediction and optimization

Figure 1.3.1: Systems Biology is often associated with iteration between biological experiments and theoretical analysis. The cycle is a strongly idealized view of systems biological research and can contain inner loops or feed-backs in reality.

There is a variety of mathematical frameworks that have been used in the past to describe biological systems and the field is still rapidly evolving. Choosing an appropriate modeling framework is a crucial step in a Systems Biology project, because this defines the scope of the model and sets limits to the development. For this step the hypothesis that shall be tested needs to be formalized and one needs to define clear goals for the study while considering available data and possible experiments. There is always a trade-off between granularity and detail of a model versus its complexity and the amount of data that is needed to test hypothesis about its behavior. Each modeling framework has its advantages and disadvantages and they all highlight different aspects of a system. Figure 1.3.2 summarizes some of the most common methods and their level of detail to visualize this aspect. Models presented in this thesis use two common forms of dynamic modeling that lie in opposite parts of this spectrum. First there are ordinary differential equation (ODE) systems which are the main focus of section 3.2 that allow for a continuous dynamic simulation of biological systems and for which there is a large amount of analysis tools available. These models often have a large number of parameters that need to be set by using experimental data. Second we concentrate on Boolean models (Section 3.3) that are the most abstract form of dynamic modeling and are suited for large systems (e.g. Figure 3.3.3) but are only coarse grained approximations of reality. Boolean models have been extended by stochastic frameworks to include uncertainties in the data into models, which we use in section 4.2.2. There are many more modeling frameworks that do not fit into the scheme (agent based models, game theory models, etc.) presented in figure 1.3.2, and it is always a difficult decision which of the many approaches fits a system best.

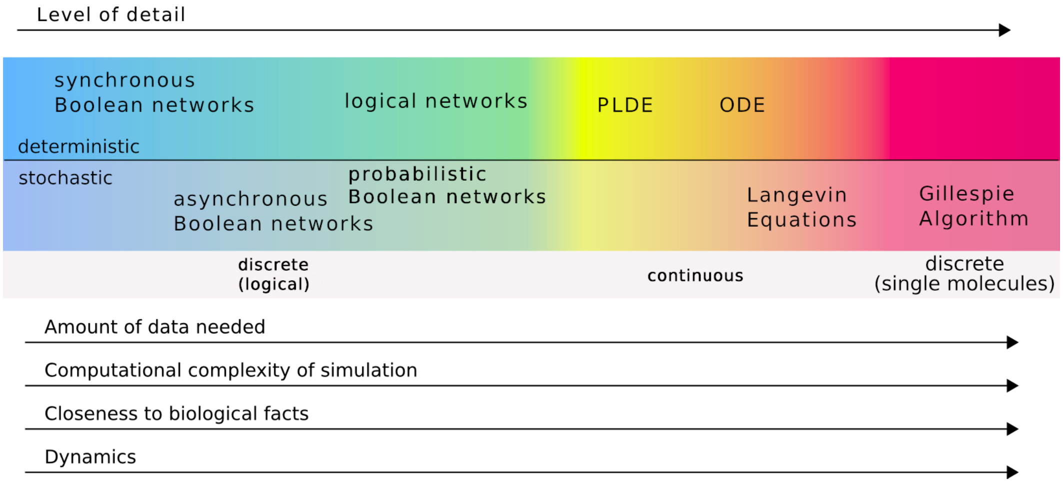

Level of detail
synchronous Boolean networks
logical networks
PLDE
ODE
deterministic
stochastic
asynchronous Boolean networks
probabilistic Boolean networks
Langevin Equations
Gillespie Algorithm
discrete (logical)
continuous
discrete (single molecules)
Amount of data needed
Computational complexity of simulation
Closeness to biological facts
Dynamics

Figure 1.3.2: A brief overview of different modelling methods and their degree of detail. Approaches are ordered by their level of detail horizontally and are divided vertically into stochastic and deterministic models. (PLDE = partially linear differential equations)

1.3.2 Data Driven Modeling

To build a mathematical model it is necessary to define relationships between the building blocks of a system (e.g. proteins and genes). This is frequently done using high-troughput association data like chromatin immunoprecipitation on microarrays (ChIP-on-chip) for transcription factor binding or affinity purification and mass spectrometry for protein interactions. Another way to create such networks is by text-mining the literature on the topic following expert curation. These two approaches are often combined and the resulting networks are stored in public databases like KEGG (Kanehisa, 2000) or Reactome (Joshi-Tope and Gillespie, 2005).

These networks become useful for modeling when annotated with Gene Ontology (GO) (Ashburner et al., 2000) and existing pathway data to see interconnections and crosstalks between systems. When connected to (ideally dynamic) expression data, these networks become testable for feasibility of the connections given the changes in expression and it is possible to estimate the strengths of regulation inside the network (Section 2.2.4). Visualization of the data on a network can also facilitate communication and interpretation.

Data analysis and annotation are needed to find out which parts of a system are modulated by a certain stimulus and which are the most pressing ones to be elucidated by modeling. We used these techniques to analyze different experiments in chapters 4 and 5.

After these first steps of defining a network structure from the data, the next task is to define the dynamics of the topological system. Once there is enough information about a system to formulate a detailed kinetic model, the challenge is to adjust the free parameters to make the model behave like the data dictates. Ultimately, the best model has to be chosen from a number of parameterizations or even structural alternatives.

On all of the previously mentioned levels there is a lot of bioinformatic and statistical data processing needed, especially when working with high-throughput data. Therefore, software to facilitate and standardize these processes is severely needed.

We should continually be striving to transform every *art into a science: in the process, we advance the art.*

Donald Knuth

2 Methods

The following chapter will describes the techniques used in this thesis in detail. It briefly explains the experimental techniques that have been used to gather the data that was analyzed. Although the experiments were not done in the frame of this thesis, their understanding is crucial for the relevance of the presented results and is therefore included.

Afterwards we give details of techniques that enable researchers to make use of the wealth of data that has become available in 21st century biology. The used techniques largely aim to extract hidden features in given data and relate them to previously known biological facts. The last section of the methods chapter deals with modeling techniques that were used in different parts of the work.

2.1 Biological Methods

The presented methods are all highly dependent on bioinformatic and statistical analysis, which we explain in later sections (Section 2.2). Results of these analysis also have to be carefully interpreted to gain insights into the biological meaning, where modeling, as described in section 2.3, plays a major role.

2.1.1 Mass Spectrometry

Mass spectrometry (MS) is a very efficient technique to identify proteins in a complex mixture. It utilizes the differences in mass and charge of peptides to identify them by comparison to known datasets.

The technique can be used to detect thousands of proteins simultaneously, but requires some preparation of the probe in advance. Mass spectrometers work best in the mass range of peptides up to 20 amino acids, therefore proteins need to be digested by proteases before the detection. This generates an even higher complexity in the probe, because proteins are cut into a possibly large number of peptides. In order to reach a high resolution in the measurement, one needs to separate the peptides by a form of chromatography. A common technique is to couple the mass spectrometer with high performance liquid chromatography (HPLC). An especially powerful method for proteomics is the reversed phase HPLC, which separates the peptides by their hydrophobicity. Chromatography adds a time dimension to the experiment, as the peptides arrive at the spectrometer at the time they leave the HPLC column. The spectrometer then detects the mass by charge (m/z) profiles measured at any given time point (Figure 2.1.1).

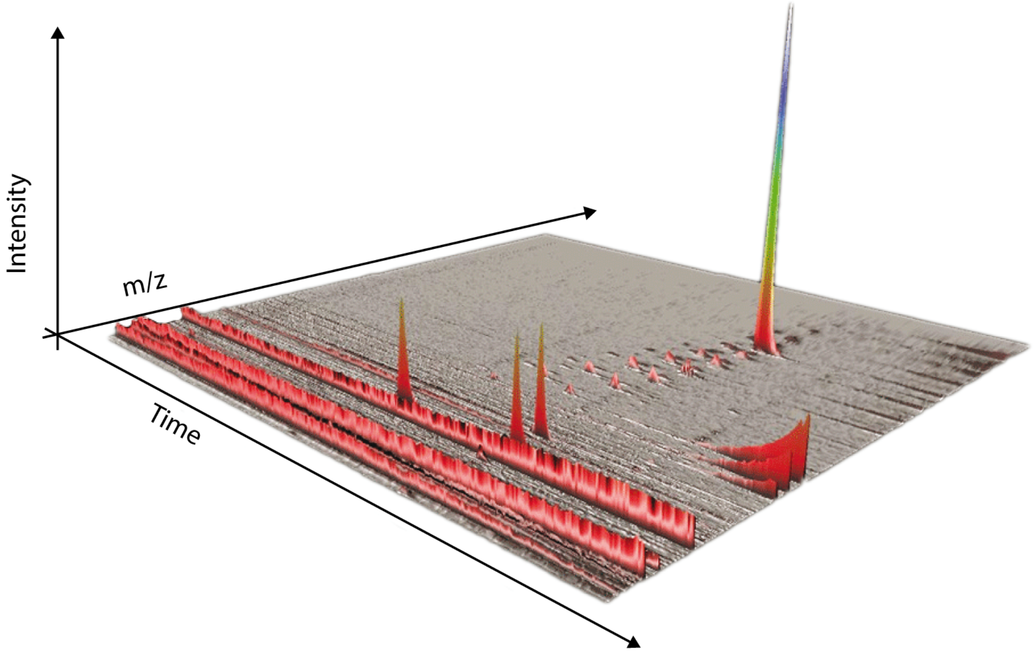


Figure 2.1.1: HPLC-MS/MS experiments generate a huge amount of data that is distributed over 3 dimensions: time, m/z, and intensity. The data measured in the m/z and intensity dimensions as a function of time is used to identify peptides that appear in each time point.

The setup of mass spectrometers can differ substantially, but by definition all of these machines consist of three parts: an ion source, a mass analyzer and an ion detector. To be able to detect the peptides in the MS, they need to be ionized by an ionization technique like for example electrospray ionization (ESI). Ionization adds charges to the peptides depending on their amino acid sequence, because amino acids differ in their susceptibility to ionization. The charged ions can enter the instrument via a capillary. The mass to charge ratio of the ions as well as the overall intensity is then measured by different detection systems depending on the setup. The total ion count (TIC) is calculated by the sum of intensities over the whole mass range at one time point. Each time point in the TIC holds information about the specific m/z ratios of the ions that enter the machine at that time. Unfortunately peptide mass profiles are not enough to deduce the present proteins with a high degree of certainty, because there can be many peptides with the same mass but different sequence. Therefore sequence information is needed to discern between these peptides. This information can be obtained by using tandem MS (MS/MS) techniques, in which the peptides are once more broken apart by collision with gas molecules after the first detection round. The m/z values of the random ion fragments are measured again and from their specific values the exact amino acid sequence can be deduced.

In this step bioinformatic techniques are indispensable to make use of the gathered data. This analysis has two levels. First there is the basic level that maps the fragment profiles to peptide sequences and subsequently identifies proteins with a high amount of certainty. As this step is required in each MS experiment, there is sophisticated software available to tackle this problem such as Maxquant or OpenMS (Cox and Mann, 2008; Sturm et al., 2008). The second step includes functional analysis and detailed analysis of the identified proteins by some of the techniques we used throughout this thesis (e.g. Section 2.2.2).

Stable Isotope Labeling by Amino Acids in Cell Culture

Measuring the complete expressed set of proteins in a sample is possible since Mass Spectrometry methods have become more common for mixed samples. The field of proteomics has seen a similar growth as genomics in the last decades and identification of proteins has become a straight forward process. A major drawback of mass spectrometry has always been a lack of quantitative data that could be used for modeling approaches. The peptide signal intensity is not directly proportional to the amount of peptide in the probe, due to a multitude of errors introduced by the many processing steps in MS experiments (Ong and Mann, 2005). Nevertheless, these errors are systematic, which enables a relative quantification of proteins between experiments or different probes within one experiment. In recent years several methods were established applying the principle of heavy isotope labeling of proteins. Stable isotope labeling introduces isotopes differing with different mass into chemically equivalent peptides that can be measured given a precise instrument.

Stable isotope labeling by amino acids in cell culture (SILAC) (Ong, 2002) applies this principle by cultivating at least two cell populations that differ only in the media they are grown on. These media contain either light, medium, or heavy amino acids that are synthesized using different isotopes (Figure 2.1.2). In a SILAC experiment, cells are grown on these media for several doubling rounds to replace even proteins with low turnover rates completely with their medium or heavy counterparts. After a sufficient cultivation time, perturbations can be done to all the cultures. To measure the dynamics of changes in

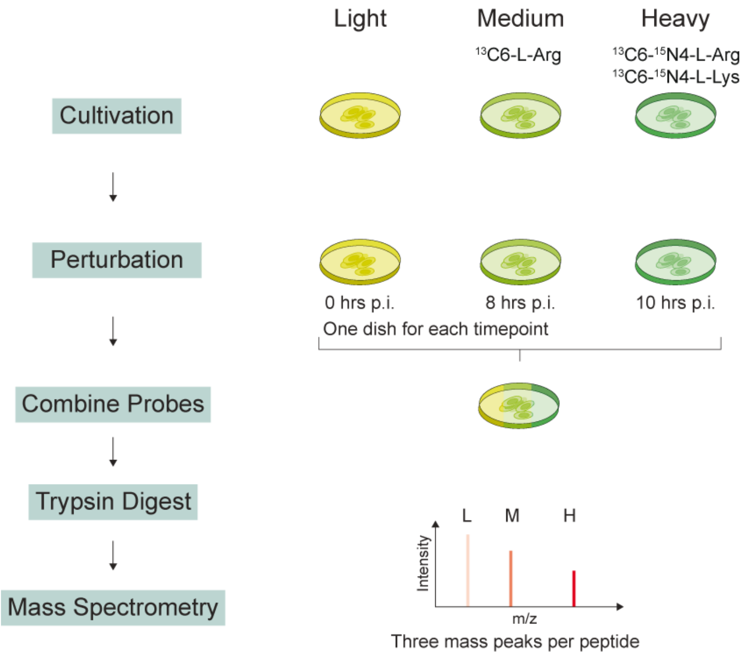


Figure 2.1.2: Stable isotopic labeling of amino acids in cell culture (SILAC). This quantitative MS technique works by growing cells in media containing isopopicaly marked amino acids, to be able to distinguish different cell lines or time points in one MS experiment.

the proteome, the effect of the perturbation can be measured by taking samples from each probe at a specific time point. These different samples can then be combined and processed (lysed, fractionated, and purified) together, which reduces the possibility of errors to a minimum. To be able to cover more than three time points at once, one can combine measurements by defining a common time point for all experiments and normalizing to it.

Intensity Based Absolute Quantification

Opposed to the SILAC approach, there are also new label free proteomics approaches. The intensity based absolute quantification (IBAQ) (Schwanhäusser et al., 2011) uses the absolute intensities and the observation that more abundant proteins are also more likely to be detected in shotgun experiments.

2.1.2 Expression Profiling by Microarrays

Gene expression analysis (profiling) is the determination of the pattern of genes expressed at the level of genetic transcription, under specific circumstances, or in a specific cell. This highly valuable information can be gathered by different techniques (e.g. RNA sequencing), but the most common method are still microarrays.

A microarray works by DNA-hybridization of a nucleic acid sample (target) to a large set of oligonucleotide probes, which are attached to a solid support surface, to detect variations in a gene sequence or in this case mRNA expression. The array is separated into tiny spots or beads, depending on the actual implementation, each filled with a lot of oligonucleotides (about 50 bases long) that are specific for a certain mRNA in the target cells transcriptome. Thereby the amount of cDNA binding to each spot is a measure for how much of the mRNA was present in the target.

DNA microarray techniques are very diverse, but all are based on the same hybridization principles. The target cells are prepared according to the array specifications typically including RNA extraction, purification, and digestion (Figure 2.1.3). The next step is to produce cDNA from the RNA probes via reverse transcriptase. The cDNA is then labeled with a fluorescent marker (e.g. cy3 or cy5) and then hybridized to the microarray. At this point lies a major difference between the specific arrays that are available. First, there are so called two channel arrays, that can compare two differently labeled target cell types on one chip by measuring the staining by two fluorophores at different wavelengths. And second, there are single channel arrays that only measure one expression profile at a time. As the arrays used for data generation for this thesis were purely single channel we will concentrate on this variant in the following.

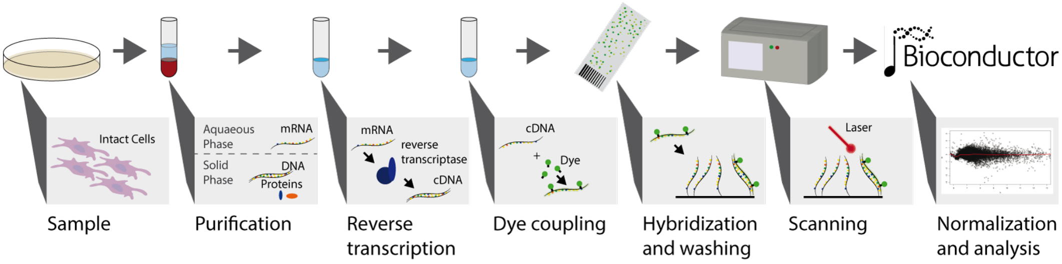


Figure 2.1.3: A typical single channel microarray experiment. RNA is purified, reverse transcribed to cDNA which is then labeled and hybridized to the chip. Each target is prepared in the same way but hybridized to a separate array. The arrays are then compared in the following bioinformatic analysis.

After hybridization the chip is washed to remove non hybridized cDNA, and then scanned to evaluate the amount of fluorescence in each spot. However, the brightness does not truly indicate abundance levels of an mRNA. Each mRNA molecule encounters protocol and batch-specific bias during amplification, labeling, and hybridization phases of the experiment rendering comparisons between genes for the same microarray uninformative. The comparison of transcriptomes between different conditions is the major strength of the approach, but it requires one array per condition. The advantage of single dye systems is the easier comparison between arrays of the same type, as they are all done in the same way.

Bioinformatic Analysis

The large amount of measurements possible on an array, and the varying precision that is influenced by many factors, make the statistical analysis of array data quite challenging.

The first of many factors that influence the outcome is the experimental techniques and bias that is introduced by the factors mentioned above. Second there are further data processing steps, like scanner sensitivity and image processing. A very important point in the analysis is the normalization of data and background correction. Data usually needs to be normalized between the single channel arrays to be able to effectively compare the different conditions. One problem for example is that differences between targets often scale with the absolute intensity in the spots. Normalization is usually done by assuming that the majority of genes did not change between the targets and using a LOESS (Cleveland, 1979) normalization to make the differences comparable between genes.

The next important step is to filter out the genes of interest, by testing whether the differences in a gene in the two arrays are significant or not. This can be done using well suited statistical tests (t-tests, empirical Bayesian methods) which take into account the large number of tests performed when calculating p-values. For all these steps there is a large amount of software available, either provided by the manufacturers of the array or freely available in the statistical programming language *R*. Most of these tools are bundled into the software collection Bioconductor (Gentleman et al., 2004).

2.2 Functional Analysis

2.2.1 Clustering of Timecourse Data

Clustering of biological data is often done to find hidden structures in large datasets. There are basically two different types of clustering that can be used to find similarities to group parts of a dataset together. (i) Hierarchical clustering methods find an hierarchical order that defines a tree of increasingly similar data points defined by an arbitrary distance metric (e.g. euclidean distance). We use such a method in the heatmap visualization to show the relationship between the displayed rows and columns (Section 2.2.3), but they did not play a major role in the presented work. In each step, the iterative clustering approach agglomerates the closest data points or clusters into a new cluster, thereby generating a tree of clusters (for details please see: Hastie, Tibshirani, and Friedman, 2009). (ii) Partitioning clustering separates the data into non-overlapping classes that form around cluster centers. These centers are defined by the data points around them. The most frequently used algorithm for partition clustering is the *k-means* algorithm (MacQueen, 1967). *k-means* is a variant of the general expectation-maximization-algorithm (EM). The algorithm uses a given number of k cluster centers which are placed randomly in the dataset. Then it computes the closest cluster center $\mathbf{c}_j$ for each of the N data points $\mathbf{x}_i$ (expectation step). In the next step the positions of the cluster centers are recalculated as the mean of the corresponding data points (maximization step). Afterwards the data vectors are reassigned to the new cluster centers and the next iteration cycle starts. The algorithm terminates when no data points change the assigned centers in one iteration. In general, the algorithm tries to minimize the function

$$E = \sum_{i}^{N} \sum_{j}^{k} ||\mathbf{x}_i - \mathbf{c}_j||^2, \tag{2.1}$$

while its outcome strongly depends on the randomly chosen start sites. Therefore the algorithm will only find local minima of the objective function E that is a measure for the within-cluster variation.

An often encountered problem in *k-means* clustering is the fact that outliers and noise strongly affect the clustering process. There is no measurement for the degree of membership to a cluster, which could be used to weaken the influence of outlier data points. This is why we chose a *fuzzy* clustering method, with the optimization algorithm *fuzzy-c-means* (FCM) (Dunn, 1973). This method is very robust against noise and additionally returns a membership value for each data point to each cluster center. These values are contained in the $N \times k$ partitioning matrix $\mathbf{U}$. The objective function for the optimization is given by

$$J = \sum_{i}^{N} \sum_{j}^{k} (u_{ij})^m ||\mathbf{x}_i - \mathbf{c}_j||^2, \tag{2.2}$$

where parameter $m > 1$ defines the sharpness of the clustering, i.e. how close the fuzzy clustering is to hard partitioning. So for very large m the level of influence of each point to each center becomes equally large close to $\frac{1}{k}$ and for $m \rightarrow 1$ the values of u_{ij} are closer to 1 or 0, which renders the FCM equivalent to *k-means* clustering. Optimization is done with the following two constraints:

1. For each point $\mathbf{x}_i$ the degree of membership to all clusters sums up to one: $\sum_{j=1}^{k} u_{ij} \; \forall \; i = \{1, ..., N\}$.

2. All clusters are non empty: $\sum_{i=1}^{N} u_{ij} > 0 \; \forall \; j = \{1, ...k\}$.

The algorithm works similar to *k-means* in an EM-like fashion alternating between the expectation step setting the partition matrix:

$$u_{ij} = \frac{1}{\left(\frac{||\mathbf{x}_i - \mathbf{c}_j||}{\sum_{l=1}^{k} ||\mathbf{x}_i - \mathbf{c}_l||} \right)^{\frac{2}{m-1}}} \; \forall \; i = \{1, .., N\}, j = \{1, ..., k\}, \tag{2.3}$$

and the maximization step setting the cluster centers:

$$k_j = \frac{\sum_{i=1}^{N} (u_{ij})^m \mathbf{x}_i}{\sum_{i=1}^{N} (u_{ij})^m} \; \forall \; j = \{1, ..., k\}. \tag{2.4}$$

These two equations are based on the first order conditions for a minimum of the Lagrange function. *Fuzzy-c-means* terminates, if the change in the partitioning matrix $||\mathbf{U}^s - \mathbf{U}^{s-1}||$ in a step s is below a threshold ε .

FCM clustering has favorable features for clustering of noisy data mainly because of two reasons:

1. The influence of outliers on cluster-centers is drastically reduced by choosing the right m values and cluster artefacts can be reduced.

2. It provides a way to do a posterior filtering of data by their membership values, instead of *a priori* filtering of datasets (e.g. by their $\log_2$ fold changes from expression profiling). The reasoning here comes from a systems perspective: If a whole cluster exists that has similar dynamics, the measurement can be related and has a higher probability to be measured correctly. Noisy vectors can be filtered after the clustering by their generally low membership values and large distances to the cluster centers.

These properties make the method well suited for the datasets we use in this thesis. The datasets we used are microarray and SILAC timecourse data, that represent the fold changes for each mRNA/protein in every time point. Especially for the proteome data (Section 2.1.1) the posterior filtering is an important feature, because the dataset shows very low fold changes in general (Chapter 5).

For the clustering we standardized the fold-changes of the time-course to mean zero and standard deviation of one, to make clustering in euclidean space possible. An FCM implementation from the Mfuzz package (Futschik and Carlisle, 2005) in Bioconductor (Gentleman et al., 2004) was used. By applying an iterative approach the number of cluster centers and a value for m was determined that gives an optimal separation.

2.2.2 Enrichment in Functional Databases

Assigning functions to the mass of biological entities found in high-throughput experiments is an important step in the analysis of datasets, because it enables researchers to interpret the data in the context to previous knowledge. There are different methodologies that either aim to find modules of genes that share a functional annotation and are similarly regulated or to find pathways that are influenced by changes in a system. We used both of these techniques to characterize the datasets in chapters 4 and 5.

Identified genes and proteins were assigned to their biological process using gene ontology (GO) (Ashburner et al., 2000) enrichment analysis. The GO database consists of three major annotation parts: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). Each annotated gene can have multiple entries in each of these classes. The classes are structured like a tree with very detailed annotations at its branches and more general annotations at its root. Given a set of genes, one can now determine, using a hypergeometric statistical test, whether certain annotations appear significantly more often in this set than expected from their distribution in the background of the whole annotation tree. The annotations that are significantly enriched are likely to describe an important part of the process. For the enrichment analysis shown in this thesis, we always used the smallest possible set of genes as a background distribution (universe), because an unnecessary large universe would lead to an overestimation of p-values for enrichments.

Enrichment for biological process (BP) was tested using a hypergeometric test from the GoStats (Falcon and Gentleman, 2007) package in Bioconductor for single clusters with the complete set of measured genes as a background distribution. This test automatically corrects for the bias resulting from the tree structure of the ontology. The analysis was performed for the KEGG (Kanehisa, 2000) pathway annotations in a similar fashion.

For the stem cell expression data (Chapter 4.2) we also performed a more sophisticated testing procedure that includes a Systems Biology based approach. This method is called signaling pathway impact analysis (SPIA). It recognizes the influence of a regulated gene on a pathway, when calculating if the pathway is influenced by the given perturbation (Tarca et al., 2009). SPIA uses very simple models of the pathways in the KEGG database that take the topology of networks into account and consider how many downstream targets are affected by each perturbation. Using these models it calculates how strong the regulation measured in the dataset affects the general behavior of each pathway.

In combination with a classical enrichment test, these values enhance the sensitivity and specificity of the tests and give a two dimensional picture of the impact on a certain pathway.

2.2.3 Proteomic Phenotyping

Based on the enrichment analysis described in the previous section, I performed a visualization techniques for changes in the proteome.

To characterize the proteomic changes provoked by a perturbation, here influenza A infection (Chapter 5), a proteomic phenotyping for GO terms was performed as previously described (Pan et al., 2009). This technique divides a skewed distribution of N measured $\log_2$ fold changes into an arbitrary number of M quantiles and does an enrichment test for each of the quantiles separately with all detected proteins as background. This leads to an $M \times N_e$ matrix $\mathbf{P}$ of p-values, where N_e is the number of GO terms that were enriched in one of the quantiles with a p-value <0.05. We did a transformation by $\mathbf{X} = -log_{10}(\mathbf{P})$ and computed a z-score by

$$\mathbf{Z} = \frac{(x_{ij} - \mu(\mathbf{x}_i))}{\sigma(\mathbf{x}_i)}, \tag{2.5}$$

where $\mu(\mathbf{x}_i)$ is the mean and $\sigma(\mathbf{x}_i)$ is the standard deviation of the GO term p-value vector $mathbfx_i$ for all of the quantiles j. This matrix is visualized in a heatmap showing the relative enrichment in each of the quantiles clustered by the z-scores.

2.2.4 Network Component Analysis

Connecting different types of data to gain mechanistic insight into a system is a very important part of Systems Biology, because it often helps the development of more precise models than using just one type of data. Combining network structure with time evolution of the network nodes is a good example for this. With the network component analysis (NCA) Liao et al. (2003) developed an elegant method that allows to estimate the strength of connections in a network with only structural information (i.e. from transcription factor binding data) by analyzing the dynamic changes in the network induced by perturbations (expression profiles). NCA computes the activity of transcription factors based on the expression of their targets (Figure 2.2.1). The method is based on matrix decomposition of the input data and optimization of the connection matrix. The method is a reverse engeneering approach to reconstruct a model of the form:

$$\mathbf{E} = \mathbf{AP} \tag{2.6}$$

where $\mathbf{E}$ is a matrix containing the expression data of the regulated genes ($N \times M$, with N time points and M genes). $\mathbf{P}$ is the expression data of the regulatory layer ($N \times L$) with L being the number of regulatory nodes ($L << N$). Matrix $\mathbf{A}$ contains the connectivity strengths between the two layers. Once the system fulfills certain criteria (for details see Liao et al., 2003), this estimation is done by minimizing the following function:

$$||\mathbf{E} - \bar{\mathbf{A}}\bar{\mathbf{P}}||^2, \text{ so that } \mathbf{A} \in \mathbf{Z}_0 \tag{2.7}$$

where $\mathbf{Z}_0$ is the allowed topology of the network. This optimization produces the estimators for the transcription factor activity (TFA) $\bar{\mathbf{P}}$ and the connectivity strengths $\bar{\mathbf{A}}$. We used this method in section 4.2.1 to estimate the effect of stem cell transcription factors on their target genes. This analysis was done using the NCA toolbox in *Matlab* (Kao et al., 2004).

2.3 Dynamic Modeling

2.3.1 Boolean Modeling

From the many approaches of dynamic mathematical modeling a biological system, Boolean modeling is the most simplistic approach there is. It ignores a lot of details but has the

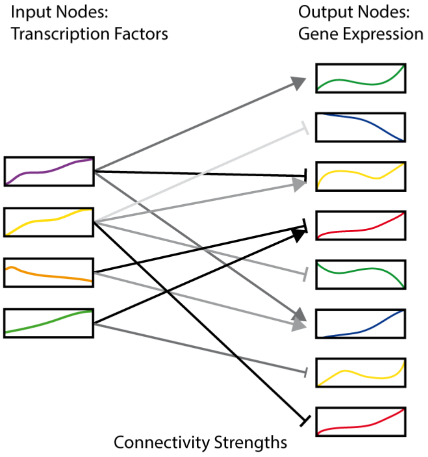


Figure 2.2.1: The connectivity strengths and transcription factor activities are estimated via the known network connectivity and the measurements for the target genes. The method takes advantage of the connectivity to reverse engineer the influences of the inputs on the outputs. This is done by minimizing the difference between the expression values and the product of the connectivity times the inputs.

ability to provide an overview of the main qualitative properties of the modeled system (Section 1.3). Boolean models were first proposed as a tool for modeling gene regulation in 1969 by Kauffmann. Kauffmann proposed the models of N genes with the degree K, which is why they are also called *N-K-models*, to mimic the dynamics of gene regulatory networks (GRNs).

A minimal example for such a network and the resulting dynamics is given in Figure 2.3.1. A Boolean network can be represented as a graph $G(V, F)$, consisting of a set of N nodes $V = \{x_1, ..., x_n\}$ and a set of N edges between the nodes that are defined by the update functions $F = \{f_1, f_2, ..., f_n\}$, which represent the transitional relationships between different time points. For every time point t, each node x_i has a state $x_i(t) \in \{0, 1\}$ denoting either no expression or expression of a gene (or absence or presence of activity of a regulatory property, respectively). A Boolean function $f(x_{j_1(i)}, x_{j_2(i)}, ..., x_{j_{k(i)}(i)})$ with $k(i)$ specified input nodes is assigned to node x_i, where $j_{k(i)}$ represents the mapping between genes at different time points. The state of gene x_i at time point $t+1$ is determined by the values of a set of other genes at time point t using the Boolean function $f_i \in F$. This way, a state transition is defined as:

$$x_i(t+1) = f(x_{j_1(i)}(t), x_{j_2(i)}(t), ..., x_{j_{k(i)}}(t)) \tag{2.8}$$

The state vector or simply the state $\mathbf{S}(t)$ of the network at time t corresponds to the vector of the node states at time t, i. e. $\mathbf{S}(t) = (x_1(t), ..., x_n(t))$. Thus, since every $x_i(t)$ can only take two possible values 0 or 1, the number of all possible states is 2^n.

Boolean models are used more and more to model GRNs and signaling networks in recent literature (Albert and Othmer, 2003; Bauer et al., 2010; Kauffman and Peterson, 2003; Orlando et al., 2008) and there is active development of new techniques.

Attractors and Statespace

The set of states that a Boolean model can possess form another network, the state space network, which has to be strictly distinguished from the Boolean network definition.

The state space network $P(S, T)$ consists of the set of state vectors S and the set of transitions T between the states. Each state has exactly one outgoing transition edge (out-degree), whereas the in-degree (i.e. number of incoming edges) can vary between the states. The state space can be divided into different classes of states:

Transient states States that are only passed once and do not occur again in the same simulation

Leaf states States with an in-degree of zero that can never be reached in a simulation, it they are not the start state

Point attractors States that have a transition to themselves and can not be left after reached once

Cyclic attractors A set of states that form a cycle and are reached periodically during a simulation

Usually the statespace is split into different attractors and the transient states that lead to them. These transient states are sometimes called the *basin of attraction* of an attractor (Figure 2.3.1 B).

Stochasticity in Boolean Models

Classical Boolean models are defined as discrete deterministic systems, which obviously far from biological reality with different timescales and stochastic processes. Therefore there have been many approaches of adapting Boolean models to include these phenomenons (Garg et al., 2009; Twardziok, Siebert, and Heyl, 2010). We will now introduce the methods that were used in this thesis.

Asynchronous Updating

The simulation of time in Boolean modeling strongly depends on the way of updating the nodes. Synchronous updating, affecting all nodes in each time step is the most simple form of Boolean simulation. There are other alternatives for defining the updating asynchronously to make the Boolean framework more flexible. In asynchronous updating only a subset of nodes is updated in each time step, and the different variants differ in the selection criteria of nodes. In this thesis we only used a randomly selected node that is updated, but there is also many other ways to choose. For example one can define a time delay for each nodes' update, to simulate different time scales of processes in a model. Stochastic

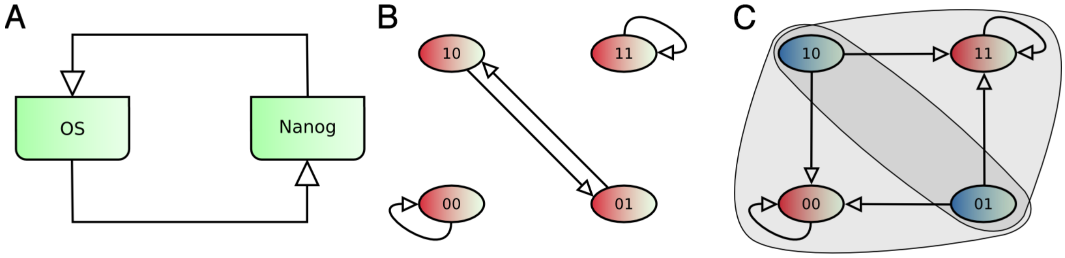


Figure 2.3.1: The basins of attraction under different updating methods. (**A**) A simple model where $OS(t+1) = Nanog(t)$ and $Nanog(t+1) = OS(t)$. (**B**) Synchronous updating leads to two point and one cyclic attractor. (**C**) Asynchronous updating leads to two point attractors that have overlapping basins of attraction.

asynchronous updating strongly alters the statespace of a model, because in this case every state **S** can have as many as N outgoing transitions that can be chosen randomly (Figure 2.3.1 C).

Probabilistic Boolean Models

Apart from the updating scheme, there are other ways to extend the Boolean framework to allow to include uncertainty (Garg et al., 2009; Shmulevich, 2002; Twardziok, Siebert, and Heyl, 2010). In chapter 4.2.2 we used the probabilistic Boolean network (PBN) approach proposed by Shmulevich (2002). Probabilistic Boolean networks were designed to represent the uncertainty in knowledge about regulatory functions and for the inference of networks from data. If there is experimental data showing that both transcription factors A and B activate gene C, but it is unclear whether they can act separately or only in combination, there is not only one determined logical function that can describe their interaction and one can train a network with data to find the most probable one. In probabilistic Boolean networks this uncertainty is taken into account by relaxing the constraint of fixed update rules F and instead permitting one or more functions per node x. Thus, function f_i is replaced by a set of functions

$$F_i = \{f_j^{(i)}\}, \text{ with } j \in \{1, ..., l(i)\}, \tag{2.9}$$

where f_j^i is a Boolean function determining the value of node x_i and $l(i)$ the total number of functions for node x_i. In each update step the functions are chosen randomly according to their given probability c_j^i. Since c_j^i are probabilities they must satisfy

$$\sum_{j=1}^{l(i)} c_j^i = 1 \tag{2.10}$$

A PBN is called independent, if the elements of different F_i are independent. Assuming independence, there are at most

$$N = \prod_{i=1}^{n} l(i) \tag{2.11}$$

possible PBN realizations, each of which is a classical BN. At any point in time t we choose one of these networks to determine the state of time $t+1$. If f_j is the j^{th} realization of the PBN,

$$f_j = \{f_{j_1}^{(1)}, f_{j_2}^{(2)}, ..., f_{j_n}^{(n)}\},\ 1 \leq j_i \leq l(i),\ i = 1, 2, ...n. \tag{2.12}$$

The probability to choose this realization is:

$$P_i = \prod_{i=1}^{n} c_{j_i}^{(i)},\ i = 1, 2, ..., N \tag{2.13}$$

As P_i is a probablity to choose one of the realizations, it is obvious that $\sum_{i=1}^{N} P_i = 1$.

Simulation of Probabilistic Boolean Networks Using Markov Chains

The statespace of a PBN can be interpreted as a homogenous Markov chain, which can be used to simulated its dynamics. A Markov chain is defined as a set of random variables following the Markov property that, given the present state, the future and past states are independent.

$$P(X_{n+1} = x | X_n = x_n), \tag{2.14}$$

where X_i are random variables from a set S of states. In time homogeneous Markov chains the probablity of transition is independent of n. The state transitions of a BN have exactly these properties and can be stated as a $2^n \times 2^n$ state transition matrix $\mathbf{A}$

$$\mathbf{A}_{ij} = \begin{cases} 1, if \exists\ s_i \rightarrow s_j \\ 0, \text{otherwise} \end{cases} \tag{2.15}$$

Using this matrix and the probability of that PBN realization P_i, as defined in equation 2.13, we can calculate the transtion matrix of a given PBN as the weighted sum of its realizations:

$$\mathbf{A} = \sum_{i=1}^{N} P_i \mathbf{A}_i, \tag{2.16}$$

where $\mathbf{A}_i$ is the transition matrix of the PBNs j^{th} realization. Due to the homogeneity of the Markov chain we can then compute the transition probablity after k steps as the k-th power of $\mathbf{A}$. We can calculate the dynamics of the PBN following a starting probability distribution of states $\mathbf{D}^{\circ}$ efficiently by

$$\begin{aligned} \mathbf{D}^{t+1} &= \mathbf{D}^{t}\mathbf{A} && (2.17) \\ &= \mathbf{D}^{\circ}\mathbf{A}^{t+1}, && (2.18) \end{aligned}$$

where $\mathbf{D}^{t}$ is the state distribution at time point t. We can then find a stationary distribution π such that $\pi = \pi\mathbf{A}$. These Markov properties were exploited in the simulation done in section 4.2.2. All simulations were carried out using the *R*-Package BoolNet (Müssel, Hopfensitz, and Kestler, 2010).

2.3.2 Ordinary Differential Equations

As already pointed out in section 1.3, the most common approach on modeling biological systems is to describe them with ordinary differential equations (ODE) (Klipp et al., 2007). In this thesis we employ these in the software *ModelMage* Flöttmann et al., 2008and the given example (Section 3.2). Here we will only give a brief introduction to the approach. More detailed explanations can be found in various textbooks on the topic (e.g. Klipp et al., 2009; Szallasi, Stelling, and Periwal, 2010).

ODEs have many advantages as a modeling framework. They are frequently used in many scientific fields and there are very good tools available to work with these systems (Hoops et al., 2006; Maiwald and Timmer, 2008) in a biological context. In an ODE system, changes in the quantity of biological entities are described by a differential equation each. These entities can be anything from an individual in a predator-prey-model in population dynamics to a protein in molecular Systems Biology.

An ordinary differential equation system describes the changes in the system depending on its current state. In reaction systems it consists of a number of terms, that describe the different processes a species is involved in. The concentration of a variable x_i in such a system is determined by an initial concentration $x_i(0)$ and a differential equation of the form

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \textit{synthesis} - \textit{degradation} - \textit{complexation} + \ldots \tag{2.19}$$

Each of the single terms of these equations represents the velocity of a single reaction x_i is involved in. They can be a function of species concentrations and are usually kinetic laws, e.g. Michaelis-Menten or mass-action kinetics. As the species of an interaction network are dependent, their fate is determined by a system of differential equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i(x_1, x_2, ..., x_n, p_1, p_2, ...p_j, t), \tag{2.20}$$

where p_j are the kinetic parameters of the function $f_i(x, p, t)$. The system can be written in vector notation as

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = f(\mathbf{x}, \mathbf{p}, t), \tag{2.21}$$

where $\mathbf{x} = (x_i, x_2, ..., x_n)^T$, $\mathbf{p} = (p_i, p_2, ..., x_j)^T$ and $\mathbf{f} = (f_i, f_2, ..., f_n)^T$.

Parameter Estimation

Parameters of an ODE model describe dynamic properties of a system like the efficiency of an enzyme catalyzing a reaction or simply the rate of diffusion in a system. Theoretically all parameters could be measured given the right experiments with infinite precision. In reality however measurements are always noisy and often limited to some components of the system, and the majority of parameters is not measured at all. Most biological experiments somehow measure the (relative) abundance of proteins or nucleotides, which leaves models unparameterized in many cases. Therefore modelers often need to resort to optimization techniques to adjust the model parameters so that the model simulations replicate the measured abundances as good as possible. There are sophisticated methods for this estimation and will only briefly explain the basic principles they build upon.

Given an ODE system as defined in Equation (2.20) and a set of noisy measurements d for n time points we can define the difference between simulation and measurements as

$$\varepsilon_i = f(t, \mathbf{p}) - d_i(t) \tag{2.22}$$

where ε is the measurement error in each time point i. Parameter estimation minimizes ε by finding a set of parameters p so that

$$\varepsilon^2 = \sum_{i=1}^{n} \varepsilon_i^2 \overset{!}{=} \min, \tag{2.23}$$

by following the least squares approach. In other words, it minimizes the squared difference between the simulated values $y_i = f(t, \mathbf{p})$ and the measured values d_i in every time point, which is called the residual sum of squares (RSS). Assuming an equal variance σ^2 of the measurement errors for every time point this can be done using the log-likelihood function

$$L(\mathbf{p}|\mathbf{d}) = -\frac{1}{2} \sum_{i}^{n} \left(\frac{d_i - f(t_i, \mathbf{p})}{\sigma} \right)^2. \tag{2.24}$$

This function expresses the probability of the parameter set p given the dataset d and accordingly has to be maximized to find the best set of parameters.

Given function f is linear and a complete dataset d, the best fit can be found by solving the system using the Gaussian algorithm. For sparse and noisy data and nonlinear system that typically occur in practice one needs to resort to iterative approaches like the Gauß-Newton method or improved variations, e.g. Levenberg-Marquardt. To find an optimum, these methods start from a given point in the parameter space , linearize the function, numerically compute a gradient, and follow the steepest decent in each step until they converge in an optimum. However, these so called local optimizers can only find the local optima that are closest to the starting conditions. Because the landscape of the objective function can have multiple local optima, the start values have to sample the whole parameter space and the optimization has to be executed for all the samples, to locate global optima. To tackle the often very large parameter spaces of nonlinear optimization problems a class of heuristic optimizers emerged that have different strategies to find global optima (e.g. Kirkpatrick and Vecchi, 1983).

Because of the scarcity of data, a common problem in Systems Biology is that the degrees of freedom of a model are too many compared to the available measurements (overfitting), or that paramaters of a model cannot be determined uniquely due to structural constraints (non-identifiability). Parameters are non identifiable, if a change in one parameter a can always be balanced by the change in another parameter b and thereby keeping the objective function at the same level. Both problems have the effect, that there will be multiple sets of parameters that fit the data equally well, which limits the power of the fitted model to predict different conditions. In this case one needs to reduce the complexity of the model to make best use of the data.

To be able to reduce an existing model, one often needs to compare structurally different model versions which is one of the points I will address in Chapter 3. This comparison between different models in one ensemble can be accomplished using the Akaike information criterion (AIC) (Akaike, 1974). The AIC can be calculated using the RSS values from the best obtained fit for each model as

$$AIC = 2k + n\left(\ln \frac{RSS}{n}\right), \tag{2.25}$$

where k is the number of parameters and n is the number of observations. The AIC is a measure for the goodness of fit of a model that takes the model complexity into account and thereby defines the best model as a compromise between low RSS ($= \varepsilon^2$) values and small numbers of parameters. This prevents overfitting and ranks simple models higher than more complex models, which lives up to the principle of parsimony.

Animals and computers are both so complex that something on the level of software explanation must be appropriate for both of them.

Richard Dawkins

3 Software Frameworks for Accelerated Modeling

Based on:

- Jörg Schaber, Max Flöttmann, Jian Li, Carl-Fredrik Tiger, Stefan Hohmann, Edda Klipp (Jan. 2011). "Automated ensemble modeling with modelMaGe: analyzing feedback mechanisms in the Sho1 branch of the HOG pathway." In: *PloS one* 6.3. Ed. by Alan Ruttenberg, e14791

- Max Flöttmann, Falko Krause, Edda Klipp, Marcus Krantz (2012). "Reaction-contingency based bipartite Boolean modelling". In: *under review*

This chapter introduces two tools that aim to speed up the idealized cyclic workflow for Systems Biology research in the theoretical part (Figure 1.3.1). We built both tools to improve the modeling process by quick generation of different model versions, and simple testing for agreement of these with data and literature knowledge. The tools differ in the way they aim to reach this goal and are designed for different modeling approaches that cover contrasting levels of detail (Figure 1.3.2). We will describe both tools and their core principles and exemplify their application using different models of mitogen activated protein (MAP) kinase signaling pathways in *Saccharomyces cerevisiae* that were developed in parallel with the presented tools. Because the modeling parts will primarily serve as examples for the presented software approaches we will keep the biological introduction in this chapter to a minimum.

3.1 Introduction

To be really able to utilize the power of the multidisciplinary Systems Biology approach, it is absolutely necessary to develop tools for scientists from different fields to cooperate more closely. This important task is often neglected by many scientists and funding bodies. Well built and freely available software tools enable accurate and reproducible scientific results mainly because they can help to soften the differences between experimental biologists and modelers. Standard software tools also have to stay in standards for data formats, which have been developed in Systems Biology in recent years (Klipp et al., 2007).

3.1.1 Available Software

As mentioned in chapter 1, Systems Biology strongly depends on supporting software. Many researchers spend a lot of their time developing software they need and that can also help others. The catalog of Systems Biology software supporting the Systems Biology Markup Language (SBML) (Hucka et al., 2003) contains 247 software packages alone (`http://www.sbml.org`).

Some of these applications are full simulation frameworks that allow modelers to build, simulate, and analyze models (eg. *Copasi*, *CellDesigner*) (Funahashi et al., 2008; Hoops et al., 2006), while many others are aimed towards more specific tasks like annotation (Krause et al., 2010) or visualization (König, Dräger, and Holzhütter, 2012). Still, there is no software that can handle ensembles of model alternatives in SBML and compare these in an automated way. Programs supporting Boolean models enable the user to simulate a network, find attractors and perform analyses on network properties, e.g. attractor search (Section 2.3.1). Although there is software available to "fit" networks to measured data and to translate Boolean models into simple ODE systems (Di Cara et al., 2007; Wittmann et al., 2009), there is no simple software available for the step-by-step analysis and visualization of Boolean simulations on network graphs with simultaneous state space visualization.

We developed our software based on our own needs and the lack of software that could help with faster iterations in model development. The given examples in the MAP kinase system show the advantages our software can offer in this process.

3.1.2 MAP Kinase Pathways in Yeast

The demonstration system for our software is the MAP kinase signaling system in baker's yeast. Therefore, I will briefly give an overview of the necessary biological background. We chose MAP kinase systems in yeast as examples for our software, because it is a well studied signaling system with a lot of available detailed data (Botstein and Fink, 2011) and existing models ranging from small focused to exhaustive large scale approaches Klipp, Nordlander, and Krüger, 2005; Muzzey and Gómez-Uribe, 2009; Schaber et al., 2010. Yeast is one of the most important model organisms for eukaryotic signaling and metabolism. Its simple cultivation and widespread industrial use make it a good target for scientific efforts. It was the first eukaryotic organism to have its genome of $\approx$ 5000-6000 genes fully sequenced by 1996. Additionally, MAP kinase signaling cascades are an abundant theme that reappear in many pathways and across a wide array of species.

Yeasts need the MAP kinase system to sense and react to environmental stresses and stimuli, like hormones or changes in nutrient supply. The high osmolarity glycerol (HOG) pathway, which we will focus on in the following sections, senses and responds to increasing extracellular osmolyte concentrations (reviewed in Hohmann, 2009). This reaction is necessary for the cells to be able to survive, because without response they would dehydrate due to osmotic water outflux. The pathway senses turgor loss (caused by water outflux) via two branches converging on a MAP kinase kinase Pbs2 and responds by increasing cytosolic glycerol concentrations (reviewed in Hohmann, 2009) to balance the osmotic pressure. The antagonist to the HOG pathway is called the protein kinase C (PKC) pathway, which among other stimuli senses increased turgor (reviewed in Levin, 2011). Under sparse nutrient conditions, yeast cells are able to transform into durable haploid spores by meiotic division. In the haploid form there are two mating types (α and a) that can mate to form a diploid cell when the nutritional conditions improve again. To locate a mating partner the haploid cells sense a pheromone in the medium that is produced by the complementary mating type and form a tip (shmoo) in the direction of the signal. This process is directed by the mating (MAT) pathway which is only active in the haploid form (reviewed in Bardwell, 2005). The pseudohyphal differentiation (PHD) pathway is the least well studied of the mentioned pathways. Its suspected function is the regulation of switching to a filamentous growth under nutrient (nitrogen) depletion.

These pathways were the subject we used to test both frameworks we developed. Some of the mentioned pathways are interconnected by cross-talks and are part of the Boolean model presented in section 3.3.3 as an example for the *rxncon* software framework. The following section focuses on the *ModelMaGe* software and presents simple ODE models of the HOG pathway.

3.2 Model Management and Generation for SBML Models

In practice, mathematical modeling includes a lot of data management and documentation of model alternatives. The idealized Systems Biology workflow (Figure 1.3.1) in reality seldom is a purely cyclic process. Most of the time one has multiple hypotheses about a mechanism that can be compared to datasets. These hypotheses need to be formalized in different structural models and modelers end up with a whole collection of models to discriminate. Problems arise whenever there are changes affecting all of the models in such an ensemble. In that case, the changes have to be applied to all the alternatives and all model files have to be edited manually. This can be a slow and error prone manual process that probably has to be done more than once during the iterative model creation.

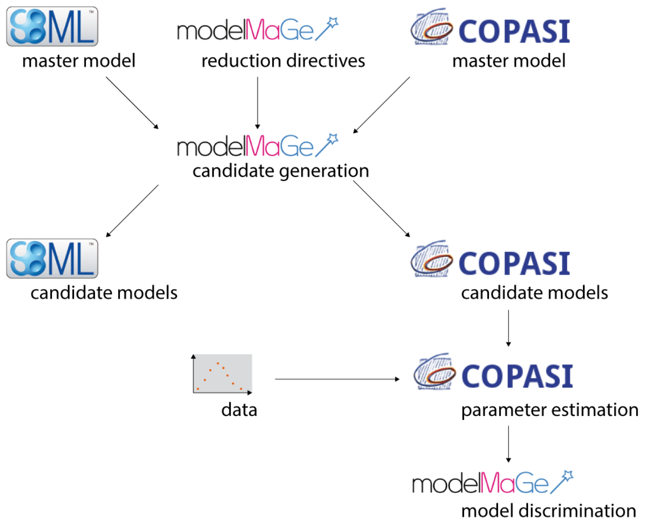


Figure 3.2.1: Generalized workflow of the *ModelMaGe* software. Input is a complete master model and output is a set of candidate models that can also be fit to data and discriminated.

ModelMaGe is aimed towards solving this problem by following a simple approach: All the alternative models are created from a master model following a defined reduction syntax. Following this approach speeds up the initial modeling process and avoids errors, as changes only have to be made in one place, the master model. Technical details of the software were already discussed in my master thesis (Flöttmann, 2008) and in (Flöttmann et al., 2008), which is why I will only briefly describe the workflow of the software and concentrate on the application of the software for acceleration of modeling in a concrete example.

3.2.1 Model Formats and Workflow

ModelMaGe is able to work with ODE models in the SBML standard (Hucka et al., 2003) and the *Copasi* (Hoops et al., 2006) file format. Both formats can be used as a master model and alternative models can also be produced in the two formats.

The first output of *ModelMaGe* is a set of model files generated following the reduction directives. The modifications to the master model are performed following simple logical directives passed to the program via the command line or a separate file (Listing 3.1). The produced files can be imported into any SMBL compliant software to be analyzed. If the master model was given as a *Copasi* file, containing a mapping for a parameter estimation, the software can automatically fit the output models to given datasets and compare the results using the AIC (Section 2.3.2). This facilitates an easy discrimination between the various candidate models generated by *ModelMaGe*. The general workflow of the software is depicted in Figure 3.2.1.

```
modelmage.py −r ‘‘species_5 & reaction_9:species_11 &
    reaction_13’’ −k ‘‘reaction_2(MA) reaction_3(mMA)’’
    Sho1Master.cps
```

Listing 3.1: An example of the commandline syntax of a *ModelMaGe* call. Models are generated by removing species or reactions from the master model (argument "-r") or setting new kinetic laws for reactions (argument "-k").

3.2.2 Analyzing Feedback Mechanisms in the Sho1 Branch of the HOG Pathway

We utilized *ModelMaGe* in a study on the adaptation mechanisms of the Sho1 branch of the HOG pathway. This study was based on a published model by Hao et al. (2007), which proposes a negative feedback mechanism of phosphorylated Hog1 onto the upstream components of the pathway and especially on the membrane protein Sho1.

The aim of the study was to systematically explore different alternative hypothesis on the mechanism of adaptation of the HOG pathway and to test which of these are best supported by data published by Hao et al. (2007) and our own experiments.

We built a master model that includes the original model by Hao et al. (2007) as well as our own alternatives, following the premise of simplification. In order to find the most simple working model, we left out major mechanisms (e.g. turgor, transcription, volume change) that are known to be involved in the HOG pathway. A wiring scheme of the master model can be found in Figure 3.2.2. In summary, this model contains the MAP kinase cascade from *Sho1* to *Hog1*, the negative feedbacks proposed by Hao et al. (2007), and a negative feedback through an integral response. This negative feedback involves the stimulation of the production of intracellular *Glycerol* by phosphorylated *Hog1* (*P-Hog1*). The component *Signal* represents the difference between intra- and extracellular osmotic potential, which is proposal to be sensed via turgor pressure (for simplicity represented by the difference between *OuterOsmolarity* and *Glycerol*). The concentration of *Glycerol* can also be influenced by the state of the membrane channel *Fps1* which, if closed hinders the outflow of glycerol and can be actively controlled in some candidate models.

From this master model we generated 12 candidate models that systematically explore combinations of the features combined in the master model. The hierarchy of these models is depicted in Figure 3.2.3. The leftmost branch of the tree contains the models that include the feedback mechanisms proposed by Hoa (*C10* is the original model) and different submodels thereof. Simplifications are done with *ModelMaGe* by removing species or using simpler reactions (For a list of candidate model structures refer to Figures A.2.1 and A.2.2). Branches one and two contain the *Sho1* desensitization feedback, in which *P-Hog1* mediates conversion of *Sho1a* to its inactive form *Sho1i*. Branches two to four contain the integral feedback via potential difference (*Signal*) as described above. These branches vary in the number of signaling intermediates present in the cascade and the simplest models

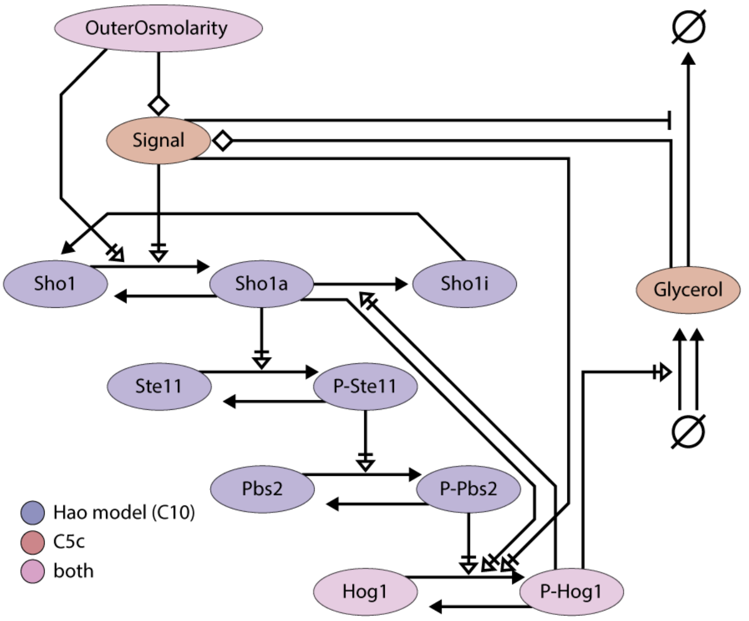


Figure 3.2.2: Structure of the master model in SBGN. Shaded components are components of the original Hao model (dark), of the C5c model (light), or both (hatched).

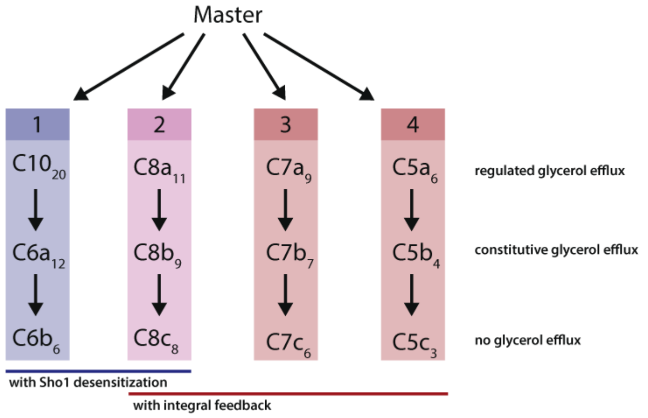


Figure 3.2.3: A tree representing the relationships between the derived candidate models and the master model. Models are named according to the number of species and their features. Numbers in the subscript indicate the number of fitted parameters.

in branch four only contain five species. The further simplifications done in each branch mainly concern the control of glycerol efflux, representing three main hypothesis: regulation of glycerol efflux by *Signal*, constitutive efflux, no efflux.

3.2.3 Model Discrimination

Model generation was automatically handled by *ModelMaGe* using the master model and the creation directives of the format given in Listing 3.1. The master model was generated in the *Copasi* format and parameter estimation to the data timecourse data provided by Hao et al. (2007) was set up using *Copasi*. The candidate models were then automatically fitted and ranked by their Akaike information criterion (AIC) values by *ModelMaGe* (Section 2.3.2). The output is listed in Table 3.2.1.

Table 3.2.1: Candidate models ranked by the AIC value. *k* number of parameters, *SSR* value of objective function for best fit, *AIC* Akaike Information Criterion.

	Model	k	SSR	AIC(c)
1.	C5c	3	0.251373	-38.045
2.	C5b	4	0.25078	-34.104
3.	C7c	5	0.258817	-31.316
4.	C6a	12	0.0614917	-29.246
5.	C7b	6	0.258223	-27.373
6.	C7a	9	0.152959	-26.465
7.	C5a	7	0.241075	-25.091
8.	C8c	7	0.258616	-23.335
9.	C8b	8	0.258023	-19.393
10.	C8a	11	0.152514	-14.537
11.	C6b	6	0.739527	-1.069
12.	C10	20	0.048542	164.842

The original model (*C10*) still shows the smallest differences to the data, as evident by its small SSR value. Nevertheless, due to its large number of parameters it is ranked in the last place. The simplified version we tested (*C6a*), still showed very small residual errors, and also ranked much higher, as it has 8 parameters less due to a shorter cascade missing Ste11 and Pbs2.

The Hao model shows sustained oscillations after an osmotic shock Figure 3.2.4, that increase with the strength of the shock. The recent literature however provides data with a higher time resolution and clearly shows that these oscillations can not be found in the experimental system (Mettetal and Muzzey, 2008; Muzzey and Gómez-Uribe, 2009; Schaber et al., 2012). Such artificial effects in the predictions of a model are a clear sign for over-fitting and weaken the models predictive power. As shown in Figure 3.2.4 B, we can represent the major features of the system, namely rapid increase upon shock and slow adaptation, with a model with only 3 parameters and a very simple structure (model C5c). Although this model has a much larger residual error, it is ranked first by its AIC value.

3.2.4 Model Predictions and Validation

We show that the double shock data from the Hao paper can be described to a large degree with the simple model we developed using *ModelMaGe* and we suspect a strong over-fitting in the original model. Although over-fitted models often show spurious effects in the data, they are usually better in predictions than too simple under-fitted models (Burnham and Anderson, 2002).

In order to find out whether our minimum model is still predictive, we did different *in* *silico* experiments to find conditions that clearly set both models apart from each other. The setup which turned out to show the biggest differences was a simple triple shock experiment. The Hao model predicts a much smaller response to the third shock and does not show adaptation after that, whereas model C5c responds in the same way it did in the first two shocks. In order to falsify one of the models we measured data of *P-Hog1* time courses after repeated osmotic shock with 0.4 M KCl for both. The amount of KCl was added to the culture three times with 30 minutes intervals. Our measurements clearly show that the C5c model predicted the experimental outcome much better than the C10 model (Figure 3.2.4).

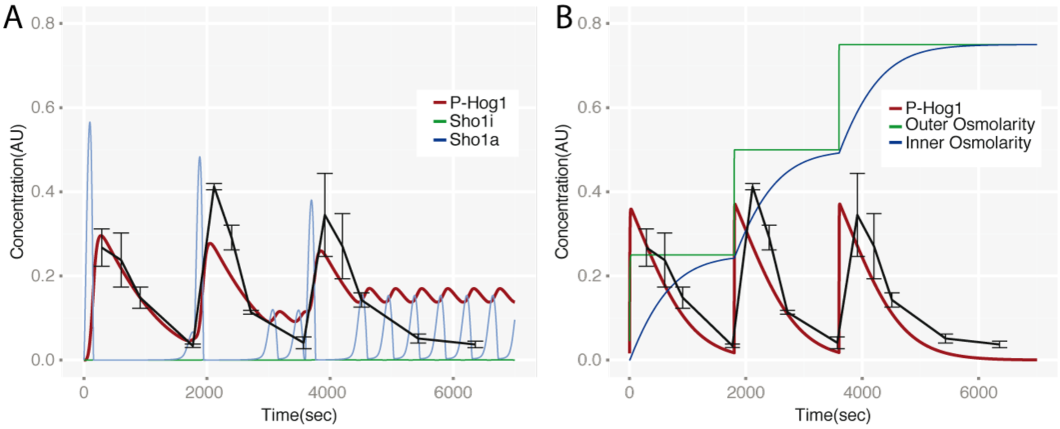


Figure 3.2.4: Triple osmotic shock experiments show the outcome predicted by the C5c model is close to reality. Salt shock was applied at t=0, t=30 min, and t=60 minutes. (**A**) The original Hao model with the best fit using our optimization. (**B**) Best fit of model C5c. The maximum of the 0.4 M KCl triple shock time series is scaled to the maximum of the 1 M KCl single shock time series. The error bars represent the standard deviation of three independent measurements.

Interestingly, the *C10* model was only able to react to a third shock at all when the desensitization mechanism was nearly disabled. The transformation reaction velocity from *Sho1a* to the inactivated form *Sho1i* was nearly zero and *Sho1i* shows no response at all (Figure 3.2.4 A).

This outcome suggests, that our approach to systematic model reduction using alternative candidates is a useful way to improve existing models and combine many datasets with whole ensembles of models.

3.3 Rxncon for Boolean Models

Mathematical modeling of large cellular networks is unfeasible or impractical, mainly due to the large number of model states and parameters needed to describe these networks.

This combinatorial complexity is particularly problematic for signal transduction networks. Their components are often influenced by multiple interaction partners and/or modifications such as phosphorylations, which rapidly combine to a large number of possible configurations – or specific states – of each component. This makes it difficult to build and parameterize large quantitative models, and computationally costly to analyze them. However, mathematical analysis of these networks is an important tool for network validation and understanding, urging a development of methods that can be used even for large complex networks. Boolean modeling provides one of the few feasible approaches to whole-network modeling. Even for systems with a lot of data like yeast signaling it can prove useful for an initial study of network properties and is often used when quantitative effects do not play a major role in the overall qualitative behavior of a network (Section 2.3.1). Therefore, Boolean models are used in a variety of signaling systems (Bauer et al., 2010; Saez-Rodriguez et al., 2007).

The classical Boolean modeling approach does not distinguish between different downstream roles played by a single component activated in different contexts. The definition of which biological entity is a variable in these networks is not standardized, but a common approach is to treat each protein as a variable. This is not sufficient for signaling networks, as modifications of proteins play a major role in these. Ignoring e.g. different phosphorylation sites that allow a kinase to play a part in two different pathways would lead to the activation of both downstream targets once the kinase is set to an "on" state and thereby lead to a wrong signal.

Our work addresses these shortcomings with a bipartite Boolean modeling approach and supporting software, which integrates model generation, simulation, and visualization. We used a state oriented modeling approach with separates update rules based on reactions and contingencies that corresponds directly to the reaction-contingency (*rxncon*) method (Tiger et al., 2012).

The bipartite Boolean model has the same structure as the *rxncon* approach with separate update rules for reactions and for states: States are a function of reactions that produce or consume them, while reactions are functions of states given by contingencies. This bipartite Boolean modeling approach retains the contextual information on activation and distinguishes distinct signals passing through the same component. We integrate this approach in the *rxncon* framework to allow automatic model generation, and benchmark the method with the previously mapped MAP kinase network in yeast. Finally, we demonstrate how this modeling approach can be integrated in the network definition process for validation purposes. Taken together, we present a bipartite Boolean modeling approach that retains contextual activation information, which supports automatic model generation from existing network definitions, and which can be used for iterative network building and validation.

3.3.1 The Rxncon Format

Briefly, *rxncon* is a network definition method which separates reaction and contingency information. The elemental reactions and their corresponding elemental states define the possible signaling events that can occur and the outcome of these events, respectively. The contingencies define the constraints on these reactions, e.g. reaction A depends on state B. Together, the reactions and contingencies can define the network completely and unambiguously, similar to a reaction system with limited kinetic information (activation/inhibition).

The input file can be created as an Excel file or as text based direct input (described further below). The Excel input consists of two lists; the reaction list and the contingency list. The reaction list defines the network topology. Each reaction is defined by two components and a relationship (reaction) between them. In the minimal format as used for the example network in Figure 3.3.1, only reaction and component names are required. Reaction and state IDs are automatically generated. Importantly, the components are always entered in their basic state even if previous modifications are required. These requirements are defined in the contingency list. Each constraint on a reaction must be defined as a contingency, and each contingency consists of three parts: A target, which identifies the reaction that is affected, a contingency, which defines how the target reaction is affected, and a modifier, which identifies the state causing the effect. More complex models may make use of Boolean statements, inputs and outputs, as described further on `http://www.rxncon.org` and in Tiger et al. (2012). A given excel file is loaded directly into the *rxncon* tool from which all export functions as well as the simulation interface are available.

Alternatively, a model can be defined directly as text input. Reactions need to be written exactly as they would appear in the Reaction list in the spreadsheet (`http://rxncon.org/rxncon/test` for examples). Contingencies would be added directly to each reaction after ";", as shown in the small model of the HOG pathway presented in Listing 3.2. We will use this simple model as an example to show the basic transitions from the *rxncon* format to Boolean networks. A visualization of the model by *rxncon* and Cytoscape is also shown in Figure 3.3.1.

3.3.2 Bipartite Boolean Models

The software built for this thesis is an extension to the already available *rxncon* software framework. It adds value to the whole system by providing the means to export the rxncon definitions to a model format that can be simulated without further input. These simulations can be used for rapid model improvements by checking whether e.g. a signal can be transmitted through the network or where exactly it is stopped. This process is assisted by a graphical user interface and a visual representation of the network (Figure 3.3.2).

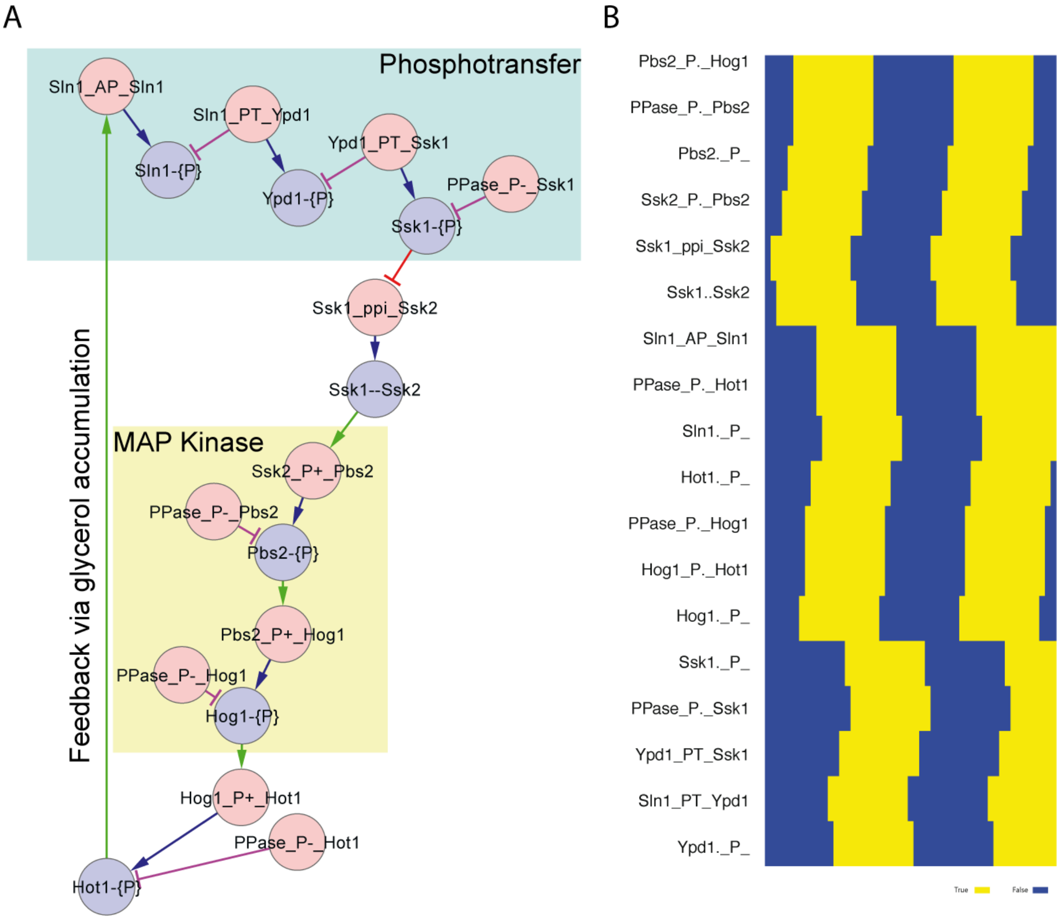


Figure 3.3.1: A small example network of the HOG pathway. (A) The *rxncon* data visualized as a regulatory graph: The network is defined as elemental reactions (red nodes); that produce (blue edges) or consume (purple edges) elemental states (blue nodes), and contingencies showing how states activate (green edges) or inhibit reactions (red edges). The elemental reactions correspond to the edges in a topological network, and the contingencies provide the contextual constraints on the reactions. The example network is a simplified version of the high osmolarity glycerol (HOG) pathway. The rxncon code for this model is given in listing 3.2. (B) Simulation of the model with the turgor feedback included. Due to the feedback from Hot1 on Sln1, the model ends in a cyclic attractor when started with our standard initial values. The exported BooleanNet code for the simulation is given in listing A.1

```
Sln1_AP_Sln1; ! Hot1-{P}
Sln1_PT_Ypd1
Ypd1_PT_Ssk1
Ssk1_ppi_Ssk2; x Ssk1-{P}
Ssk2_P+_Pbs2; ! Ssk1--Ssk2
Pbs2_P+_Hog1; ! Pbs2-{P}
Hog1_P+_Hot1; ! Hog1-{P}
PPase_P-_Ssk1
PPase_P-_Pbs2
PPase_P-_Hog1
PPase_P-_Hot1
```

Listing 3.2: The small example HOG model shown in Figure 3.3.1 formulated in the *rxncon* format. The format uses a very compact notation to describe reactions and contingencies that does not require declarations of e.g. components that other formats need.

Export Logic

Tiger et al. (2012) have previously shown that a *rxncon* network unambiguously defines a model structure and can be exported to SBML, rule based or agent based formats. While these models can be generated automatically, their behavior relies heavily on parameter values that must be estimated from empirical data. Our approach complements these export options with a new Boolean format that is able to capture the qualitative network behavior without any further parameterization.

The Boolean model structure directly corresponds to the *rxncon* regulatory graph (Tiger et al., 2012). This bipartite graph consists of elemental reactions and states as nodes, reaction effects as reaction-to-state edges, and contingencies as state-to-reaction edges. Our approach of encoding the reaction information into Boolean logic uses the same bipartite partitioning and has separate update functions for the reactions, states, and input and output nodes. To be able to use a standard translation from the *rxncon* format to the Boolean format, we had to make certain assumptions about the dependencies that are described in the following.

In our Boolean models, reactions depend on the states that are given as their contingencies and the components that are involved. Contingencies c_a giving quantitative and absolute requirements (k+/! in *rxncon* notation) as well as components d are all needed for the function to be true. States c_n given in negative contingencies (k-/x) simply are negated with a NOT ($\neg$) operator:

$$r(t+1) = \bigwedge c_a(t) \bigwedge d(t) \bigwedge \neg c_n(t). \tag{3.1}$$

Components are part of the Boolean model, but are not influenced by any other entities and are therefore considered constant. Boolean nodes defined in the *rxncon* format are flattened in the update function of the reactions in the Boolean format by adding them recursively to the function. Protein-protein interaction between Ssk1 and Ssk2 (Ssk1_ppi_Ssk2) is inhibited by Ssk1 phosphorylation (Ssk1-{P}). This yields:

$$\text{Ssk1\_ppi\_Ssk2}(t+1) = \text{Ssk1}(t) \wedge \text{Ssk2}(t) \wedge \neg\text{Ssk1-}\{\text{P}\}(t). \tag{3.2}$$

Update functions of states are built up from the producing reactions, the consuming reactions, the involved components, and the state itself. Components are absolute requirements for the state to be true, while the exact structure of the update function depends on the reaction types the state is involved in. Reversible production reactions r_r need to be set to *true* to keep the state active, because the states are considered to decay the state when set to false.

$$s_i(t+1) = \bigwedge d(t) \bigwedge r_r(t) \tag{3.3}$$

In contrast, irreversible reactions r_{irr} cannot switch produced states to false:

$$s_i(t+1) = \left(\bigwedge d(t)\right) \wedge \left(s_i(t) \vee \bigvee r_{irr}(t)\right) \tag{3.4}$$

Output nodes are treated in the same way as states, while input nodes are constantly either true or false.

Updating states can be exemplified by the reactions depicted in Figure 1 B. The state Sln1-{P} of protein Sln1 is produced by auto-phosphorylation and consumed by phosphotransfer to Ypd1. This would be updated by the following rule:

$$\text{Sln1-}\{\text{P}\}(t+1) = \text{Sln1\_AP\_Sln1}(t) \vee \text{Sln1-}\{\text{P}\}(t) \wedge \neg\text{Sln1\_PT\_Ypd1}(t). \tag{3.5}$$

Once the state is true, it cannot be set to false by the producing reaction anymore, because the reaction is irreversible. A different example is the Ssk1-Ssk2 dimer (Ssk1–Ssk2) that is produced by the protein-protein interaction between Ssk1 and Ssk2 (Ssk1_ppi_Ssk2). It follows the update rule: $\text{Ssk1—Ssk2}(t+1) = \text{Ssk1\_ppi\_Ssk2}(t)$. The state would decay if the reaction was false, as protein-protein interactions (ppi) are defined as a reversible reaction.

Simulation and Visualization

Boolean simulation in the software extension is handled by the BooleanNet Python library (Albert et al., 2008) and the biographer library is used for visualization (`http://biographer.biologie.hu-berlin.de/biographer/`). The simulation interface visualizes the network as an activity flow (AF) diagram according to SBGN (Le Novère et al., 2009). The SBGN-AF representation contains the reactions and states from the *rxncon* regulatory graph, but also includes the nodes for each of the network components themselves. It comes in two different styles: the default style visualizes all influences according to the Boolean update rules, while the alternative style mirrors the regulatory graph format. The regulatory graph is more easily accessible as it leaves out the influence of components on reactions and a large number of Boolean operators. Both styles include all components, reactions, states, inputs, and outputs, which can be turned on or off individually by the user to alter the initial state of the simulation.

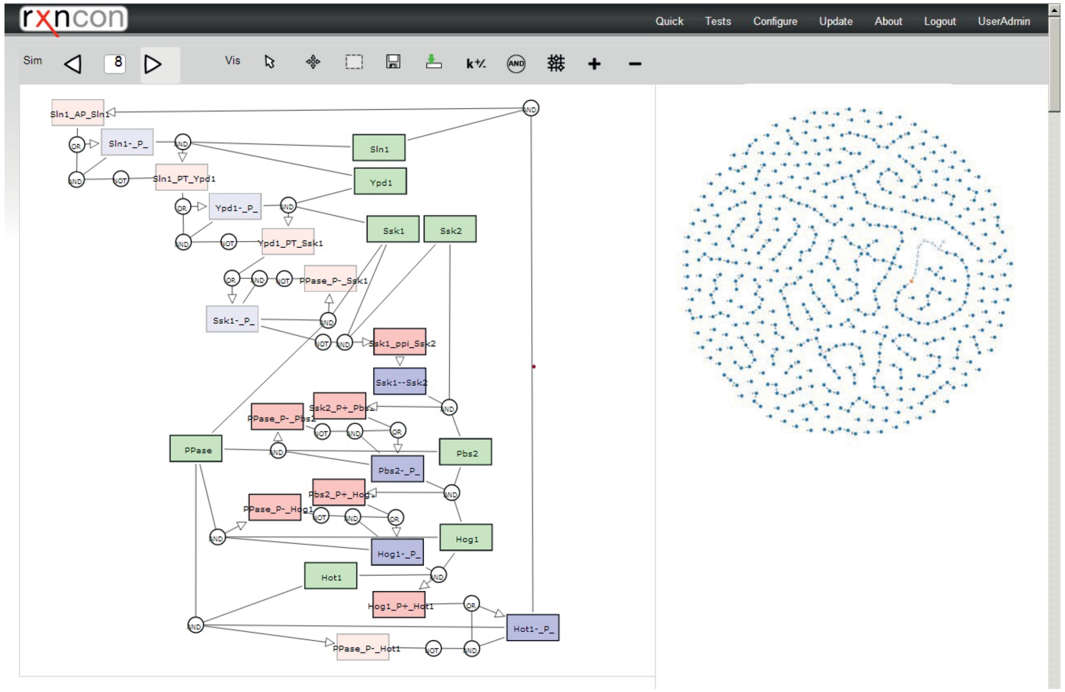


Figure 3.3.2: Screenshot of the simulation interface of the *rxncon* extension. The Exported BooleanNet files can be directly simulated and visualized by the software and users can interact with the model by going through simulations step by step and setting the desired starting state by selecting the nodes in the displayed visualization (left). Users can also check for the correct behavior in the statespace (right) and select states that are then shown in detail on the network.

The network layout can be imported from an Extensible Graph Markup and Modeling Language (XGMML) file from *Cytoscape* and/or edited manually. The software includes different layout algorithms to arrange the network. Possible state trajectories are calculated automatically and visualized within the simulator (Figure 3.3.2 right). The complete state space can only be calculated and visualized for small models, while for larger models the calculation is limited to states reachable from a limited set of starting states (all permutations of the input nodes). The state space visualization allows the user to access a specific state by simply selecting it with the mouse, and also clearly identifies point and cyclic attractors. The modeling interface includes layout algorithms and the node positions can be saved to let previously existing nodes retain their positions if the model is modified.

Hence, this extension provides support for iterative model generation, visualization and simulation. Thereby it facilitates integration of the three steps in the network reconstruction process. As mentioned above, the bipartite Boolean simulation provides a powerful albeit qualitative validation tool. The iteration between model creation and qualitative model validation provides for quality assurance in the model creation process without the need of expensive – if not unfeasible – parameterization and quantitative simulation.

3.3.3 Iterative Model Building and Validation

The potentially most important contribution of the integration of Boolean model generation and simulation in the network definition framework is that it enables iterative model building and validation (Figure 3.3.4 A). The idealized work flow starts from an existing model or a small network reconstruction, which is translated into a bipartite Boolean model and simulated to confirm that the current reconstruction can reproduce the networks' *in vivo* function qualitatively. Ideally, the iteration uses small steps to immediately identify missing and/or erroneous features and to constantly keep the model consistent with *in vivo* observations. This can be done without any overhead due to Boolean model creation, as the network definition format is identical to that used in all other *rxncon* features (Figure 3.3.4 B). The input used to create the bipartite Boolean model can also be exported to the standard SBML format or to formats for rule or agent based modeling, as well as to a range of visual formats, including the SBGN formats. Hence, the Boolean analysis can easily be integrated as a validation step in a modeling effort aiming for a quantitative model without duplication of work.

Validation and Extension of the Yeast MAP Kinase Network

To test our approach on a larger scale, we revisited the carefully curated MAP kinase network of baker's yeast, *Saccharomyces cerevisiae* (Tiger et al., 2012), henceforth referred to as Tiger network. To assess the accuracy and completeness of this network curation, we generated the corresponding bipartite Boolean model to determine which additional features would be needed to (qualitatively) capture the physiological behavior of the network

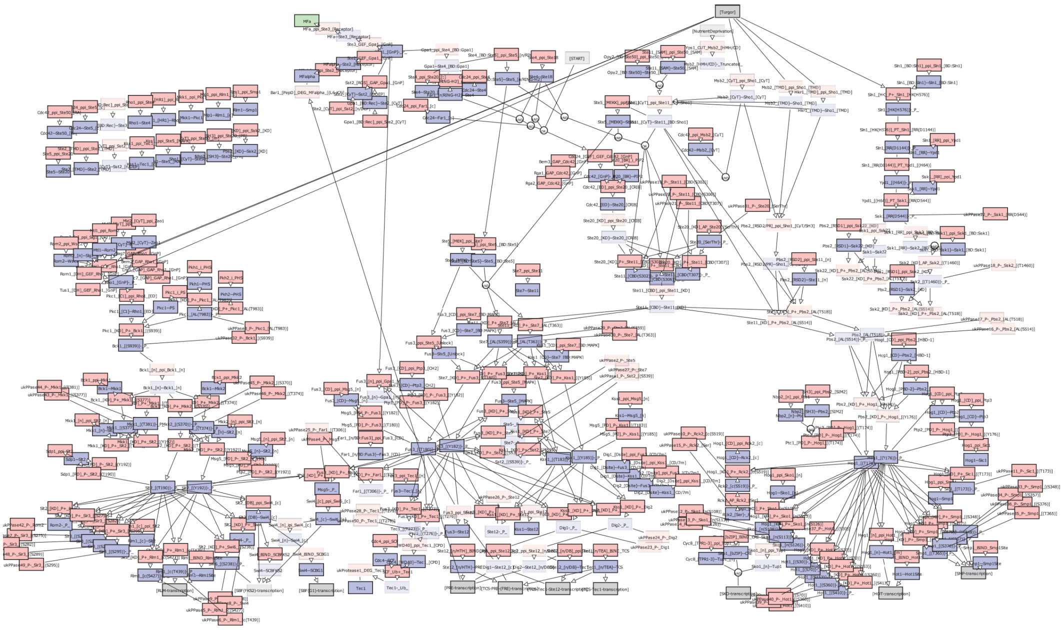


Figure 3.3.3: The complete network structure of the MAPK model. This visualization was done by the *rxncon* extension we implemented. For a larger version of the model see Figure A.2.3 or `http://www.rxncon.org` to see a zoomable version.

(Figure 3.3.4). The network was translated into a bipartite Boolean model assuming all contingencies were absolute, as Boolean simulations cannot deal with quantitative modifiers (Figure 3.3.3, Table A.2.1). Not surprisingly, we found that this network definition is insufficient to predict the network behavior and proceeded to identify the missing features. Most importantly, the Tiger network contains 50 phosphorylation reactions that lack a corresponding dephosphorylation reaction.

To address this shortcoming, we added 50 hypothetical dephosphorylation reactions to make all phosphorylation states reversible (Table A.2.2), which is likely the case *in vivo*. This modification alone was enough to make the Sln1 branch and hence the HOG pathway functional, as measured by its ability to respond to turgor (Figure 3.3.5). Next, we turned our attention to the PKC pathway. It has been reported to respond to increasing osmolarity (García-Rodríguez et al., 2005), although the sensing mechanism remains unclear. To make it turgor sensitive, we simply added a turgor requirement for the guanine nucleotide exchange (GEF) of Rho1. While mechanistically unsatisfactory, this is sufficient to make the PKC pathway responsive to turgor. Importantly, no additional modifications are needed downstream for the signal to reach its targets.

The MAT pathway required more complex adjustments, in part due to the interconnection with the HOG and PHD pathways. Yeast mating only occurs between haploid yeast cells of complementary mating types MATa and MATα. To simulate the well stud-

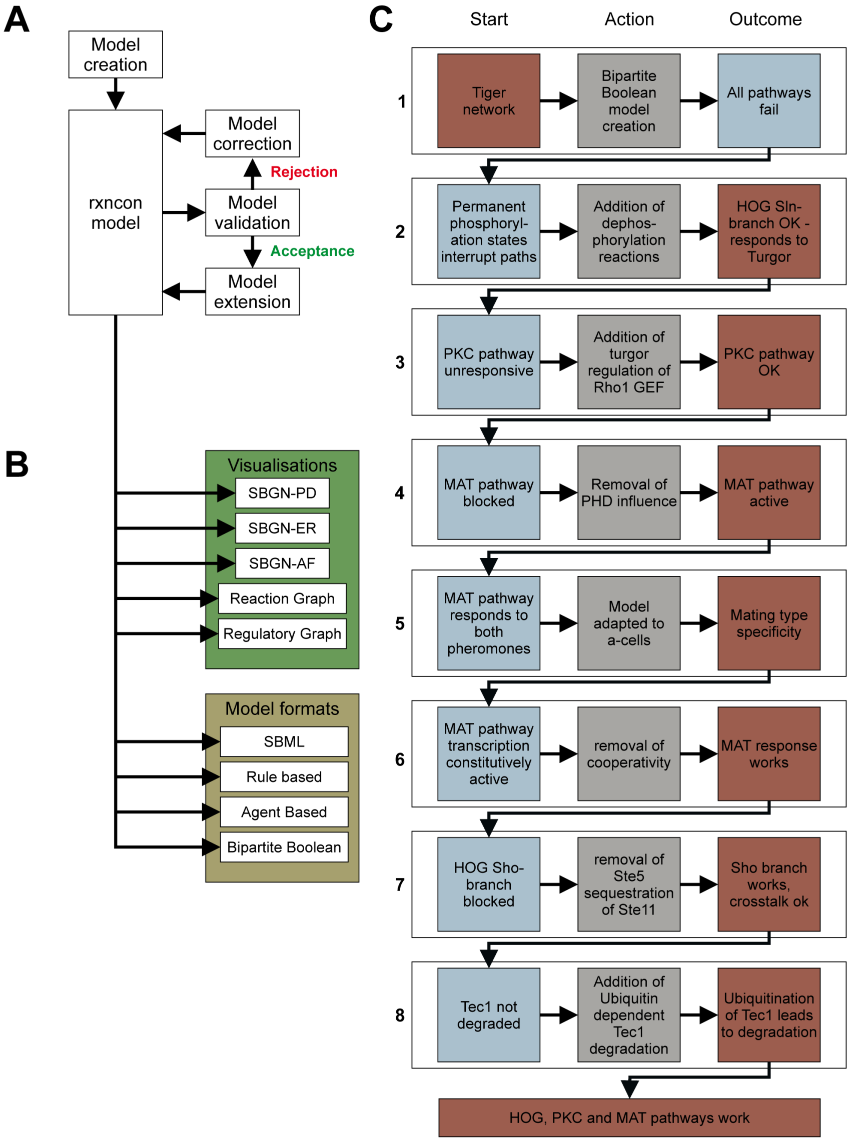


Figure 3.3.4: Iterative model building and validation as a tool to guide and validate network reconstruction. (A) Idealized workflow for model building: Model extensions and improvements are done in small steps, with each step being evaluated as a Boolean model. (B) The *rxncon* database underlying the Boolean model is fully compatible with the other *rxncon* features, including a range of visualizations and automatic model generation in formats suitable for quantitative modeling. (C) The iterative improvement applied on the yeast MAP kinase network. Only a limited number of changes were needed to make the HOG, PKC and MAT pathways functional (Table A.2.1). The single largest change was the addition of 50 hypothetical dephosphorylation reactions (Table A.2.2).

ied MATa-cells, we removed the MATα-cell specific mating receptor (Ste3), and added a negative feedback loop on the pheromone response by allowing degradation of alpha factor only after gene induction of Bar1. Next, we eliminated the interference from the only partially defined PHD pathway. In the Tiger network, the PHD and MAT pathway stimulates some of the same components, which was translated as absolute requirements hence blocking these reactions completely in the Boolean model. To remove this block, we simply removed the influence of the PHD pathway by removing the effect of four contingencies and corrected the requirement for two others (Table A.2.1).

Finally, we removed the cooperative binding of the downstream transcription factors (which again were interpreted as absolute requirements and hence blocking reactions unduly) and added the ubiquitination dependent degradation of the Tec1 transcription factor, which was missing in the Tiger network. In total, we needed to adjust only ten out of 281 contingencies and add one reaction and one contingency to make our Boolean model of the MAT pathway work according to our current understanding (simulations Figure 3.3.5).

We resolved the HOG-MAT crosstalk by removing one final contingency, namely the ability of Ste5 recruitment of Ste11 to block the interaction of Ste11 and Sho1. While this block is likely true for each Ste11 bound to Ste5, the amount of Ste11 in the cell vastly exceeds that of Ste5, making a complete inhibition by stoichiometric binding impossible (Thomson et al., 2011). Taken together, the main changes were addition of 50 new dephosphorylation reactions and turgor regulation of Rho1. Additionally, we corrected the assumption of absolute effects of 4% of all contingencies and added transcriptional induction of Bar1 (Manney, 1983) and Tec1 degradation after ubiquitination (Bao, Schwartz, and Cantin, 2004). With the addition of these small changes, the model was able to reproduce physiological behavior of the included pathways in a qualitative fashion.

3.4 Discussion

In this chapter we described two software applications that we developed and tested in research projects connected to the MAP kinase pathways in yeast. Both applications have shown their value in accelerating the modeling process on different levels but in similar ways. While *ModelMaGe* assists the user in model generation and discrimination of ODE models, Boolean *rxncon* helps the user in an earlier phase to build and test qualitative models. In the end both tools are a way to ease the often frustrating model refinement steps researchers have to repeat often, and thereby reduce errors and speed work up in an early phase of the project.

In the *ModelMaGe* project we were able to reject a hypothesis formulated by a former publication (Hao et al., 2007) about the regulation of the HOG pathway and systematically explore new hypothesis and models using the software. We could show that the formerly proposed model simulated spurious effects that are not supported by data which is prob-

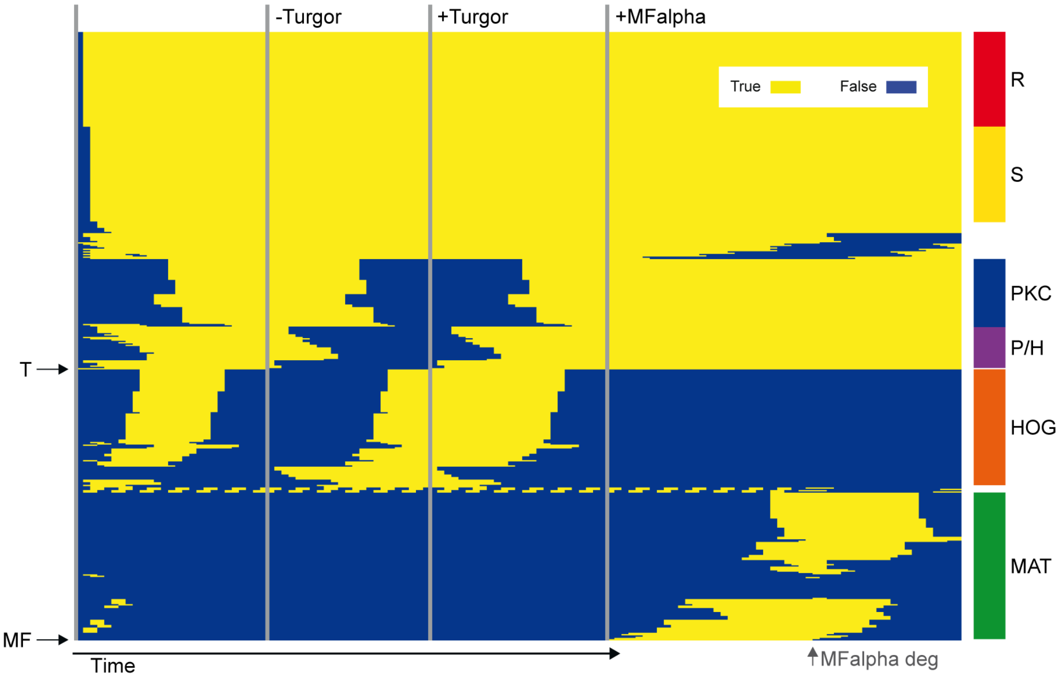


Figure 3.3.5: Time evolution of the Boolean version of the MAPK model under different conditions displayed as a heatmap. States and reactions are hierarchically clustered on the vertical axis by their activation profile over time (horizontal axis). Blue represents inactive states and yellow active states, respectively. The model starts in the standard state assigned by the exporter plus changes in Table A.2.1. Vertical grey lines indicate external stimulus changes after an attractor is reached. Turgor is turned off after steady state is reached (t=27), which activates the Hog pathway. Turgor is switched on again at the next attractor (t=50) and the HOG pathway is inactivated again. From the following attractor MFα is activated (t=75), which clearly turns on the mating pathway, and adapts by degrading MFα. The pathway components cluster together in their state evolution, including a group of early PKC genes that cluster with the (negative) Sln-branch of the HOG pathway (P/H). The unregulated reactions (R) and states (S) are turned on at time step 1 and 2, respectively, and stay constitutively active.

ably caused by over-fitting. Additionally we could fit our simpler alternatives to the published data, albeit with larger residual errors. Nevertheless our simpler model represents the data much better in terms for parsimony and therefore gets lower AIC values. In order to see which model is more predictive for the pathways behavior under altered conditions we did *in silico experiments* to find conditions in which both models differed sufficiently in their predictions. We found that a third shock after 60 minutes was enough to clearly distinguish between the two predictions and therefore carried out an additional experiment which confirmed the predictions of our model. This validation means we completed one cycle of the Systems Biology workflow during which we learned new properties of the system. The use of the *ModelMaGe* software makes the results completely transparent and easy to replicate, as only the master model and reduction directives are sufficient to create and test the validated as well as the rejected models. This facilitates reproducibility of the modeling method and helps to communicate the outcome.

Biologically we could show with this example that the inherent feedback of the HOG pathway, consisting of the adaptation of the intracellular glycerol levels is sufficient for the signal adaptation after osmotic shock. Even though the phosphorylation of Sho1 by Hog1 has been shown by Hao et al. (2007), this does not seem to be the main desensitizing mechanism. Glycerol accumulation mediating adaptation and Hog1 de-activation probably acts via removal of the stimulus, which in turn might be volume or membrane related, e.g. turgor pressure (Schaber et al., 2010). It has been shown for the wild type and the Sln1 branch of the HOG pathway that such an integrator feedback is probably responsible for the adaptation response (Klipp, Nordlander, and Krüger, 2005; Mettetal and Muzzey, 2008). Here, we provide computational as well as experimental evidence that this is also the case for the Sho1 branch.

The *rxncon* project aims in a different direction. It does only represent a small part of the whole cycle, that deals with the initial model generation and testing. Its use of data is only indirectly through conclusions drawn in the literature about qualitative network behavior. It helps to check whether or not a structural model can represent this knowledge in a qualitative Boolean way in a quick and easy manner integrated into a standard workflow. The software implements a new bipartite Boolean modeling approach supported by automatic model generation, simulation and visualization using the *rxncon* framework. Our Boolean approach retains contextual activation information and avoids inappropriate pathway crosstalk, even when the signal passes through shared components. The interactive visualization the software uses, enables model validation at a glance even for large models. By using the MAP kinase network as an example, we demonstrate the use of Boolean modeling for a preliminary model validation and show how it can be integrated in the model construction process. We envisage this iterative process of model building and qualitative validation to be a useful tool in construction of network maps and even quantitative mathematical models. By using this approach we found a number of reactions missing from the model that prevent it from being simulated, which we then added in a step wise manner.

This shows that we are close to a functional understanding of the HOG, PKC and MAT pathways, that this functional understanding can be expressed within the *rxncon* formalism, and that the iterative model building and bipartite Boolean simulation is a potent tool to identify and correct missing or erroneous features in even large models.

In summary both developed tools show their applicability to signaling networks in the presented examples and *ModelMaGe* was has already been used in a different context (Klotz et al., 2011). Both tools help to find inconsistencies by comparison or fitting to data. They accelerate the modeling process and facilitate communication of results and thereby fill existing gaps in the Systems Biology software landscape.

4

Gene Regulatory Networks in Pluripotency

Based on:

- Max Flöttmann, Till Scharp, and Edda Klipp (Jan. 2012). "A stochastic model of epigenetic dynamics in somatic cell reprogramming." In: *Frontiers in physiology* 3.June, p. 216

- Nancy Mah, Ying Wang, Mei-Chih Liao, Alessandro Prigione, Justyna Jozefczuk, Björn Lichtner, Katharina Wolfrum, Manuela Haltmeier, Max Flöttmann, Martin Schaefer, Alexander Hahn, Ralf Mrowka, Edda Klipp, Miguel a Andrade-Navarro, James Adjaye (Jan. 2011). "Molecular Insights into Reprogramming-Initiation Events Mediated by the OSKM Gene Regulatory Network." In: *PloS one* 6.8, e24351

In this chapter we examine a system on a longer time scale and a different level of regulatory mechanisms than in chapter 3. Induced pluripotent stem cells — the system under examination — received an increasing amount of attention in recent years, which culminated in a Nobel price for their inventor Shinya Yamanaka in 2012. Fast technical advances have been made in their production, but the underlying processes are still unclear. In order to propose possible answers to open questions about these mechanisms, we developed a model of the interplay between different levels of regulation in this system. Before constructing the model, we analyzed the behavior of developmental gene regulatory networks (GRNs) and their reaction to external perturbations based on different high-throughput datasets.

4.1 Introduction

4.1.1 Stem cells and somatic reprogramming

Multi-cellular organisms are the most complex form of life that ever existed on earth. The cell as a form of organization is still a riddle to be solved, and communication between these highly complex systems that leads to the formation of higher organisms, is so intriguingly complicated that its decoding seems nearly impossible. At this stage it is a useful model to examine the development of the organisms to understand how the system is slowly built up from its various parts. This is where stem cell research and Systems Biology fit perfectly together.

The term *stem cells* was first introduced in 1909 by the Russian histologist Maximov on a congress in Berlin (Maximov, 1909). He characterized hematopoetic stem cells as the common ancestor of all blood cells. A similar definition still holds today. Stem cells are characterized as a class of cells that have two special properties in common: (i) they are able to proliferate infinitely by mitosis and (ii) they can differentiate into different specialized cell types.

Since the discovery of a special class of stem cells and its extraction from mouse embryos in the 1980s (Evans and Kaufman, 1981) stem cells are divided into adult- and embryonic stem cells (ES). In 1998, the first human ES cell line was derived by Thomson (Thomson, 1998) who proposed the following definition for ES cells:

> (...), we proposed that the essential characteristics of primate ES cells should include (i) derivation from the preimplantation or periimplantation embryo, (ii) prolonged undifferentiated proliferation, and (iii) stable developmental potential to form derivatives of all three embryonic germ layers even after prolonged culture.

ES cells are derived from the early stages of embryonic development. After fertilization, the zygote starts to divide and form a morula. After a number of division this "lump" of cells is transformed into a sphere, called the blastocyst (Figure 4.1.1). In the early blastocyst, the cells have already differentiated into trophectoderm cells that form the extra embryonic tissue, and the inner cell mass (ICM) that later differentiates into the three somatic lineages. The ICM from these blastocysts is used to derive ES cells in culture.

Compared to adult stem cells which are limited in their differentiation capabilities to certain lineages, ES cells have the great advantage to be able to differentiate into all cells of the body. In general stem cells are classified by their developmental capabilities into the following hierarchy of potential:

Totipotent Cells that are able to produce a full viable embryo, including extra embryonic tissues. A potency of this level is only observed in the fused egg- and sperm cells and a few divisions after that.

Pluripotent Cells that are able to form all body cells. Examples for these are ES cells and the artificially produced induced pluripotent stem cells (iPS) (Sec. 4.1.1).

Multipotent Cells that are found in the adult organism and can differentiate into a number of cells from different lineages. A good example are the hematopoetic stem cells mentioned above.

Oligopotent Progenitor cells which are able to differentiate into a small number of closely related cells.

Unipotent Progenitors (also called precursor cells) that can only differentiate into one cell type, for example hepatocytes in the liver.

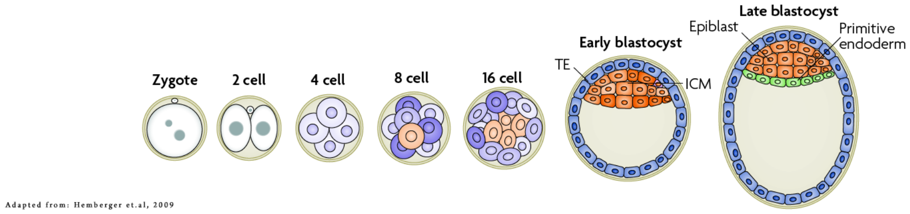


Figure 4.1.1: Development of germ layers from the zygote over the early and late blastocyst. ES cells are extracted from the inner cell mass in the late blastocyst stage.

These definitions from cell biology are reflected by molecular differences in DNA structure and gene expression between these cell states. The potency of a cell seems to be strongly coupled to the general structure of epigenetic marks on the chromatin like DNA methylation and histone modifications (Laurent et al., 2010; Lister et al., 2009, 2011). These epigenetic changes are governed by gene expression and protein levels in the cell and vice versa, giving a perfect example of the need for a Systems Biology approach combining different levels of regulation. The model we present in section 4.2.2 is based on these effects to a large extent. The feedback from the expressed proteins on the expression of genes becomes strikingly obvious when looking at the first techniques for cellular reprogramming. This technique is called somatic cell nuclear transfer (SCNT) and is also used for reproductive cloning. The somatic nucleus is transferred into an enucleated oocyte, and thereby forms an ES cell by restructuring its chromatin to the pluripotent state (Hochedlinger and Jaenisch, 2002; Rideout, Eggan, and Jaenisch, 2001). When transplanted into a surrogate mother this oocyte can also give rise to a genetically equivalent animal (clone). This is a great example to show that proteins present in the cytoplasm are sufficient to alter gene expression and also epigenetic regulation in the nucleus radically.

The epigenetic landscape

The definition of a differentiation potential of cells is strongly influenced by the vivid image of a so called epigenetic landscape first described by Waddington (Waddington, 1940, 1957). He depicted the process of development as a ball rolling around in a sloping landscape full of hills and valleys. In the course of development cells take different paths guided by the shape of the surface and end up in different valleys in the end. Spontaneous dedifferentiation is prevented by "gravity" keeping the cell in the valley it went into. His view was that there are epigenetic barriers keeping cells in their valley (lineage), but given strong enough perturbation cells are able to leave their valley again. This visual description of the differentiation process is surprisingly similar to the modern view of biological processes as complex dynamical systems and their attractors and critical points (Figure 4.1.2). The attractor can be seen as a "balanced" (steady) state of a dynamical system in which, while unperturbed, all forces are in balance and the state does not change (Section 2.3). Pluripotent cells like ES cells are in a balanced, but unstable state, which can be left due to minor perturbations, while fully differentiated cells are in a stable state that can only be left by major perturbations of the whole system. To provide a holistic view on this epigenetic landscape, I have developed a model that shows these differences between the stability of cell states (Section 4.2.2).

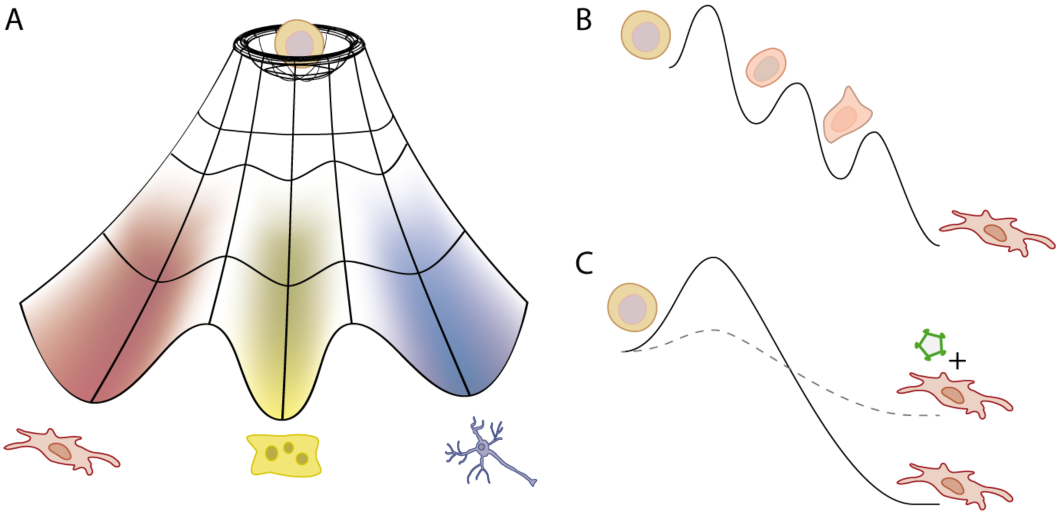


Figure 4.1.2: (**A**) An example of an epigenetic landscape showing the different stabilities of a state depending on its location. (**B**) Differentiation shown as cross-section of a path through the epigenetic landscape. Different progenitor cells can be sustained in local minima of the landscape. (**C**) Changes in the epigenetic landscape are produced by the Yamanaka factors and the difference in potential between differentiated and pluripotent cells becomes less pronounced.

Pluripotent cells, like ES cells, are the most universal cells that can be cultured in the lab. They hold great promise for a number of medical applications and their study has become a large research field. However, working with these cells poses ethical problems due to their embryonic origin, at least for human cells. Their discovery has led to a still ongoing public debate about the ethic implications of their use, as embryos have to be destroyed to harvest the cells from ICM and cultivate them. Due to these problems, research in the field is difficult and highly regulated in many countries.

Aside from the ethical problems, the clinical use of stem cells is limited to the donor himself due to immune system reactions. Until recently, there was no way to derive pluripotent stem cells for the adult patient. This changed with the discovery of iPS cells in 2006 by Takahashi and Yamanaka, who derived pluripotent cells that could contribute to all three germ layers from mouse embryonic fibroblasts. Only one year later the same process of somatic cell reprogramming (SCR) was completed with human fibroblasts (Takahashi et al., 2007). This surprisingly simple process is only based on the viral induction of the same four transcription factors in mouse and human. These factors, Oct4, Sox2, Klf4, and cMyc are sometimes also referred to as the Yamanaka factors. Since their discovery, iPS cells have been derived from a large number of different somatic cell types using the same cocktail or variations and subsets of it (Okita et al., 2008; Zhou et al., 2009). The genes used in the reprogramming cocktail are all master regulators of transcription that have a lot of targets throughout the genome. They all influence many targets and their target sets have large overlaps (Boyer et al., 2005).

Around these factors there is a large gene regulatory network influencing cell state and potential that is only partly known (Kim et al., 2008). Oct4 is at the center of this network and has been studied as a stem cell factor long before the discovery of iPS cells. Together with Nanog, another important pluripotency factor, it forms the core pluripotency gene network. This network, when activated, keeps cells in a pluripotent state. To enable fully differentiated cells to return to this state one needs to activate other factors as well. To reprogram partially differentiated progenitor cells to pluripotency it has been shown that the induction of only Oct4 is sufficient if certain small molecules are added to the medium. Also all the other factors except Oct4 can be replaced in the process by using small molecules (Huangfu et al., 2008a,b; Mikkelsen et al., 2008; Shi et al., 2008). Another interesting development around more efficient reprogramming was the generation of a mouse with stable integrated Yamanka factors that are inducible by administering doxycycline (DOX). These secondary iPS systems have a much higher reprogramming efficiency of about 2%, compared to 0.01% of the standard approach.

Pluripotency is regulated on various levels

Although direct reprogramming enabled us to study the interplay of the networks regulating pluripotency in a defined environment, it is still not understood how the transition happens in detail. However, it has become clear that the reprogramming potential is not limited to specific cells in a culture, but rather that essentially every cell can be reprogrammed given enough time and the appropriate method (Hanna et al., 2009). A high proliferation rate seems to be beneficial to the process of overcoming the barriers in reprogramming (Hong et al., 2009; Kawamura et al., 2009; Marión et al., 2009). As mentioned above, efficiency could be improved by the addition of small molecules (Wang and Adjaye, 2010), some of which are also capable of replacing KLF4 and cMYC or even SOX2 (Ichida et al., 2009) in the process. Most of these small molecules act on the epigenetic modifications that fix the cells in their current developmental state. One of the most prominent drugs improving reprogramming is the histone deacetylase 1 (HDAC1) inhibitor valproic acid (VPA) (Huangfu et al., 2008a). The inhibition of HDAC1 seems to lower the epigenetic barrier between the cell states and facilitates the transition from one state to the other.

Pluripotency in general is regulated by an interplay of different mechanisms that we will outline in detail in the following. First, transcriptional regulation, i.e. activation or inhibition of target gene activity by specific transcription factors, controls the expression of master regulators of pluripotency or differentiation. A second layer of control consists in DNA-methylation of promoters of genes. Finally, the activating or repressive modifications on histones represent the third mechanism (reviewed in Meissner, 2010).

The core transcriptional regulatory circuitry of pluripotency in human embryonic stem cells (hESCs) was first established by (Boyer et al., 2005) and contained the master regulators of pluripotency OCT4, SOX2, and NANOG. These three transcription factors were found to interact in a mutually- and auto-activating fashion thereby promoting and maintaining pluripotency (Boyer et al., 2005; Loh et al., 2006). This regulatory circuitry has been extended in further studies to yield different larger networks regulating pluripotency (Chavez et al., 2009a; Ivanova et al., 2006; Zhou, Chipperfield, and Melton, 2007).

DNA-methylation of regulatory sequences, which silences gene promoters, is one of the known mechanisms in epigenetic regulation. This methylation is a major hindrance in reprogramming, because methylation marks cannot easily be removed, although there is evidence for active demethylation in reprogramming cells (Bhutani, Burns, and Blau, 2011), which we will further discuss below.

With the advent of next generation sequencing techniques there is a wealth of data accumulating on DNA-methylations ("methylomes") in different cell types (Laurent et al., 2010; Lister et al., 2009, 2011). These studies reported large differences between ES/iPS and differentiated cells in the methylation states of promoters of key pluripotency and developmental genes. Moreover, they identified a very slow reprogramming of methylation states with aberrant methylation persisting in reprogramming cells, which can thus be distinguished from fully reprogrammed or ES cells. These remaining DNA-methylation differences also limit the differentiation potential of the iPS cells and restrict their applications. A recent study also reported the occurrence of newly methylated aberrant sites that did neither occur in the source nor in the target (ES) cells (Nishino et al., 2011).

Comparative studies were not limited to DNA-methylation. Histone modifications were also studied extensively, suggesting a close connection between DNA-methylation and histone modifications (Hawkins et al., 2010). It has been found that there is a strong correlation between gene silencing histone modifications and DNA methylations in promoters of pluripotency regulators (Cedar and Bergman, 2009). However, the relationship between the two is still not fully understood. The connection is probably established by histone binding proteins such as G9a, which have histone methylation activity (HMT) and therefore facilitate the formation of heterochromatin. G9a can also recruit the *de novo* methyl transferases DNMT3A and DNMT3B to the nucleosome which in turn can methylate the gene promoters on the DNA. DNA-methylation is thought to stabilize chromatin structure during mitosis through differential binding of proteins for closed or open chromatin (Cedar and Bergman, 2009) and it can also inhibit methylation of H3K4, an activating histone mark. Inheritance of histone modifications is coupled to the methylation pattern as it guides binding of certain HDACs (Fuks et al., 2000). DNA-methylation itself is sustained throughout DNA replication and mitosis by virtue of DNMT1 and other associated proteins like NP95 by copying the methylation pattern of the template strand to the copied strand. Though this process is quite efficient, methylations can be lost in rapidly dividing cells and cells lacking DNMT1 (Monk, Adams, and Rinaldi, 1991).

Existing models of pluripotency and reprogramming

The consequences of the complex interplay of the three mentioned regulatory mechanisms, i.e. transcriptional regulation, histone modifications leading to changes in chromatin structure, and DNA methylation, are not easy to understand. Mathematical modeling can help to unravel these complex interactions and explain how cellular behavior is linked to the molecular mechanisms. Since we are dealing with an enormously complex system, we need to reduce its complexity in order to discern the basic underlying features of the network. There have been various attempts to model certain parts of regulatory networks in great detail, which gave valuable insights into the dynamics of these subsystems (e.g. MacArthur, Please, and Oreffo, 2008).

All the above mentioned regulatory processes only work correctly in an orchestrated manner. Regulatory structures in stem cells have been described by various models using different modeling approaches. There is a number of detailed models describing the interplay of regulatory genes in pluripotency and reprogramming, which help to understand the gene networks in detail and have elucidated the bistability of decisions taken in development and the influence of expression noise (Chickarmane and Peterson, 2008; Chickarmane, Troein, and Nuber, 2006; Kalmar et al., 2009; MacArthur, Please, and Oreffo, 2008). These models use ordinary differential equations to show the dynamics inside a small part of the whole machinery. There are also many studies describing regulation of differentiation into different lineages (Duff et al., 2012; Huang et al., 2007; Roeder and Glauche, 2006; Wang et al., 2010). Bigger networks were just recently modeled using dynamic Bayesian networks and were used to predict improved reprogramming factor combinations (Chang, Shoemaker, and Wang, 2011).

A second class of more coarse grained models deals with transitions between cell states and how they are shaped by self-organizing systems in the cells (Halley, Burden, and Winkler, 2009; Qu and Ortoleva, 2008). These models are very conceptual and refrain from describing single gene interactions. There have also been efforts to characterize the processes in chromatin remodeling in a theoretical model, which showed that there must be a positive feedback in the formation of heterochromatin structure to explain its observed behavior (Dodd et al., 2007).

Looking at the experimental evidence in the literature it seems that the progression of reprogramming is governed by stochastic processes that prohibit or permit activation of pluripotency genes. For that reason, there have also been attempts to model it with noisy ordinary differential equations (MacArthur, Please, and Oreffo, 2008) or even as a stochastic process of state transitions (Hanna et al., 2009). In a more general approach Artyomov, Meissner, and Chakraborty (2010) explicitly modeled the space of cellular states as a binary tree with nodes for each cell state and the pluripotent state as the root of the tree. This study was the first to include gene regulation and epigenetic changes in one model and it could, among other things, explain the low efficiency of reprogramming.

4.2 Results

4.2.1 Gene Regulatory Networks that Govern Pluripotency

As a starting point for our work on stem cell gene regulation we used two networks that have been generated by the two different approaches mentioned in section 1.3.2, i.e. literature mining and high-throughput data. The first network was assembled by an extensive literature mining and database search for transcription factor binding of the Yamanaka factors and was presented in Mah et al. (2011). The second network was derived from ChIP-on-chip experiments for a number of pluripotency factors by Boyer et al. (2005) combined with some activation/inactivation information from Chavez et al. (2009b). Henceforth we will denote these as Mah network and Boyer network. The two networks are completely different in their focus and we used them for different purposes. The Boyer network is an exhaustive listing of all genes with promoters binding one of the transcription factors in the study (topology partly shown in figure 4.2.8), whereas the Mah network is a much smaller network centered around the Yamanaka factors and all their interactions, including signaling pathways (Figure 4.2.5A).

The second piece of information that was a prerequisite for the analyses in this section, was a set of microarray measurements gathered before, during, and after a reprogramming experiment with human embryonic fibroblasts and the Yamanaka standard procedure (Figure 4.2.1). This dataset was generated by transduction of human foreskin fibroblast cells (HFF1) with retroviral constructs and culturing them on a feeding layer of mouse embryonic fibroblasts. we took samples from the cells before transduction and at three early time points afterwards (24h, 48h, and 72h). These samples were hybridized on an Illumina whole genome microarray and also compared to HFF1 derived iPS cells and ES cell lines. The cells showed a stable expression of the exogenous Yamanaka factors already after 24h, although some of these were not captured on the array.

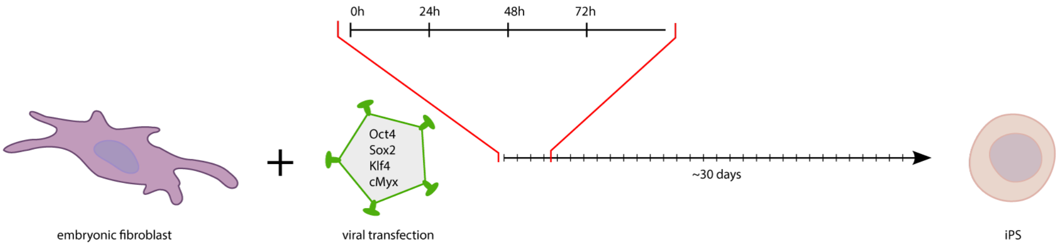


Figure 4.2.1: Experimental procedure for gathering reprogramming timecourse data. Embryonic fibroblast cells were transduced with the reprogramming factors and expression profiles by microarrays were taken after 0h, 24h, 48h, and 72h. Additionally profiles were measured for reprogrammed iPS cells from this experiment and H1 cells as a comparison.

The basic analysis of raw data was done by Nancy Mah using standard Bioconductor (Gentleman et al., 2004) tools. As expected the expression changes get stronger over time and the largest set of differential expression was found in iPS cells (Figure B.0.1).

Functional Analysis of Expression Data

For a further analysis of the timecourse data we only considered probes that showed a p-value < 0.05 for differential expression in at least one of the time points (2636 of 13708 probes). To get a separation of different dynamics of the genes we performed a fuzzy-c-means clustering (Section 2.2.1) of these genes with nine cluster centers (Figure 4.2.2). The clustering showed different classes of dynamics. Most showed a consistent pattern during the early phase of reprogramming with either constantly declining or increasing abundances. As expected, differences to the reprogrammed iPS state or the ES (H1 cells) state were the biggest in every class, but interestingly the trend of the early stages was sometimes reversed (e.g. cluster 3 and 4) and there were strong differences between iPS and ES cells for some classes (e.g. cluster 9).

To find out which genes were grouped into the clusters, we conducted a Gene Ontology (Ashburner et al., 2000) enrichment analysis for biological processes by comparing the genes belonging to each cluster to the whole set of genes present on the chip. We used a hypergeometric test to screen for significantly over-represented GO-terms in the clusters

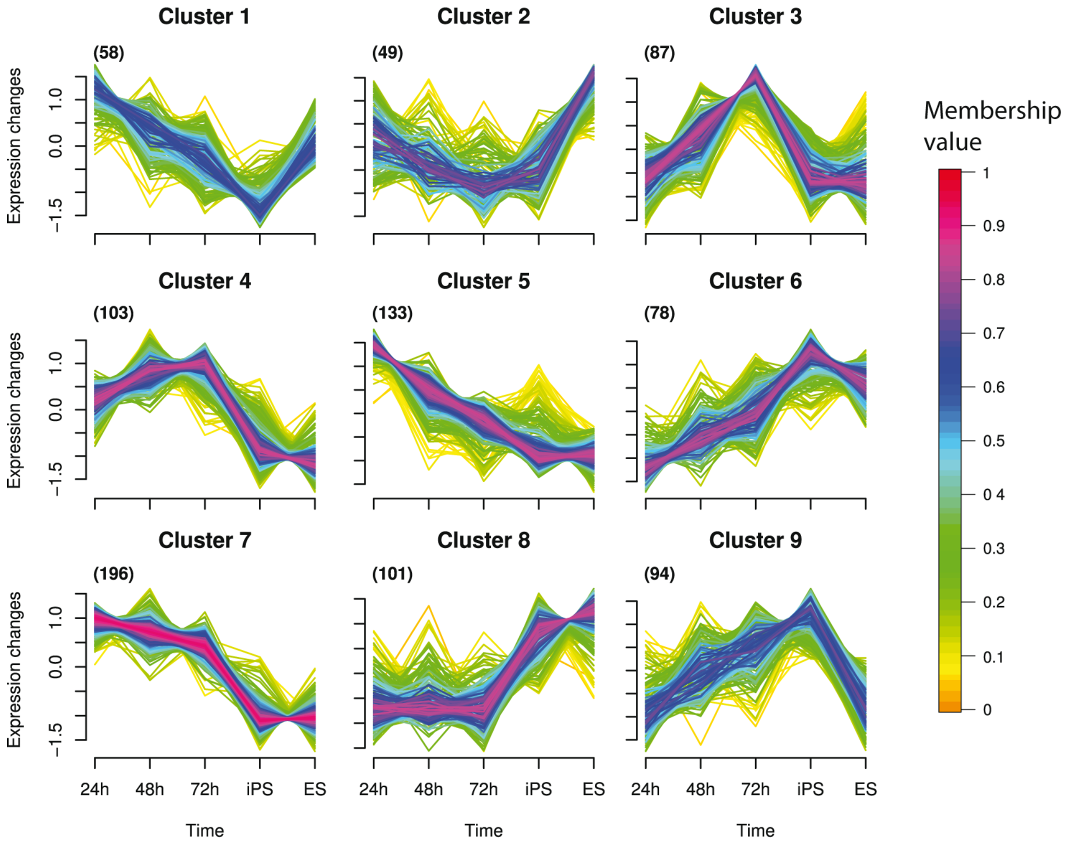


Figure 4.2.2: Clustering of timecourse data. I clustered all differentially expressed genes ($p < 0.05$) by their expression profile over time using the fuzzy-c-means method. Expression profiles of genes are color-coded by their membership values for the cluster (Section 2.2.1). Cluster core size ($a > 0.5$) given in brackets for each cluster.

using a correction for the tree structure of the GO-terms (Falcon and Gentleman, 2007) (Section 2.2.2). This analysis showed that there was no enrichment for developmental genes in any of the classes and that the main changes in the dataset laid in immune response and mitosis. Cell division differences were most significantly enriched in cluster 2 that showed strong differences between iPS and ES cells, suggesting that ES cells are much more potent in proliferation than the reprogrammed iPS. Cluster 9 showed an opposite behavior and is enriched with many ribonucleotide biosynthetic processes, suggesting a fast metabolism in the iPS cells compared to the other cell lines.

Cluster 3 and 4 are interesting, because they show a completely different trend in the beginning than in iPS and ES cells. They show a strong to moderate activation in the beginning of reprogramming and much less expression in the stable iPS and ES cells. Both clusters are enriched with apoptotic and immune system genes. The early activated Cluster 7 is related to immune and anti viral response even clearer with "response to virus" being the first ranking GO term. This cluster also contains genes responsible for cell-matrix-adhesion. Cluster 5, which was constantly declining in expression, shows an enrichment in substrate junction assembly and actin related processes, anticipating the results of the next analysis.

A Shift of Pathway Activation Happens Early in Reprogramming Cells

To complement the GO enrichment analysis, we also searched for pathways that were activated or inhibited in the early time-steps of reprogramming. The pathway analysis over all KEGG (Kanehisa, 2000) signaling pathways was performed for each time point on all differentially expressed genes (24h:294 genes, 48h: 1299 genes, 72h: 1901 genes). For the analysis we used the Bioconductor package SPIA (Tarca et al., 2009). This analysis combines a standard enrichment analysis in signaling pathways with an analysis of the impact of the regulations on the pathway, which makes it more specific and sensitive as simple enrichment testing (Section 2.2.2). The analysis showed a clear picture of pathways that are influenced by the transcriptional changes during viral infection and the beginning of reprogramming (Figure 4.2.3).

After 24h there are only pathways impacted and activated that are associated with immune response and various viral diseases (Table 4.2.1). This is not surprising, because the cells are transduced with the Yamanaka factors using retroviral vectors. Already after 48h the Focal adhesion pathway is significantly inhibited while viral response is still active. The viral response is loosing significance in the 72h time point while focal adhesion is strongly significant at 72h as well as in the iPS state. The focal adhesion pathway is responsible for the formation of cell-matrix adhesion points termed focal adhesion, where bundles of actin filaments are anchored to trans-membrane receptors of the integrin family. Integrin signaling events culminate in reorganization of the actin cytoskeleton; a prerequisite for changes in cell shape and motility, as well as gene expression. The early inactivation of this pathway shows a fast activation of morphology changes and downregulation of cell-cell

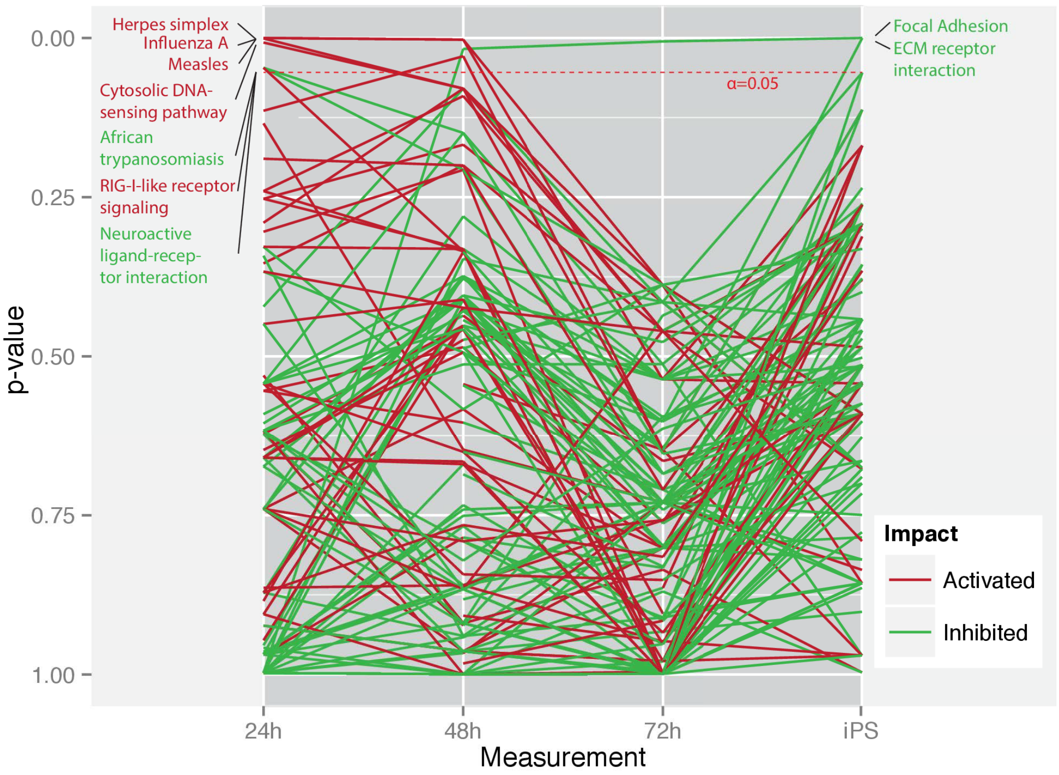


Figure 4.2.3: Pathways returned by SPIA analysis for all time points. The y-axis shows the p-values corrected for false discovery rate. Only pathways annotated with their names are significantly enriched in any time point. In the first time points these are primarily viral response pathways, which are all activated. After 72h this reaction is already superseded by inhibition of cell-cell-adhesion and communication (Table 4.2.1).

and cell-matrix-contacts. In the iPS state there is also a significant inhibition of the ECM receptor interactions, which serve an important function in organ morphogenesis and cell differentiation. The evidence for downregulation of the focal adhesion pathway is really strong. Looking at the regulation of pathway proteins over time, our experiment shows a clear downregulation for a large part of the pathway proteins on all levels (Figure 4.2.4).

Table 4.2.1: Top three pathways that were enriched in the differentially expressed genes at each time point.

Pathway	p-value	Impact
24h		
Herpes simplex infection	3.04e-06	+
Influenza A	5.15e-06	+
Measles	5.15e-06	+
48h		
Measles	2.58e-03	+
Influenza A	2.58e-03	+
Focal adhesion	1.71e-02	-
72h		
Focal adhesion	5.45e-03	-
Measles	3.87e-01	+
Pathways in cancer	3.87e-01	-
iPS		
Focal adhesion	3.46e-07	-
ECM-receptor interaction	5.39e-05	-
Type II diabetes mellitus	5.43e-02	-

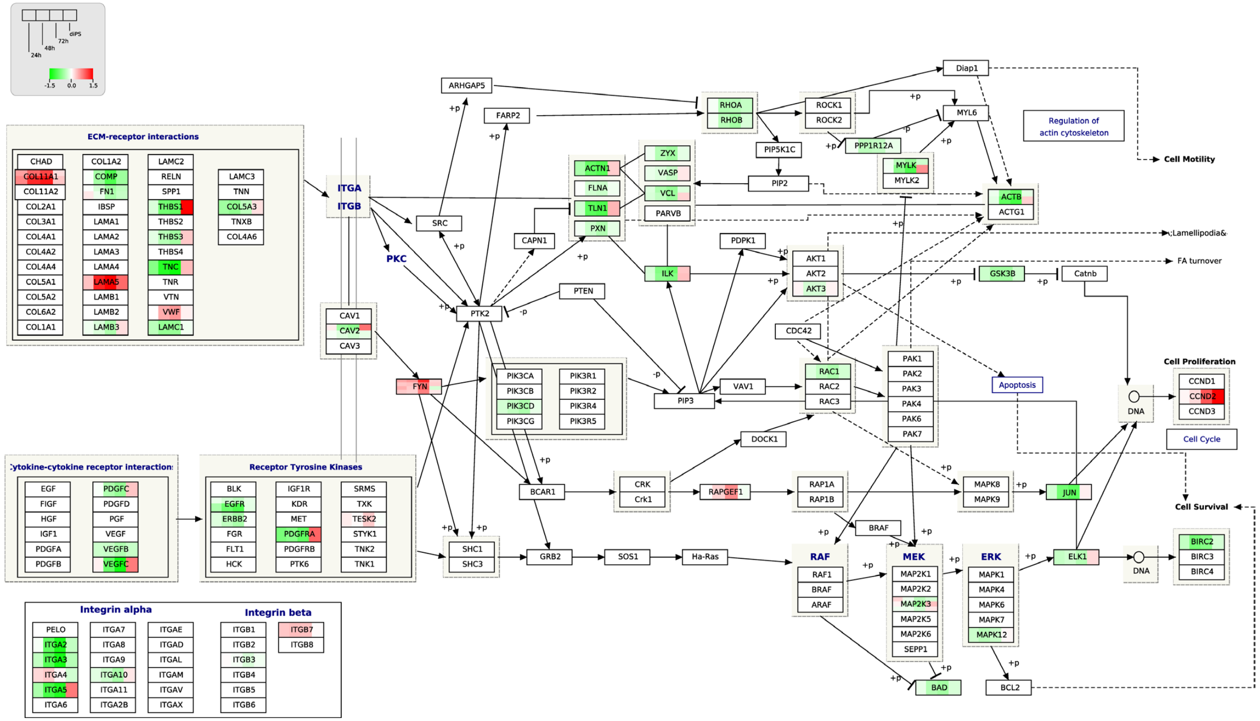


Figure 4.2.4: KEGG representation of the focal adhesion pathway with projected expression data for the genes that were present in our dataset. Each present gene is divided into four blocks colored representing its expression changes in the measured time points. It becomes clear that regulation happens on all levels of the pathway, but especially on the receptor level and regulation of the actin skeleton.

The focal adhesion pathway and the ECM receptors are usually active in HFF1 cells, because fibroblasts show strong matrix interactions. Its downregulation already after 48h shows an early step in reprogramming that has not been reported before. The transition from a mesenchymal to an epithelial cell type (MET) plays a major role in SCR. ES cells are epithelial cells that undergo epithelial-mesenchymal-transition (EMT) during differentiation to the mesenchymal fibroblast lineage. This process needs to be reversed in SCR and is then called MET. One of the major differences of these cells types is their attachment to their environment. While mesenchymal cells are strongly connected to the extracellular matrix, epithelial cells are directly bound to their neighboring cells to form a dense layer. Another hint towards the activation of MET is that we see a downregulation of N-Cadherin (*CDH2*) within the first three days, which is proposed to be a functional switch between focal adhesion and cell-cell-adhesion (Lehembre et al., 2008). The analysis shows a clear shift away from matrix associated mesenchymal fibroblasts by inhibition of the focal adhesion pathway.

Mapping Timecourse Data onto the Network

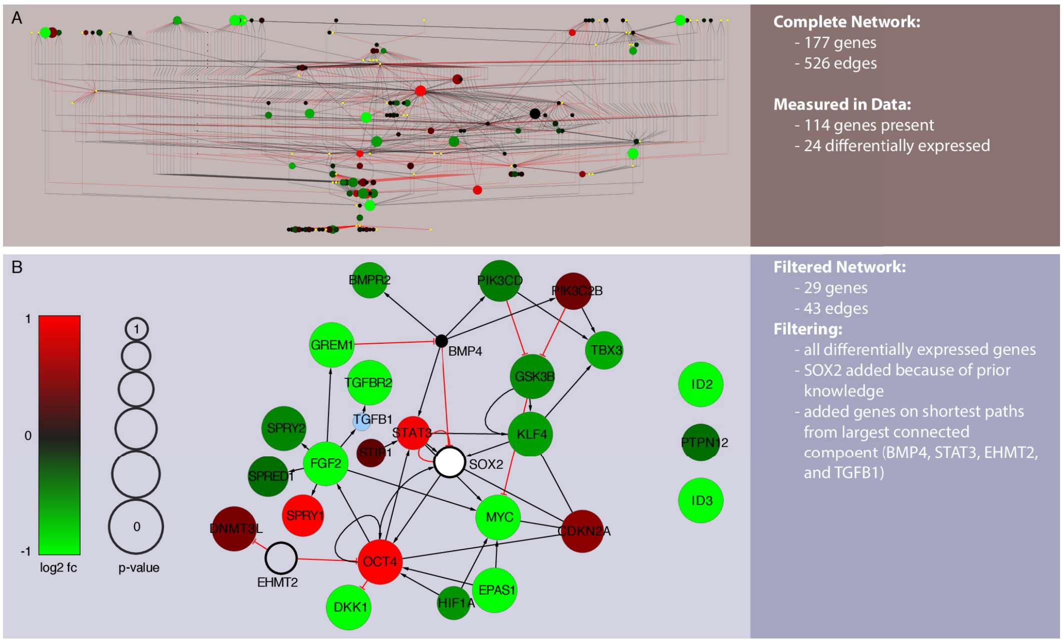
To combine the data measured in the experiment (Figure 4.2.1) with the literature knowledge in the Mah network, we used the timecourse of expression profiles to filter out a more specific network (Figure 4.2.5).

First we filtered for genes and gene products, as nodes in the literature network represent different kinds of biological entities and also include small molecules, extra-cellular signals, etc. In cases where the literature was not completely clear which gene was referred to (i.e. use of common names), we used databases to identify all possible genes and treated them as different nodes with the connections pointed out in the reference network.

Every gene from the original network that showed a differential expression (p-value < 0.05) in either 24h, 48h, or 72h was used to produce a network of the most important genes during the beginning of reprogramming. We furthermore added SOX2 to the network, as it was one of the exogenous factors. It was not measured correctly by the chip, because the probe on the chip matches an untranscribed region that is not transduced by the vector. This processing lead to a network consisting of 24 differentially expressed genes, which had one large connected component and 7 unconnected nodes.

To transform the network into one large connected component, we added the 4 genes (BMP4, STAT3, EHMT2, and TGFB1) that were on the shortest paths between the unconnected nodes and the connected component. We did not include additional nodes if the original nodes had a distance greater than one from the connected component, which leaves 3 genes unconnected (ID2, ID3, and PTPN12) (Figure 4.2.5 B).

We assembled a Boolean model from this resulting network, to be able to explain parts of the data and show inconsistencies in the network. The aim of this work was to find the update functions that would produce a Boolean simulation consistent with the data for the topology given by literature data. When examining the network structure with data



Complete Network:
- 177 genes
- 526 edges

Measured in Data:
- 114 genes present
- 24 differentially expressed

Filtered Network:
- 29 genes
- 43 edges

Filtering:
- all differentially expressed genes
- SOX2 added because of prior knowledge
- added genes on shortest paths from largest connected compoent (BMP4, STAT3, EHMT2, and TGFB1)

Figure 4.2.5: Visualization of the networks, color of the nodes according to their expression changes at 72h and setting node size relative to their p-values for differential expression. (A) The complete Mah network with the reprogramming data mapped onto it. (B) The filtered model of all differentially expressed genes in one of the time points during reprogramming.

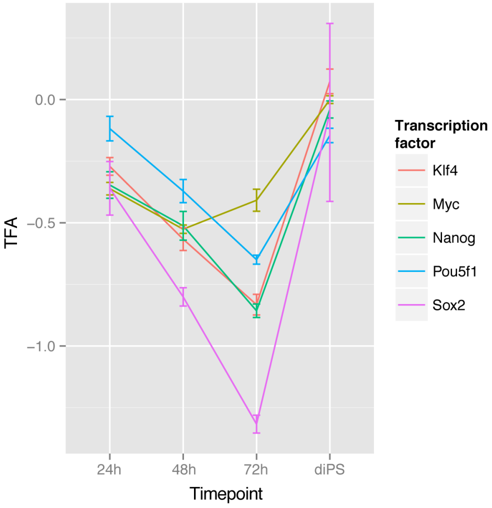


Figure 4.2.6: Transcription factor activation profiles as estimated from the expression data and transcription factor binding data. The three core factors Oct4, Sox2, and Nanog show different dynamics from cMyc, which is already more active after 72h, whereas the other factors are inhibited during the early reprogramming phases.

mapped onto the nodes, it became obvious that the discrepancy between data and network structure was too big to make this approach feasible.

To understand why the approach was not working, we highlighted the differences between the observed expression changes in the Yamanaka factors — and other pluripotency factors in the beginning of reprogramming — and their actually observed effects. We did this using the Boyer network in combination with the timecourse data.

Estimating Influences on target genes

In the experiment we observed protein expression of the four exogenous Yamanaka factors already after 24h post transduction (Mah et al., 2011). One of the big questions in reprogramming is what the roadblocks are that hinder greater efficiency and faster reprogramming. If the factors are already there after 24h why does reprogramming happen so slowly? We investigated which changes one could observe in the target genes and how they relate to the activity of the single factors. To be able to see the direct effects of the four Yamanaka factors over the course of the experiment, we used the network component analysis (NCA) method described in section 2.2.4. NCA is a method to calculate transcription factor activation (TFA) and connectivity strengths (CS) for a set of transcription factors and their targets. As described above, we first combined the transcription factor binding data with the computationally derived activation and inactivation data to get a network with some directed edges better suited for NCA (Boyer et al., 2005; Chavez et al., 2009b). From this large network we only used the Yamanaka factors and Nanog and all their target genes as input for the NCA.

The estimated activity of the Yamanaka factors did not resemble the measured expression patterns (Figure 4.2.6) at all. The mRNAs were present, the proteins were expressed, but the expression of the target genes did not change accordingly (Mah et al., 2011). Only in the iPS cells, where we saw an endogenous activation of the Yamanaka genes, the TFA profiles showed higher levels. Directly after transduction the TFAs even declined. This result suggests the presence of epigenetic regulation prohibiting the action of the Yamanaka factors in the beginning of reprogramming on many genes they would act on in ES cells.

Driven by these results, we asked the question which genes are influenced in the beginning, and which are not. There is a second outcome of NCA that can partly answer this question. The connectivity strengths define how strongly a gene is influenced by a certain transcription factor. Combined with the TFAs, the connectivity strengths of all acting TFs predict the behavior of a gene (Section 2.2.4). These important factors are shown in Figures 4.2.7 and 4.2.8.

The heatmap in Figure 4.2.8 shows the similarities between the connectivity strengths on the different genes by a hierarchical clustering. The similarity of the four core pluripotency factors is clearly visible, as they influence many genes in the same fashion, whereas cMyc and Klf4 show opposite behavior in many cases.

4.2.2 A Probabilistic Boolean Model of Somatic Reprogramming and Differentiation

Somatic cell reprogramming has dramatically changed stem cell research in recent years. As obvious from the results in the previous sections, the many levels of regulation make it a challenge to isolate core principles of the process. In order to analyze such mechanisms, I developed an abstract but mechanistic model of a subset of the known regulatory processes during cell differentiation and production of induced pluripotent stem cells.

This probabilistic Boolean network (Section 2.3.1) describes the interplay between gene expression, histone modifications, and DNA-methylation. The model incorporates recent findings in epigenetics and reproduces experimentally observed reprogramming efficiencies and changes in DNA-methylation and chromatin remodeling. Using the model simulations, I could investigate how the temporal progression of the process is regulated. Guided by the results of the previous section that clearly highlight the importance of cellular responses to the delivery method, the model also explicitly includes the transduction of factors using viral vectors and their silencing in reprogrammed cells. Viral transduction is still a standard procedure in somatic cell reprogramming and it was also used in the experiments for the previous analysis. Based on the model, we calculate probability landscapes of cell states for different starting conditions.

The model differs from the existing models of pluripotency described in section 4.1.1, because it is an abstract representation of the combined networks that govern pluripotency and reprogramming using well established modeling frameworks in a novel way (Figure 4.2.9). The model is based on a standard Boolean networks approach comparable to the one used in section 3.3. This has the advantage that it can easily be modified and combined with other results.

Boolean models have the convenient property that a cell state is defined as a binary vec-

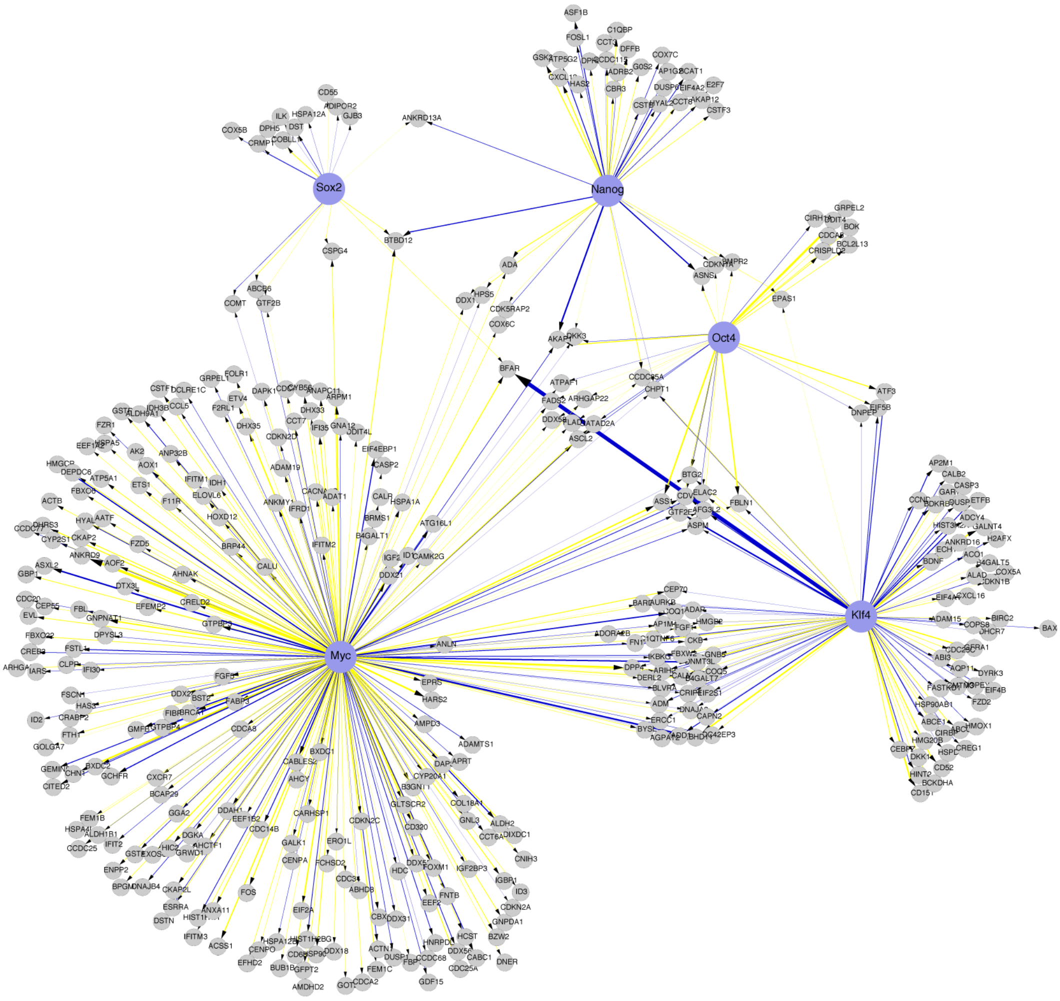


Figure 4.2.7: A network of transcription factors Oct4, Sox2, Nanog, Klf4, and cMyc generated from ChIP-on-chip data from Boyer et al. (2005) and a computationally generated network from Chavez et al. (2009a). The network was combined with the dynamic expression data from Mah et al. (2011) to estimate the control strengths of the factors on their target genes. The corresponding TFA profiles can be found in Figure 4.2.6

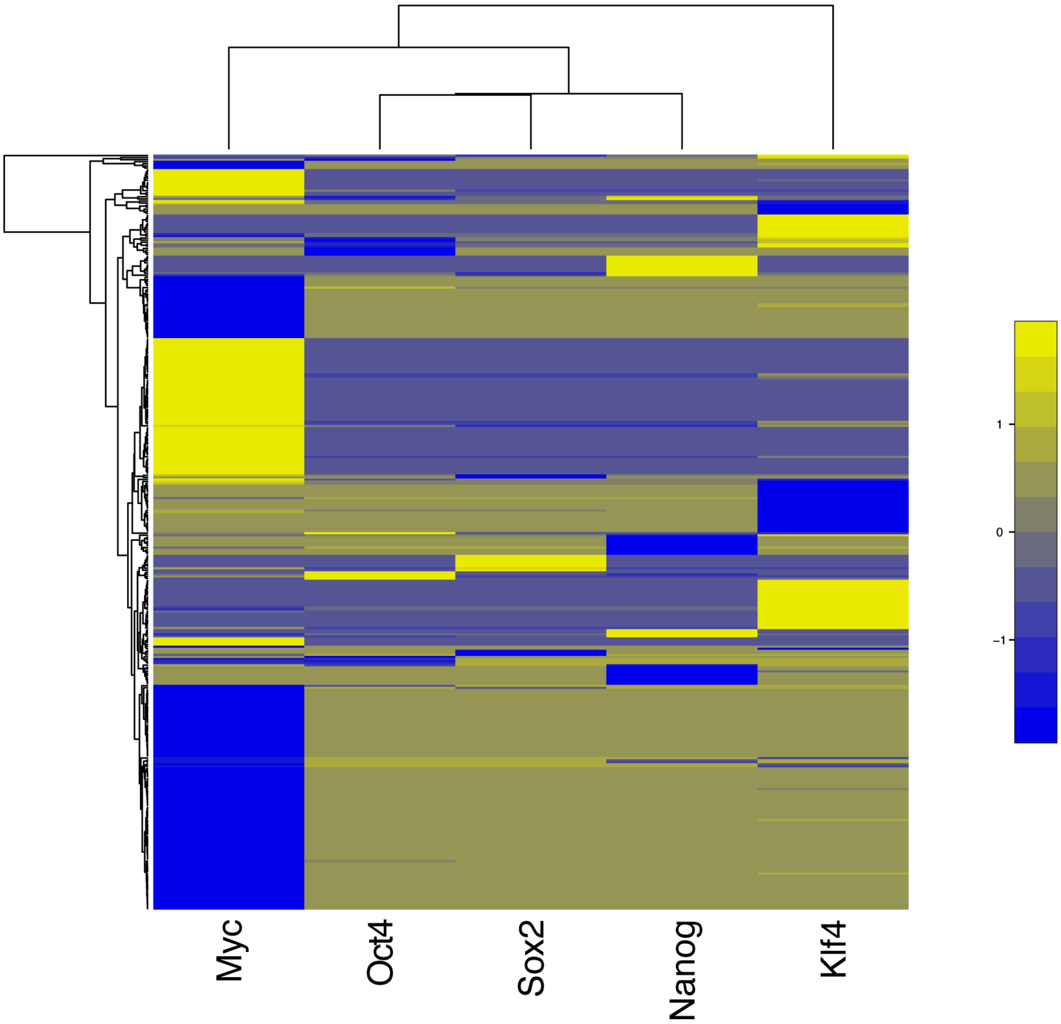


Figure 4.2.8: The connectivity strengths of the reprogramming factors on their target genes. Again the core pluripotency factors cluster together and exert seemingly similar influences on the majority of their targets.

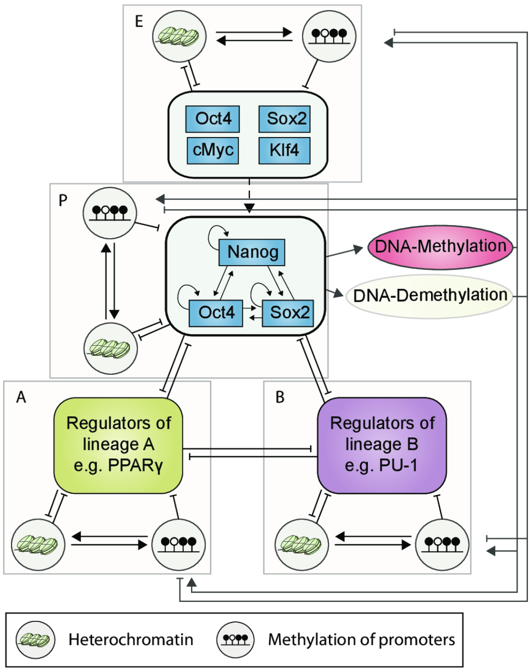


Figure 4.2.9: General model structure. The model consists of sub-modules representing classes of transcriptional master regulators. The full model has four modules: Two differentiated cell lines A and B, the pluripotency module P, and the exogenous factors E. These inhibit each other and thereby create a bistability with different steady states.

tor of the states of all variables, making it easy to compare states without further complicated definitions. Since the processes the model should elucidate are non-deterministic, we chose a probabilistic approach. The model is also much smaller compared to the networks presented in the previous sections. It only has 14 species that are simulated, whereas the smallest network derived from the Mah network has 27 nodes. The exact model structure will be derived in Section 4.2.2.

Simulation results show good reproduction of experimental observations during reprogramming, despite the simple structure of the model (Section 4.2.2). An extensive analysis and introduced variations hint towards possibilities for optimization of the process that could push the technique closer to clinical applications.

Model Structure

With this model we analyze the interplay of three different regulatory layers, as we include histone modifications, DNA-methylation, and transcription factor DNA interaction. Due to the different properties of these mechanisms it needs to keep a fairly high level of abstraction to combine them in one simple model (Figure 4.2.9).

For the sake of simplicity, we combine the single genes and regulatory factors that are responsible for the activation of a certain cell state into modules. This can be justified by the strongly correlated behavior of these genes (Berg et al., 2010) and their large number of shared targets (Kim et al., 2008). It has also been used in other theoretical models before and has proven to be successful (e.g Artyomov, Meissner, and Chakraborty, 2010).

The modules contain many activating interactions between their members. A good example is the network of core pluripotency transcription factors *OCT4*, *SOX2*, and *NANOG* that is responsible for sustaining pluripotency. These transcription factors bind a large number of shared targets as well as their own promoters. This leads to their mutual and auto-activation (Boyer et al., 2005). Similar interactions have been reported for master regulators of differentiated cell lines like PU-1 for erythrocytes (Nishimura et al., 2000; Okuno et al., 2005) or PPARγ for adipose tissue (Wu et al., 1999).

Interactions between these modules are often mutually repressive, as it was reported for GATA-1 and PU-1 (Rekhtman et al., 1999). The pluripotency module also represses differentiation factors. This mutual antagonism paired with auto-activation of the single modules is the basic structure of the transcription factor regulations in my model.

Basic biological findings underlying the logical rules of the model are summarized in table B.0.13. The epigenetic regulations that influence the expression level in general and specifically for each module are described in detail in the following.

As explained in the introduction of this chapter, there has been immense progress in the field of epigenetics in recent years. Nevertheless many of the regulatory mechanisms and their interactions are still enigmatic to researchers (Cedar and Bergman, 2009; Djuric and Ellis, 2010). In our model we explore different motifs of the epigenetic marks governing gene expression in development and reprogramming.

The general mechanism implemented in the model follows the approach suggested by Cedar and Bergman (2009) and others. Epigenetic dynamics emanate from the more rapidly changing states of the proteome of the cell. The expressed regulatory proteins and RNAs not only govern future expression profiles by direct action on promoters, but also change the more persistent epigenetic marks which then in turn define a new set of transcribed genes and, thus, of cellular proteins. In the model, the expression of genes that belong to the same module increases the chances of removing silencing marks on histones. Once the chromatin is in a less dense conformation by changes in histone modifications, we assume there is a possibility to remove DNA methylation if it is also suppressing gene expression in the module. The process of silencing can happen if the genes of a module are not expressed. The module is then prone to methylation and formation of repressing histone marks. If one of the silencing marks (either negative histone or DNA-methylation) is set, it increases the chances of keeping it and setting the second mark as well. As described above, histone modifications and DNA-methylations are strongly interconnected (Epsztejn-Litman et al., 2008; Thomson et al., 2010). This collaborative aspect of silencing creates a positive feedback loop, which promotes bimodality of the epigenetic states. Thus there is only a low probability to stay in a state where only one of the marks is set when the gene state is constant.

There is evidence for a co-regulation of DNA-methylation in genes with similar functions, which was found by comparison of whole genome methylation patterns between HFF1 and ES cells by bisulfite sequencing (Lister et al., 2011). Here, we assume that the

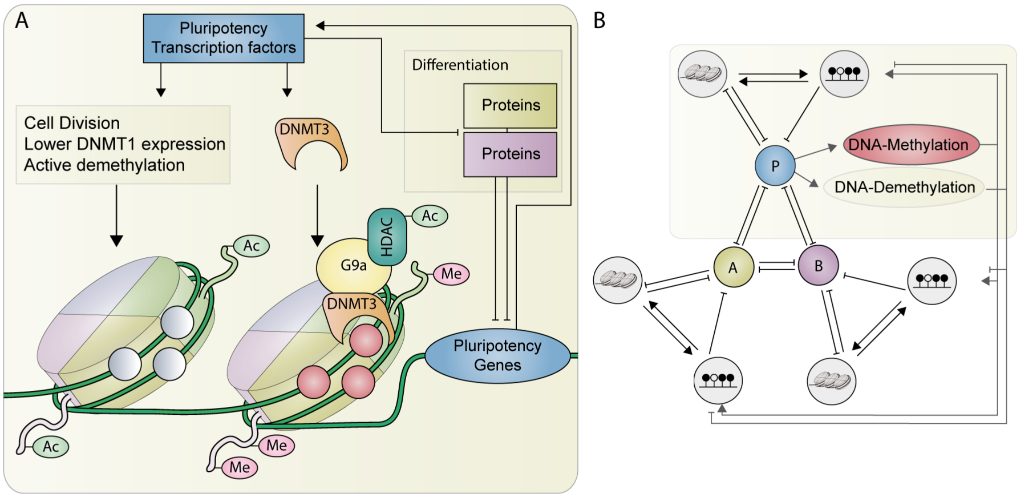


Figure 4.2.10: A schematic representation of the processes described by our model. **(A)** Connection between DNA-methylation, histone modifications, and the pluripotency master regulators. Pluripotency transcription factors activate their own expression and can be suppressed by factors regulating differentiation. The pluripotency factors themselves increase the expression of *DNMT3* which enables *de novo* methylation of DNA preferably in combination with repressive histone modifications such as methylation or deacetylation (right nucleosome). Additional activation of pluripotency genes also leads to a higher cell division rate, a suppression of methylation maintenance, and probably active demethylation, which also increases the chances of euchromatin formation. **(B)** Without external influences (e.g. retroviral genes or signaling molecules), the structure of our model consists of three gene modules (P,A,B) inhibiting each other and each governed by their specific epigenetic states. The pluripotency (P) module regulates the activation of methylation and demethylation.

DNA-methylations in the promoter regions of the genes in one module are co-regulated to a large degree, and are thus also characterized by one variable only. This variable follows rules which we derived from literature (Table B.0.13 and Figure 4.2.10). Activation of this variable means that the promoters are methylated, which leads to inhibition of gene expression. The activation of the DNA-methylation status is governed by the presence of *de novo* methyl transferases DNMT3A/B that are summarized in the variable *dnmt*. However, *dnmt* is not the only variable influencing the methylation state of a module. As described above, there are also other chromatin binding proteins influencing the likeliness of DNA methylation. We assume that all of these proteins combined are responsible for the current local chromatin structure and set the histone marks of the module as a modifying factor of the DNA methylation. This defines the activating update function that – if chosen – can only activate the variable.

If the DNA is already methylated, it can be demethylated by different mechanisms. For example, inefficiency of DNMT1 in copying methylation patterns is considered as passive demethylation (Monk, Adams, and Rinaldi, 1991). This process can only happen when cells are dividing, as it depends on DNA replication. However, there might as well be active demethylation processes influencing the DNA-methylation state as discovered recently (Table B.0.13). In our model, we summarize these processes leading to demethylation of DNA in the variable *demeth*. All of the mentioned processes happen very slowly compared to transcription factor mediated changes in the regulation of expression. Because of this, we also introduced a function that does not alter the variable if chosen. This function gets a high probability compared to the rest to guarantee that there are only slow changes in the variable over time. A combination of the above yields the following update functions for methylation of pluripotency genes:

$$\begin{aligned} m_m^A(t+1) &= m_m^A(t) \vee dnmt(t) \wedge m_{hc}^A \\ m_m^A(t+1) &= m_m^A(t) \wedge (demeth(t) \vee m_{hc}^A) \\ m_m^A(t+1) &= m_m^A(t) \wedge demeth(t) \\ m_m^A(t+1) &= m_m^A(t) \end{aligned} \tag{4.1}$$

where m_m^A and m_{hc}^A are the methylation and histone modification states of module A, respectively. Similar rules hold for modules B and P. Note that probabilities of the formulas must sum up to 1.

The *dnmt* and *demeth* variables are governed by the following rules:

$$\begin{aligned}
dnmt(t+1) &= m_e^P(t) \vee m_e^E(t) \\
dnmt(t+1) &= m_e^P(t) \vee m_e^E(t) \vee dnmt(t) \\
demeth(t+1) &= m_e^P(t) \vee m_e^E(t) \\
demeth(t+1) &= m_e^P(t) \vee m_e^E(t) \vee demeth(t)
\end{aligned} \tag{4.2}$$

where m_e^P and m_e^E are the expression of the pluripotency and the exogenous modules, respectively. Switching off these factors is very slow (small chances), because we assume that the influences implemented here are not the only impact on these variables and that they are active in many cell states.

Histone modifications are strongly simplified in our model. We consider neither single modifications on different sites nor different numbers of methyl groups on the residues. Chromatin changes are dependent on the expression of the module's genes. If these genes are expressed, it is impossible to remodel the chromatin to a closed form. If they are not present, there is a chance of negative histone modification, which is increased by present DNA-methylation marks. In Boolean formulas these processes are described as

$$\begin{aligned}
m_{hc}^A(t+1) &= m_{hc}^A(t) \vee m_m^A(t) \wedge \neg m_e^A(t) \\
m_{hc}^A(t+1) &= m_{hc}^A(t) \vee \neg m_e^A(t) \\
m_{hc}^A(t+1) &= m_{hc}^A(t) \wedge \neg m_m^A(t) \\
m_{hc}^A(t+1) &= m_{hc}^A(t)
\end{aligned} \tag{4.3}$$

where m_e^A is the expression of module A, m_{hc}^A the histone modification state, and m_m^A is the DNA-methylation of the module, respectively. Following these rules the DNA-methylation in a module increases the chance of forming and keeping heterochromatin independently of the chosen parameters. The same is true for the chances of DNA-methylation, which are dependent on the histone modification state of the module as well. Thus, the epigenetic states are mutually dependent on each other, and are also reigned by the states of their expressed genes.

In turn, the expression of a module is governed by its epigenetic states. If the gene is located in heterochromatin and methylated it is marked inactive and can not be activated by any composition of transcription factors. If both epigenetic sub-modules are inactive, the expression of the genes in the next time step only depends on the transcription factors. If the gene is in heterochromatin and not methylated or vice versa, there is still a chance that it is expressed, given the right transcription factors. We implemented these rules for all modules by the following Boolean formulas:

$$\begin{aligned} m_e^A(t+1) &= m_e^A(t) \wedge \neg(m_e^B \vee m_e^P(t)) \wedge \neg m_m^A(t) \\ m_e^A(t+1) &= m_e^A(t) \wedge \neg(m_e^B \vee m_e^P(t)) \wedge \neg m_{hc}^A(t). \end{aligned} \tag{4.4}$$

Activation of the pluripotency network by the transduced gene cocktail is also modeled explicitly. The pluripotency network has a small chance of being activated by the artificially introduced genes. Exogenic factors are deactivated when the cell has reached a pluripotent state with the pluripotency module turned on and all differentiated modules turned off. The probability of activation is rather small compared to the probability of the pluripotency module activating itself. Since only a small subset of pluripotency regulators is transduced in reprogramming experiments (Yamanaka factors) we assume that the activation is happening rarely. Section 4.2.1 clearly showed that the exogenous factors do not activate many of their targets effectively. The probability of activation is directly connected to the number of reprogramming factors transcribed, and it can be increased to model the influence of e.g. additional *NANOG* transduction, which has been shown to be beneficial to the efficiency of reprogramming (Hanna et al., 2009).

The deactivation of the transduced genes is achieved by silencing of their promoters through methylation and histone deacetylation (reviewed in Hotta and Ellis, 2008). In our model this process is triggered when the cell exhibits the iPS expression profile.

Since the transduced genes differ from the endogenous pluripotency genes in their promoter region, some changes regarding their transcriptional repression and interactions of DNA-methylation and chromatin remodeling are required compared to the other modules. Due to the constitutively active viral promoters the expression of the transduced genes only depends on their histone modifications and DNA-methylation state and not on any transcriptional inhibitors or activators.

$$\begin{aligned} m_e^E(t+1) &= m_{hc}^E(t) \vee m_m^E(t) \\ m_e^E(t+1) &= m_{hc}^E(t) \wedge m_m^E(t). \end{aligned} \tag{4.5}$$

The rules for DNA-methylation of the promoter of the exogenous genes are very similar to the ones of the other modules except for the probabilities which we chose to be smaller for *dnmt* and histone modification dependent DNA-methylation. This is due to the finding that after reprogramming, the retroviral genes can either be active (class I iPS cells) or silenced and thus producing fully reprogrammed class II iPS cells (Mikkelsen et al., 2008; Niwa, 2007). This suggests that methylation of the viral promoters is not fast and complete, as this necessarily would lead to quick silencing. Moreover, methylation does down-regulate the activity of the retroviral genes according to Pannell et al. (2000) which accounts for these low probabilities as well. Similar to the other modules, we also introduced slow, cell cycle dependent DNA demethylation induced by variable expression of DNMT1 after mitosis (Li, Bestor, and Jaenisch, 1992) (Table B.0.13).

Since the remaining update rules for DNA-methylation stay the same with the sole difference of lower probability in comparison to the other modules, this is the only structural difference:

$$m_m^E(t+1) = m_m^E(t) \wedge (\neg demeth(t) \vee dnmt(t)). \tag{4.6}$$

The histone modification rules of the retroviral genes mainly depend on their own methylation state (just like the other modules) and on the expression of the endogenous pluripotency genes. We hypothesize this interaction to be mediated by the *NANOG* and *OCT4* associated deacetylase (NODE) complex. It consists of a histone deacetylase (*HDAC*) and *NANOG* or *OCT4* (Liang et al., 2008) and was found to catalyze histone deacetylation on developmental target genes, thereby leading to heterochromatin formation. The hypothesis that this complex or at least one with very similar properties and behavior is responsible for retroviral gene silencing is based on the fact that *de novo* DNA-methylation is not necessary for retroviral silencing as mentioned above (Pannell et al., 2000). An additional hint into this direction is that a complex of NANOG and HDAC exists, which has shown to be active in silencing (Hotta and Ellis, 2008). Thus, the only update rule differing from the other modules regarding the chromatin structure depends on the expression of the pluripotency module P:

$$m_{hc}^E(t+1) = m_{hc}^E(t) \vee m_e^P(t) \tag{4.7}$$

which completes the set of update rules.

Stable Cell States and Differentiation of Combined Modules

When combining the single modules to a model of two differentiated states and the pluripotency network (*A*,*B*,*P* model) (Figure 4.2.10 B), already quite complex dynamics of state transitions exist. Gene expression in each module is mutually exclusive with all other modules and a module that is shut off once can only be activated by an external signal combined with epigenetic activation. The steady state of the pluripotency module consists of a number of different states that represent the hyperdynamic characteristics in epigenetic factors of the pluripotent cells (Meshorer et al., 2006). These states have different probabilities to differentiate, depending on the current epigenetic configuration. Similar kinds of population differences have been shown for pluripotent cells and the expression of *NANOG* (Kalmar et al., 2009).

We have already included three different levels of regulation in the model, we refrain from adding detailed signaling pathways to the system to regulate differentiation. We simulate differentiation by simply activating the gene expression of gene module *A* with a small probability. This causes the system to leave the pluripotent state very fast. After about 300 time steps it reaches its steady state with the differentiated state being the main attractor (Figure 4.2.11). The system also reaches a state, in which no proteins are expressed. This state is probably reached because the differentiation signal is strongly simplified and does not guarantee the correct timing of events. If the pluripotency genes are switched off before the correct methylation pattern is in place, differentiation related genes may not yet be properly expressed while pluripotency genes and thus *de novo* DNA-methylation and pluripotency related DNA demethylation mechanisms are already silenced. This behavior could be prevented by a proper regulation of gene expression by signaling molecules.

Nonetheless, de-regulation occurs in biological systems as well, caused by i.e. transcriptional noise, epigenetic variability, or external factors. The undefined cell state could be identified with cell death or other fatal events caused by the introduction of the stimulus. Such a simple differentiation could even be compared to the deregulation that happens in reprogramming and leads to senescence and apoptosis in many cells (Section 4.2.1). Despite the simplicity of the mechanism, the model differentiates very quickly and produces stable differentiated cell lines.

Modeling Roadblocks in Reprogramming

To be able to analyze reprogramming, we combined the four single modules, i.e. the retroviral transcription factors *E*, the endogenous pluripotency genes *P*, and the two model cell lineages *A* and *B* into one model (Figure 4.2.9). We simulated the model in a Markov simulation (Section 2.3.1) for various starting distributions and systematically analyzed the temporal dynamics of the model for typical start scenarios.

First, we analyzed the situation in which the system is initialized with only one defined state that corresponds to either one of the two cell lineages *A* and *B*. This is the state, where the set of master regulator genes associated with lineage *A* is expressed, unmethylated, and in an open chromatin configuration. At the same time the module for lineage *B* and for exogenous (*E*) and endogenous (*P*) pluripotency genes has the opposite configuration, i.e. the genes are downregulated, methylated, and in a closed chromatin formation. For this specific start state, the network remains in its differentiated cell lineage over the complete time of simulation, i.e. the defined cell lineage is stable without any outer perturbation (data not shown).

Second, when the simulation starts from a state that corresponds to the fully reprogrammed cells, i.e. where module *P* has the active configuration, while the other modules are silenced, we observed a shift into a set of states closely related to pluripotency. This behavior can be observed in cultures of iPSCs and ESCs as well and is often referred to as a hyperdynamic plasticity. The cells have a fast changing chromatin structure in general and different methylation states on several loci (Meshorer et al., 2006). This plasticity leads to

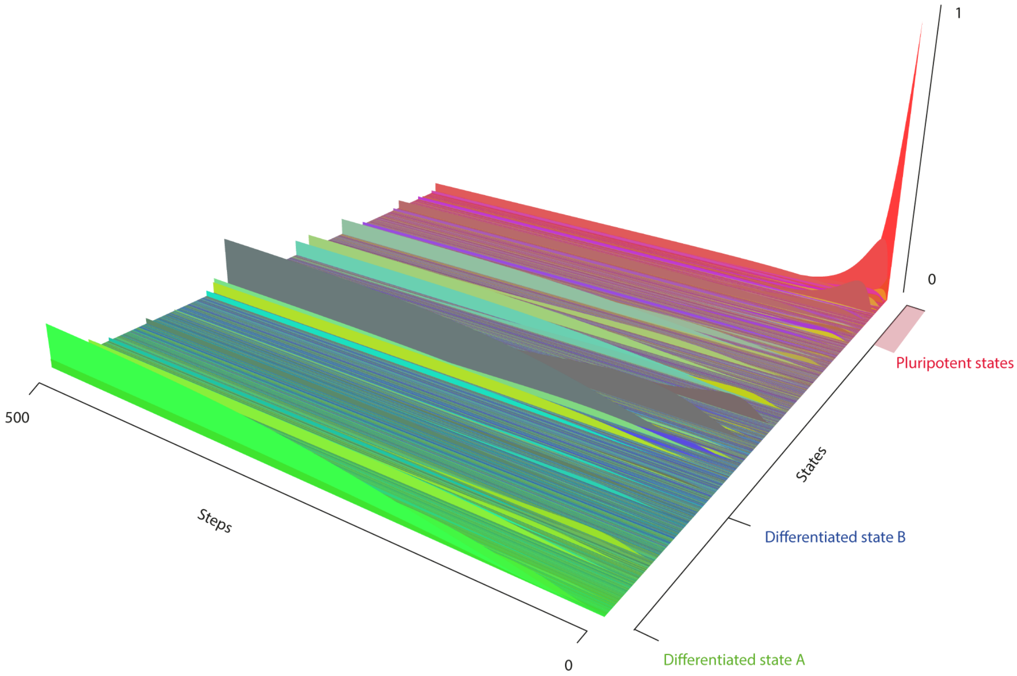


Figure 4.2.11: The epigenetic landscape of differentiation by the activation of module A through a weak signal. All possible states of the model are sorted along the x-axis by similarity. The y-axis corresponds to simulation time steps, and the z-axis to state probabilities.

a distribution across different states in our model as there is no single point attractor. This effect may also be responsible for the priming of iPSCs to quickly differentiate into various different cell types upon external signals (Ang et al., 2011), as we also observe states that can more easily differentiate than the defined pluripotent state.

Third, the focus of the simulation was the model starting in the sharp states of the differentiated cell lineages when the retroviral transcription factors are expressed, unmethylated, and in open chromatin. These simulations can be related to classical direct reprogramming. As shown in figure 4.2.12, the starting state will be left quickly for transient states that lie along the path to pluripotent cells. On inspection of the landscape it becomes clear that cells will transit into states that resemble pluripotent cells more and more until they eventually reach the fully reprogrammed state with a certain probability. This probability can be considered as the reprogramming efficiency that increases with time (or cell cycles) as demonstrated before by Hanna et al. (2009).

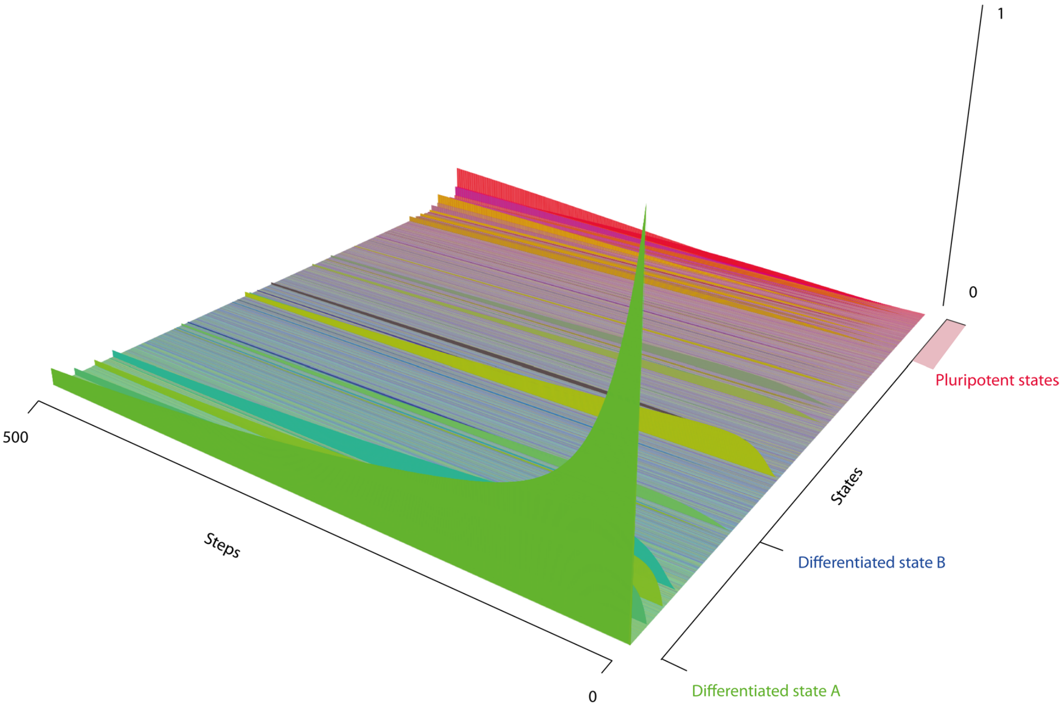


Figure 4.2.12: Landscape of the reprogramming experiment. Reprogramming starting from one clearly defined state where module A is active and the reprogramming factors are present. Compared to Figure 4.2.11, the transition happens a lot slower and the probabilities of the reprogrammed state are lower.

Routes through the state space

The state space of the simulation (Figure 4.2.13) reveals some more details about the timing and order of the states that are passed on the way to reprogrammed cells. Since we are dealing with a model of 14 variables, the whole state space has 2^{14} states, in a probabilistic approach the state space could in theory be fully connected, i.e. every node of the state space could possibly have 2^{14} outgoing edges. Therefore, we only show those states that can be reached from our starting state, and reach a probability larger than 0.0001 in the course of the simulation. These most relevant states are the ones that some cells will probably pass during the process of becoming iPS cells. Surprisingly, these states clearly show different events that are crucial in the reprogramming process and resemble the order of events described in the literature (Papp and Plath, 2011).

We simulated the model over 500 time steps, until the state probabilities assumed a steady distribution. The states that have the highest probability to be observed in the beginning (time steps $\sim 1 - 100$) show a slow unpacking of the pluripotency genes, but about 10% also show modifications to the genes in the other lineage. After this stage all cells can enter the next phase which lasts for about 150 time steps. Nearly all states in this phase share the property that the differentiation genes are already shut down, but endogenous pluripotency genes are still silenced. From this stage there is a non-negligible possibility that the cells enter a non-functional state where nothing is expressed anymore, which cannot be left. This state creates a small attractor that prevents the cells from successful reprogramming. The phase that follows with a much higher probability is the first stage of reprogrammed states. Cells in this state have been characterized as class I iPS cells (Mikkelsen et al., 2008; Niwa, 2007), as they express the endogenous pluripotency genes, as well as the exogenous reprogramming cocktail that is not epigenetically silenced yet. From these states there is a slow transition to the states with a stable silencing of the exogenous factors, expression of pluripotency genes, and a hyperdynamic state in the differentiation modules. Cells exhibiting one of these states can have taken any existing path through the state space. In the visualized subspace, which makes up about half of the states reachable from the start state, there are 146 states and 2473 edges with only one connected component. As there are so many possibilities, the probability for each path is infinitesimally small. The most probable single path from the start state to the iPS state only has a probability of $9.3 \cdot 10^{-12}$ and consists of 7 state transitions. When looking at the state space structure it becomes obvious that the phases described above cluster together in the graph and that some states are much more central to the transition than others. Most states are not essential to the reprogramming, because there are paths that can avoid them. But there is one transition that is absolutely essential for reprogramming. This is the transient activation of the pluripotency module relatively early in the process after removal of their methylation marks. This enables the suppression of the differentiation genes and makes further reprogramming possible.

At the end of the process there is a large probability that cells are in the reprogrammed state. This corresponds to the findings by Hanna et al. (2009), who showed that in a drug inducible reprogramming system all cells are able to reprogram given enough time.

As illustrated by the most probable path, not all of the described phases will be passed by all cells, shortcuts like the one shown are possible, e.g. cells with a demethylation of all modules in an early phase, the pluripotency genes can be activated much faster compared to the rest.

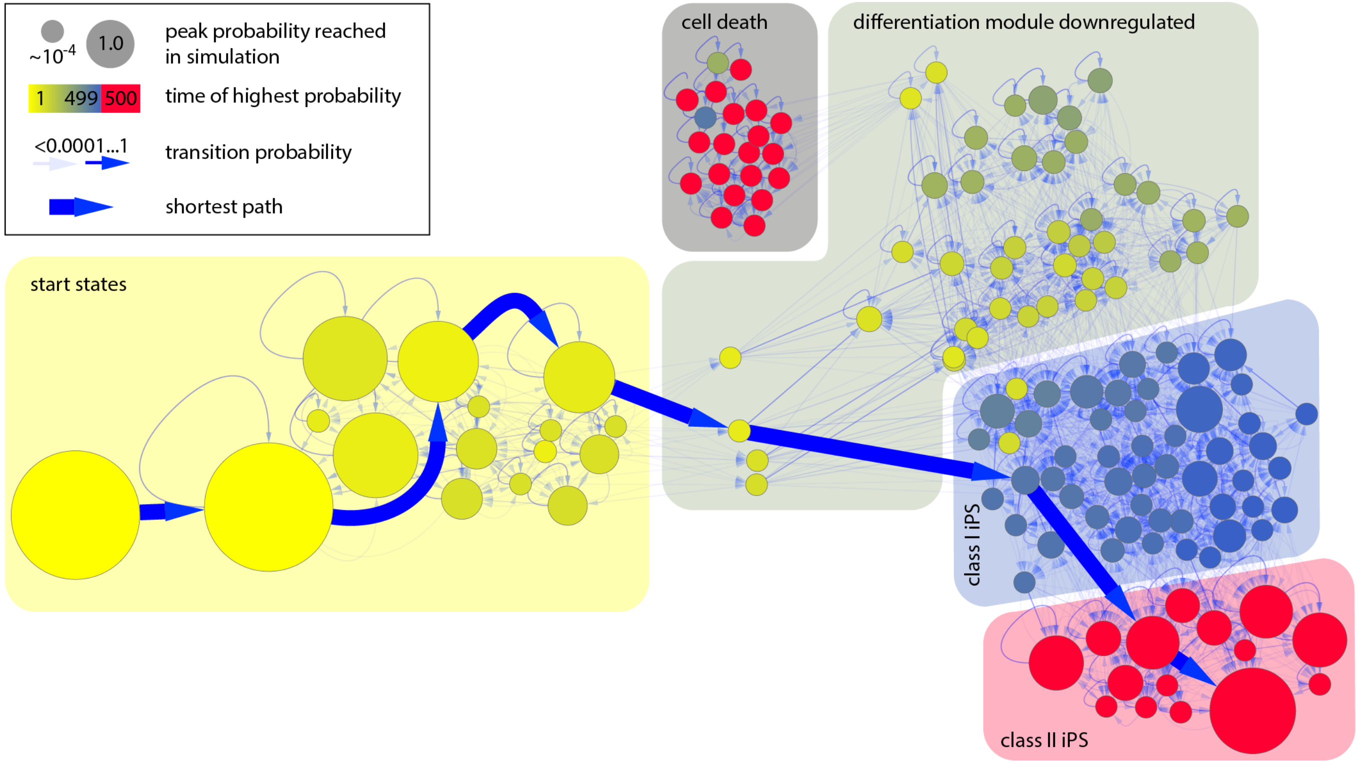


Figure 4.2.13: State space of reprogramming. Time evolution of the model starting with an active differentiation network and active reprogramming genes. The figure only shows the states that are reached with a probability of $p \geq 10^{-4}$. The model has 2073 possible state transitions between these 149 states. Different phases can clearly be separated in the reprogramming process. In the beginning (yellow area) the epigenetic factors of the different modules are modified, but there is no change in gene expression yet. The second phase (dark yellow) represents the down-regulation of the differentiation module followed by the activation of the pluripotency module (blue area). The last step consists of the silencing of exogenous factors that produces stable iPS cells (red area). There are some states that can lead to nonviable cells, in which no regulators are expressed at all (grey area). The bold blue arrows represent the shortest path to the main pluripotent state.

Parameter Variations of the Model

In order to analyze the stability of our model and its behavior upon parameter variation, we varied the strength of the epigenetic modifications, i.e. DNA-methylation and histone modification changes. We chose these parameters, because they are present in most update functions (Table B.0.13) and thereby probably also have the biggest influence. We defined a parameter range including the parameters of our main model, a decreased and an increased probability of changes in methylation and deactivating histone modification formation and analyzed the effect on the reprogramming efficiency (Figure 4.2.14). Interestingly, we can observe that in the time range of 2000 time steps our main model nearly seems to have a maximal saturation for its reprogramming efficiency. This is only slightly surpassed by increasing the probability for removal of negative histone marks. The increase could experimentally be reached through a heterochromatin formation inhibiting agent such as VPA (Section 4.1.1 and Table B.0.14).

However, the timing of reprogramming can obviously be influenced by parameter variations. While an increase in methylation dynamics, i.e. faster demethylation, speeds up the reprogramming process with a reprogramming efficiency peaking at approximately 0.8 after 2000 time steps, we observe slower reprogramming for increased probabilities of the formation of negative histone modifications and DNA-methylation. To check the sensitivity of the model to structural modifications and how its behavior corresponds to responses of reprogramming cells in reality, we searched the literature for various experiments that can be mimicked by slight modifications (Table B.0.14). In the following sections we describe such modifications and their effects on the reprogramming process with a focus on efficiency.

Spontaneous methylation

Since the exact mechanism of action of DNMTs in DNA-methylation is still not fully understood, we modified the model to include spontaneous methylation. We introduced an interaction that accounts for methylation of the different modules by *dnmt* independently of other factors.

We found that in comparison to the original model, there was an overall decrease in the reprogramming efficiency, i.e. probability to be in a reprogrammed state is approximately 10 times lower than in the original model after 500 time steps. However, the spontaneous methylation model reaches its maximum distribution slightly faster. Another interesting feature of the spontaneous methylation model is the fact, that it assumes a new specific state with a high probability. In this state, which we will call the undefined state, all modules are silenced except for the retroviral ones. We will discuss this state below.

Spontaneous Histone Deacetylation

Similar to DNA-methylation, the exact mechanisms of histone modifications are still a matter of debate. In our model, introducing spontaneous formation of negative histone marks as an independent term is a general de-regulation of these mechanisms. In principle it could happen during reprogramming due to Yamanaka factor induction.

The steady state reprogramming efficiency in the spontaneous methylation model is more than 40 times lower than in the main model. Interestingly, in the first 50 time steps the probability to be in a reprogrammed state is higher than in the original or the methylation model and it is only at later time points that this ratio is reversed. This may be due to the fact that the differentiation related state is downregulated much faster (results not shown). Similar to the methylation model, the undefined state mentioned above is also attained with a high probability.

Spontaneous Demethylation

In contrast to DNA-methylation and histone modifications, which have been already in the focus of research for many years, active DNA demethylation has long remained in the dark until recent discoveries have unraveled a new perspective. DNA demethylation seems possible via the intermediate 5-hydroxymethylcytosine and different enzyme driven modifications which transform it back to unmethylated cytosine (Bhutani, Burns, and Blau, 2011). To account for uncertainty in spontaneous demethylation, we transformed the modules of the model in order to be able to randomly lose their methylation with a certain probability.

Of all the structural model variants, spontaneous demethylation leads to the highest reprogramming efficiency after 500 time steps although it is still 3-4 times lower than in the original model. Although its reprogramming efficiency stands back behind the original model in every time-point, its differentiated state shows a fast decrease in probability at the beginning, followed by a much longer second phase of slow decay. This process is very similar, to the reprogramming experiment of the original model (Figure 4.2.14).

Stronger Interaction Between Methylation and Heterochromatin

We analyzed the effects of the debated interaction between methylation and negative histone mark formation which we explained above and in Table B.0.13. The overall reprogramming efficiency after 500 time steps was approximately reduced by the factor 5. The dynamics of the differentiated state are similar to the ones of the original model although it decreases even slower and remains with a higher probability at the end. Interestingly, another state is found with high probability. It is similar to the differentiated state of the other cell lineage, except that the pluripotency module is already demethylated and in an open chromatin formation but not expressed. Yet, this state is transiently present with a high probability which slightly decreases over time. This phenomenon could be interpreted as

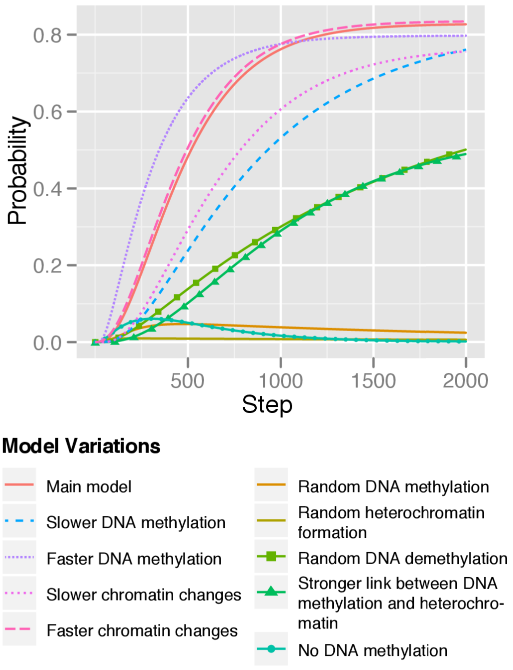


Figure 4.2.14: Reprogramming efficiency seen as the sum of all states that end up in the reprogrammed state can be quite high, and follows a saturation kinetic. The original model has the highest efficiency, compared to all variants. Differences between the original model and structural changes are quite drastic, whereas parameter variations don't influence the outcome very much.

trans-differentiation during reprogramming without passing the pluripotency state (Vierbuchen et al., 2010).

No Methylation

As a final model variant we assume that methylation has no influence on gene expression or heterochromatin formation. Without methylation effects, the model is neither able to reprogram anymore nor to differentiate. What we can observe instead is a re-distribution of the different start states, i.e. the pluripotency related or the differentiation related states into similar states but no transition to any states that are further away in the state space. This is most likely due to the fact that methylation is needed in the long run to determine the heterochromatin structure after cell division and to fully silence gene expression. Without these features active modules cannot be silenced and thus inactive modules stay transcriptionally inhibited although they might be demethylated and have positive histone marks structure.

Summary

The effects of the analyzed model variants on the reprogramming efficiency are summarized in Figure 4.2.14. For every model variant the reprogramming efficiency increases over time except for the model without methylation. What becomes apparent at first sight is that obviously all structural model variants seem to have a strongly decreased reprogramming efficiency after 2000 timesteps.

Overall, we find that all variants resulted in de-regulating modifications of the original model, i.e. modifications that reduced the level of tight regulation of the epigenetic processes involved, which in turn have a direct effect on the expression of important genes. In the landscape of these model variants (not shown here), one could observe a general transition from a few defined states that could be reached in the original model, to a dramatically increased number of states. In the original model, we can reach a total of 2592 states after 500 time steps in a reprogramming simulation, while the spontaneous methylation model variant could reach 10240 states in the same time. However, the efficiency of reprogramming was approximately 10 times lower (Figure 4.2.14) in the spontaneous methylation model. Nevertheless, the 366 pluripotency related states in both models are the same. Only their probability to be reached after 500 time steps is much lower in the variant.

4.3 Discussion

The goal of the work outlined in this chapter was to improve our understanding of the intricate regulatory mechanisms that govern the process of reprogramming somatic cells to a pluripotent state. A first step was the analysis of high-throughput and literature data to generate a working hypothesis for a model of reprogramming. The model we implemented afterwards includes key findings of the recent literature, but additionally uses the outcome of the analysis as a guideline.

We were able to extract crucial processes from the microarray time series dataset comparing the first three days of a reprogramming experiment to the source cells (HFF1) and the desired cell states (iPS/ES) (Mah et al., 2011). By structuring the dataset, employing clustering methods in combination with database enrichment, we identified key expression changes that play a role in the commencement of reprogramming.

Different database analyses highlighted different aspects, but generally showed a consistent picture of the changing biological functions and pathways. Both, GO and KEGG inspection, showed an activation of immune related genes in the first day of reprogramming compared to either HFF1 and iPS cells, but also a fast adaptation of the immune pathways. Already after two days we saw a lower expression of response to virus genes which continued to decline. This identifies innate immunity as the first barrier that cells have to overcome to become reprogrammed. It hints towards immune suppression as a possible candidate for optimization of reprogramming protocols. This observation would also partly explain the differences in efficiency between stable inducible reprogramming systems and viral transduction (Hockemeyer et al., 2008).

The second set of genes we identified in both analyses were those related to cell-matrix adhesion and actin fibers. These genes start to decline in expression after two days of reprogramming and are continuously less expressed in iPS and ES cells. The focal adhesion pathway includes a number of these genes and is constantly negatively impacted (Figure 4.2.4). The downregulation of these genes is an example for the early onset of cellular changes towards a stem cell like state in SCR. The loss of focal adhesions can be seen as the onset of the MET, which is a necessary step towards a successful reprogramming (Papp and Plath, 2011). Epithelial genes were only upregulated in the iPS and ES cells and did not change in the first phases, which shows that the MET is not completed on the third day.

The fact that there are nearly no pluripotency related genes activated in the beginning led to a further inspection of direct influence of the Yamanaka factors on their targets. We observed a difference between the expression changes of the exogenous factors and the expected changes in their targets in the experiment. The Yamanaka factors were all highly expressed after 24h, but there were nearly no pluripotency markers present in the first three days of reprogramming, which was reflected in low TFA values for the early time points. This difference is probably caused by epigenetic modifications (Hawkins et al., 2010; Lister et al., 2009) that hinder the direct transfer of the information from the Yamanaka factors to their targets. Epigenetic modifications were not subject of the experiment, but we could observe some downregulation of HMTs in early reprogramming as well as in iPS and ES.

The extra layer of regulation between our stimulus (Yamanaka factors) and our readout made it impossible to build a quantitative model that could reproduce the data without including epigenetic modifications. This forced us to take a more abstract approach, because a simulatable stochastic model of the whole network with included methylation and histone modifications would not be feasible. However, since the whole dynamic process cannot be understood by looking at its single pieces (Section 1.2), we want to take a more holistic approach in this work and combine gene expression and epigenetic principles in one abstract model.

We developed (to our knowledge) the first model of processes in somatic cell reprogramming that explicitly includes the virally transduced factors and their regulatory interactions. The model is also unique in its representation of the different epigenetic factors that regulate cell states and their interactions. Our modeling approach qualitatively reproduces experimental results from reprogramming as well as differentiation experiments. The probabilistic Boolean state space in combination with the epigenetic landscape plots of the simulations gives insights into different possible ways reprogrammed cells take in this scenario. In combination these visualizations can be related to the potential landscapes that have been developed for continuous modeling approaches (Wang et al., 2010). They show the direction the system is moving towards as well as the probability for reaching each state under specific conditions. The stategraph also enables us to identify different phases during reprogramming that are important milestones. These simulation phases are coherent with the sequence of events reported in experiments (reviewed in Hanna, Saha, and Jaenisch, 2010; Papp and Plath, 2011). This sequence is also supported by our experimental data, e.g. by the early inhibition of the native HFF1 pathways, and the later activation of pathways typical for ES cells.

The reprogramming efficiency of the system seems high ($p = 0.8$ after 2000 time steps) compared to experimental results from transduction experiments, but one has to keep in mind that the model leaves out major experimental hurdles and regulatory mechanisms. We neither include the immune response of cells — which we previously identified as reprogramming roadblocks — nor varying transduction rates. The general efficiency shows a similar behavior to experiments done in inducible stem cell systems, which also showed sigmoidal efficiency curves with saturation at high levels (Hanna et al., 2009). After a long simulation time we see a high steady state of reprogrammed cells in a relatively broad distribution of states. Nevertheless, this high reprogramming rate indicates that there are mechanisms not included in our model that suppress reprogramming in differentiated cells.

Differentiation in our model is also possible and happens a lot faster than reprogramming, although it is impaired by the lack of regulatory factors. In order to improve the representation of the course of differentiating cells the model would need to be extended by signaling pathways. This would enable a more precise modulation of the activity of the important model components and would moreover enable the system to sense external factors. Another interesting extension of the model consists in the integration of further branches of differentiation for other cell lineages to depict the path from stem cells over progenitors to fully differentiated cells. These extensions are simplified by the modular structure of our model, but would nevertheless increase the statespace drastically.

The model is centered on the mutual inhibition of master transcription factors and their connection to epigenetic factors, which is an important mechanism. However, other regulatory processes, not captured by this approach, certainly play crucial roles during phenotype transitions (e.g. miRNAs). Cell types are generally viewed as different steady states of gene regulatory networks. This is reflected in the model by various attractors that represent different cell types. Nevertheless, it does not account for cell types that express a mixture of regulatory genes, as for example progenitor cells would. These states can occur as cyclic attractors, but are unstable to stochastic perturbations and are therefore left quickly. One could, in an extended version, include conditions that stabilize these states to generate a progenitor cell.

Because the model is so abstract, predictions cannot aim to reflect more than a small part of biological reality, but they can show trends and general effects that hold true for the modeled entities. The modifications we introduced show how the system reacts to perturbations in the epigenetic regulations. Most of the structural changes showed a devastating effect on the reprogramming efficiency, demonstrating the need for tight regulation of the process. The only two modifications, in which efficiency can be sustained at an adequate level, are those that increase the influence of the gene regulatory networks on the epigenetic factors.

Improvements of the reprogramming efficiency can only be achieved by two modifications. First, higher probabilities for changes in DNA methylation status lead to a faster increase of the reprogrammed cells, but also to a lower probability in the steady state. Therefore a de-regulation can have beneficial effects on the process, but also has drawbacks. Higher probabilities for changes in histone mark formation lead to a mild increase in efficiency, resembling the effects of small molecules like VPA.

A better understanding of the underlying processes of somatic cell reprogramming is the key to a clinical application of iPS cells in the future. The proposed model, although abstract and limited, extends our knowledge into this direction. It outlines the possible epigenetic regulations that play a role in reprogramming, elucidates their connections, and partly explains experimental observations in reprogramming although it ignores large parts of the complex gene regulatory network of developmental genes. Additionally, the analyzed experiments hint at the possibility of improving reprogramming efficiency by inhibiting immune response in the early phase of reprogramming.

5

Proteomes of Influenza Virus and its Host in the Course of the Infection Cycle

The influenza A virus still poses a serious health issue worldwide, and has caused millions of deaths in different pandemics in the last century. Due to its virulence it attracts a lot of research and a lot of its lifecycle has been uncovered already (reviewed in Bouvier and Palese, 2008). Nevertheless, a lot of the cellular processes the virus influences during its lifecycle are still unknown. In this chapter, we present the results of a study that tries to unravel the connections between host and virus, utilizing recent advances in proteomic research. We were able to measure virus and host proteomes in parallel at different time points of infection. With this data we were able to refine the knowledge about interactions between host and virus and propose new levels of interconnection.

5.1 Introduction

5.1.1 Influenza A Infection and Host Virus Interaction

Influenza A infection is a complex process, in which the virus uses the host cell's metabolic machinery to produce new virions. On an organism level, the virions enter the body via the respiratory system, and primarily infect cells in the upper and lower respiratory tract. Influenza viruses belong to the family of *Orthomyxoviridae*, a term containing the greek word for mucus (*myxa*), which describes the symptoms caused by influenza infection. *Orthomyxoviridae* are single stranded, negative sense RNA viruses with a fragmented genome. The genome of influenza A consists of 8 independent viral RNA (vRNA) fragments that are translated to messenger RNA (mRNA) and complementary RNA (cRNA) in the course of virus replication.

The virus particles are coated with a lipid envelope, attained during budding from the former host cell membrane. They can differ in shape, although most are spherical with about 100 nm in diameter. Embedded in the lipid hull are two of the viral proteins — Hemagglutinin (HA) and Neuraminidase (NA) — that are presented on the outside of the virus. These proteins are crucial for both ends of the infection cycle. While HA, one of the best studied viral proteins, initiates the lifecycle by mediating virus binding to the cell surface and its escape from the endosome, NA enzymatically cuts the newly produced virions loose from sialic bonds on the cell membrane and enables them to infect other cells. HA and NA are so obviously important for the virus that their subtypes are used to describe the virus subtype (e.g. H1N1, H5N1). These virus classes differ strongly in virulence and there are only 6 subtypes that have been isolated from humans (Cheung and Poon, 2007), although there are 16 types of HA and nine types of NA (Fouchier et al., 2005; Laver et al., 1984). Other types of viruses infect birds or pigs.

Another reason why these proteins are so extensively investigated is that they are presented on the viral envelope, and are therefore a target for antibodies. The problem for immune system recognition is the so called antigenic drift that these proteins exhibit. They accumulate point mutations quickly and generate new subtypes every year. On top of that, a so called antigenic shift can happen, when two different influenza A subtypes infect the same host (Bouvier and Palese, 2008). This co-infection can lead to a genome re-assortment, due to the fragmented nature of the influenza genomes. This often happens across species boundaries in hosts and leads to words like "swine-flu". The resulting subtypes can have a dramatically increased virulence and can cause pandemics of catastrophic proportions ("spanish flu" 1918).

The influenza genome codes for 9 other proteins, which we will describe to some extend in the following section.

Genome and Structure of the Influenza Virus

As stated above, the influenza A virus has a fragmented genome, that consists of 8 negative-vRNA strands that are present in each virion and encode for 11 proteins in total. The vRNA is encapsulated in Nucleoprotein (NP). This structure is referred to as the ribonucleoprotein vRNP and it is complexed with the viral polymerase subunits (PA, PB1, and PB2). The polymerases seem to interact with both ends of the vRNP and it probably forms a cyclic or supercoil structure inside the virion (Hsu, 1987). NP is the second most abundant protein in the virions ($\approx$ 1000 proteins per viral particle), connecting to the vRNA in a one protein per 11 bases ratio (Baudin et al., 1994). It is only topped by the Matrix Protein 1 (M1) that forms the connection between the membrane proteins and the vRNP ($\approx$ 3000 proteins per viral particle). M1 has plenty of described functions throughout the lifecycle of the virus which we will describe in detail further below. It has to be distinguished from the third membrane protein, matrix protein 2 (M2) that is transcribed from the same genome segment as M1 by alternative splicing. M2 is an ion channel and is responsible for the acidification in the viral particle during endocytosis inside the endosome, followed by uncoating of the vRNA. HA comprises about 80% of the proteins in the viral membrane, and is therefore much more abundant than NA or M2 (Table 5.1.1).

Two proteins were characterized as nonstructural proteins (NS1 and NS2). Nonstructural in the sense that only the coding RNA fragments are present in the virions, but the protein is only needed for reproduction in the host cell and only translated "on the fly". In case of NS1 this characterization still holds, but NS2 has been found in low amounts in the virion (Richardson and Akkina, 1991). NS2 is active in the export of vRNA from the nucleus, and it has therefore been proposed to rename it to nuclear export protein (NEP) (ONeill, Talon, and Palese, 1998). NS1 also plays an important role in the cap snatching mechanism and in the modulation of host mRNA.

Recently there have been findings of new post transcriptional modifications producing protein variants which have not been observed before. PB1-F2 serves an anti inflammatory function and increases the virulence of virus strains where it was detected (Varga et al., 2011). PA-X, another frame shift variant of the viral protein PA, falls into the same category and has been described to modulate the host immune response (Jagger et al., 2012).

Infection Cycle

Viruses reproduce by using the cellular machinery of their hosts to their advantage. The influenza virus replication is a cycle that in the end produces new viral particles that bud from the host cell, which can then in turn infect new host cells and organisms (Figure 5.1.1). To start the cycle the virus has to bind to the cell membrane and initiate the process of endocytosis. Animal cells are covered with glycopeptides or glycolipids, which often contain sialic acids at their ends that can differ in their exact chemical structure. The HA protein on the virus surface has a binding pocket for special sialic acids that differ from subtype to subtype ($a - 2, 6$ or $a - 2, 3$, depending on the preferred host species).

After the virus is bound to the cell surface, it is endocytosed and enclosed in an endosome via cellular mechanisms. The low pH values in the endosome trigger two important processes for the virus uncoating: (i) The M2 ion channel in the virus envelope lets H^+ ions enter into the virus matrix, which weakens protein-protein interactions between the M1 protein and the vRNP by lowering the pH; (ii) the HA protein changes its conformation and thereby exposes the so called fusion peptide. This peptide facilitates the fusion of the viral envelope with the endosomal membrane, releasing the vRNP into the cytosol.

The vRNP exhibits nuclear localization signals, which direct cellular proteins to import them into the nucleus (Cros and Palese, 2003). Inside the nucleus, two major processes in virus replication are driven by the viral polymerases complexed with the vRNP: (i) transcription of the vRNA to polyadenylated and capped mRNA; (ii) transcription of the negative sense vRNA to positive cRNA templates that are used to generate copies of the vRNA.

Table 5.1.1: The viral proteins present in the H1N1 variant that was used in our experiments (*Influenza A Puerto Rico/8/H1N1*). For proteins where it is known, the approximate proteins per particle number is given. To provide an overview, the RNA segment they are transcribed from and the most prominent functions are mentioned.

Protein	RNA	AA	≈p/p	Function
M1	7	252	3000	Virus budding; inhibition of transcription; vRNP nuclear export
NP	5	498	1000	formation of vRNP with RNA;
HA	4	565	500	host cell surface binding; endosome escape;
NA	6	454	100	detachment of virus from host cells;
PA	3	716	??	polymerase subunit; helicase;
PB1	2	757	??	functional polymerase subunit;
PB1-F2	2	87	??	anti-interferon function; pro apoptotic;
PB2	1	759	??	cap-snatching;
M2	7	97	<100	ion channel; pH regulation
NS1	8	230	-	transcription regulation; host interaction;
NS2/NEP	8	121	<100	vRNA nuclear export

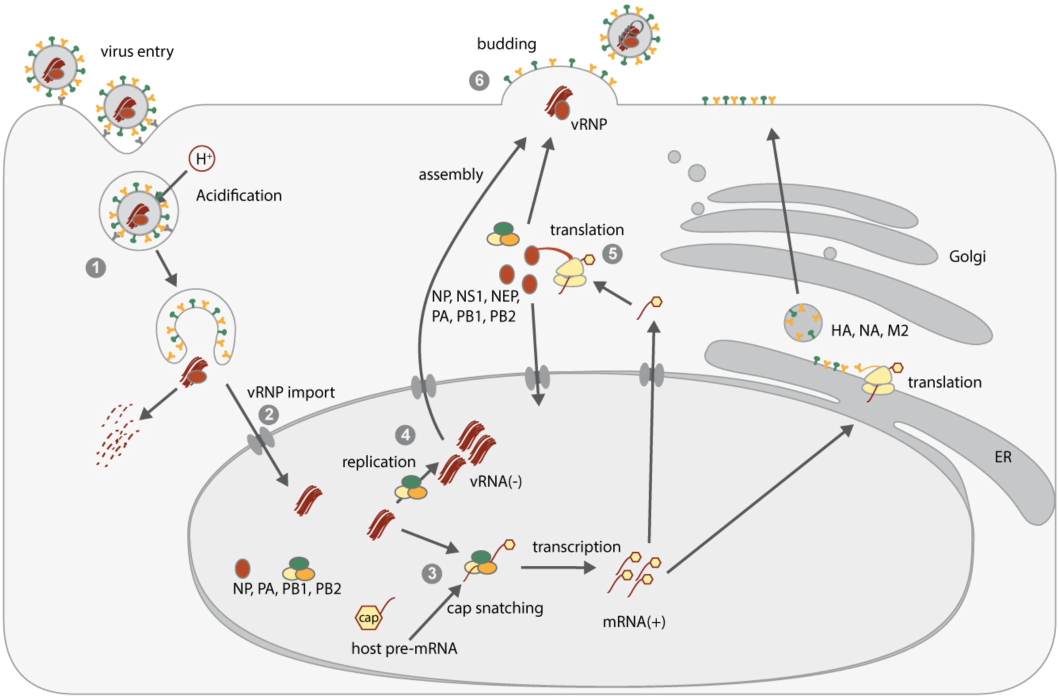


Figure 5.1.1: The influenza virus and its infection of the host cell. (1) The virus HA membrane protein binds cellular sialic acids and is endocytosed. (2) Induced by low pH in the endosome, conformation changes lead to fusion of the viral membrane with the endosomal membrane and release of the vRNP into the cytosol. The vRNP is imported into the nucleus by the nuclear pore, because of its localization signal peptides. (3) Viral mRNA is transcribed and provided with 5'-caps of host mRNA by the viral polymerase. (4) vRNA is replicated and exported from the nucleus. (5) Viral proteins are translated and transported to the membrane on different ways, depending on whether they are integral membrane proteins or not. (6) New viral particles bud from the cell and are cut loose by the NA proteins on the membrane.

Regulation of the balance between these two processes is proposed to be dependent on the level of NP protein that seems to act as a switch from transcription to replication (Shapira et al., 2009; Shapiro and Krug, 1988). Viral mRNA capping is a unique process, as it includes a mechanism called cap snatching. The endonuclease domain of PB2 cuts the 5'-cap from cellular mRNA to use it as a primer for viral mRNA transcription. This mechanism protects the mRNA from endonucleolytic degradation (Plotch et al., 1981). Additionally it ensures that the cellular RNAs robbed of their caps are degraded and thereby are prevented from nuclear export to the cytosol for translation.

The viral proteins that need to be embedded into the viral envelope are translated by membrane bound ribosomes and trafficked through the Golgi apparatus to the cell membrane for budding, making use of the actin cytoskeleton of the cell. The mRNA coding for internal and non structural proteins are exported via the normal cellular pathways to the cytosol and translated by free ribosomes. vRNA however — which has to leave the nucleus to reach the membrane for budding of new viral particles — needs the NEP and M1 proteins to mediate its export via nucleoporins (Bouvier and Palese, 2008).

When there is a sufficient amount of vRNA and viral proteins present, the viruses are packaged and start budding from the cell membrane. This part of the replication cycle is the least well studied. The budding probably starts by accumulation of M1 at the cell membrane and recently discovered signaling sequences ensure the packaging of the whole genome in each particle (Fujii et al., 2003). At this state the NA protein is of utter importance for the virulence and effectiveness of the virus. The NA protein cuts the sialic bonds that are formed by the HA protein and the surface glycoconjugates of the host.

5.2 Results

The viral infection cycle has been studied in great detail, but the dynamic host reaction has not been the target of a greater interest, especially looking at the proteome.

In our global study of proteome dynamics of the host and the influenza virus, we measured 1262 cellular proteins and 9 viral proteins over 4 time points spread over 12 hours (4, 8, 10, and 12 hours). To be able to measure this many proteins at once, we applied a high-throughput quantitative MS technique. **S**table **I**sotope **L**abeling of **A**mino acids in **C**ulture (SILAC, Section 2.1.1), is a sensitive way to measure the relative changes of protein concentrations in a probe. Using **I**ntensity **B**ased **A**bsolute **Q**uantification (IBAQ) we could also measure absolute abundance values of proteins, albeit with a lower precision.

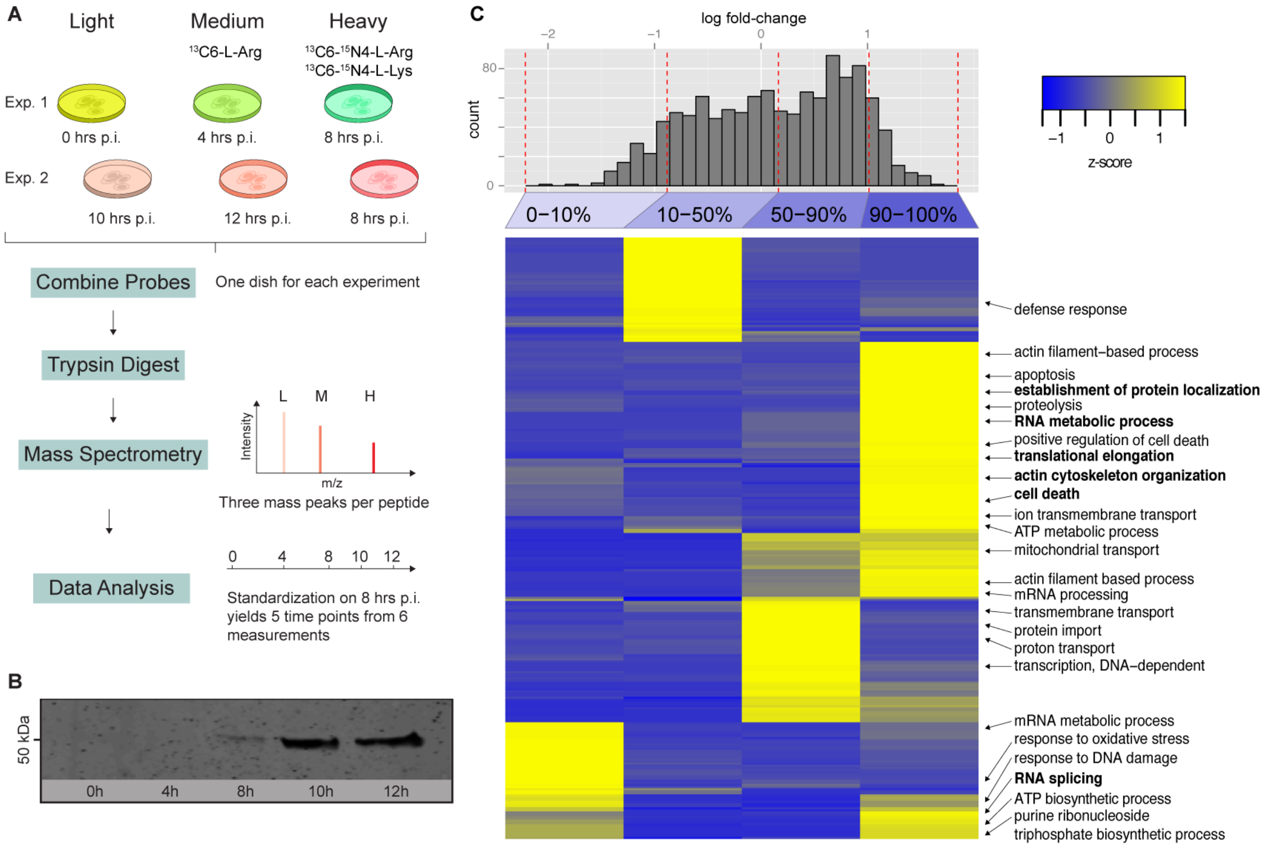


Figure 5.2.1: **A** Experimental procedure following the SILAC protocol. Two SILAC experiments were combined by normalization on a common time point creating 4 time points with relative data. **B** Western blots of viral NP protein from the same cells used for the SILAC experiment. **C** Proteomic phenotyping of the influenza A/PR/8 infected MDCK cell proteome using GO annotations. Quantiles of the quantification histogram are indicated at the top of the heatmap. Each quantile was separately analyzed for gene ontology pathways and clustered for the z-transformed p-values. The most prominent representatives of all over-represented biological processes of each quantile were selected and annotated.

The experiments were carried out using MDCK cells (*Canis familiaris*), which we infected with H1N1 (*Influenza A Puerto Rico/8/H1N1*) virus particles at a multiplicity of infection (MOI) $\approx$100. We prepared probes in different SILAC media and then measured at 4, 8, 10, and 12 hours post infection (Figure 5.2.1 A).

An in depth bioinformatic analysis we performed showed the strong interconnection of the viral and the host proteome and suggests interesting feedback loops between the two that can explain the data (Section 5.2.2).

5.2.1 Virus Proteins Show Strong Differences in Dynamics of Translation

As expected, virus proteins show much bigger changes in abundance than cellular proteins, as they are not present in the cells before infection and then gradually increase with virus entry and replication. The dynamics of the observed virus proteome are presented in Figure 5.2.2. The method was able to identify nearly all (9 of 12) known viral proteins that can be found in the strain (Table 5.1.1). The only standard protein that was not detected is the envelope ion channel M2.

Quantifying the changes in the viral proteins was not as straight forward as the SILAC approach for the cellular proteins. The reason is, that for the SILAC approach we use a number of probes that are grown in different media to include the labeled amino acids into their proteome (Section 2.1.1). This makes the total protein of the different time points visible as different peaks in the spectrum (Figure 5.2.1 A). For the virus particles used for the infection however, this is different. The particles were all cultured in normal (i.e light) medium and so we have to interpret the SILAC results for the viral proteins differently. The ratios we see between the different timesteps are not the ratios between the total protein present, but the ratio between the newly produced protein and the protein still present from the infection batch. We therefore need to be very careful with interpreting the SILAC results for the viral data, but we can include IBAQ data to show the large scale changes in the viral proteins.

The 9 quantified proteins can be separated into two groups relating to their dynamic expression. The first group — including HA, M1, NS1, and NP — shows comparatively high protein counts already at the 4 hour time point and gradually increases up to 10 hours where it saturates (Figure 5.2.2 A). The second group on the other hand shows a different behavior. NA, NS2, PA, PB1, and PB2 stay more or less constant until the 8 hour time point and then increase dramatically at 10 hours(Figure 5.2.2 B). But not only the time profile is a major difference between the groups, also the absolute abundance differs significantly. The first group is always expressed more strongly than the second by approximately an order of magnitude in every time point. The differences in protein abundance nicely reflect the previously reported protein abundances in the virion for most proteins (Table 5.1.1). For example for M1:HA:NA we found a molar ratio of 12:3:1. This is in good agreement with previous studies reporting about 3000 molecules of M1 protein per particle forming the inner core while 500 HA and 100 NA molecules are embedded in the viral membrane

(Lamb and Krug, 1996).

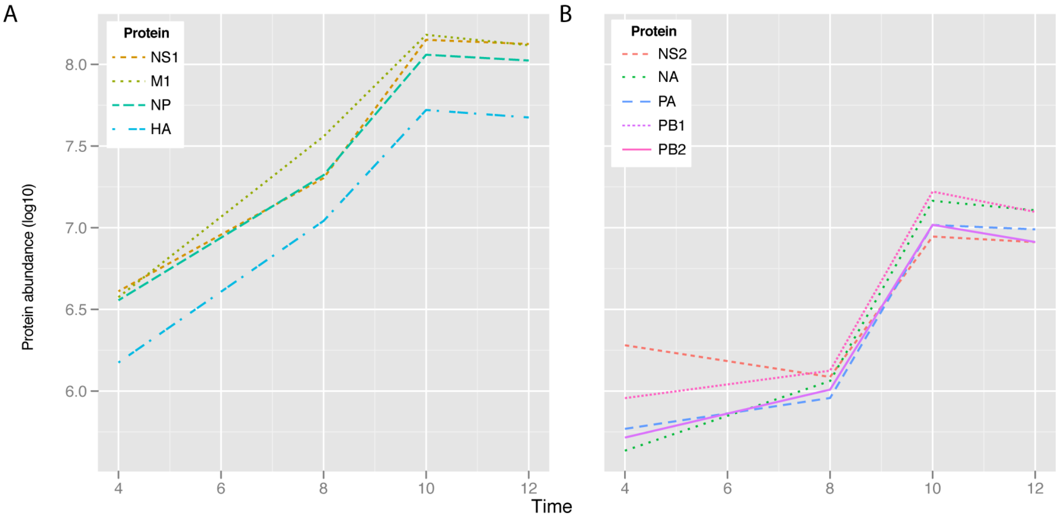


Figure 5.2.2: Viral proteins are mass produced upon infection, although with varying dynamics between the different protein groups.

5.2.2 Influenza Infection Influences Protein Production in the Host

While we could interpret the changes in the virus proteome by analyzing the temporal evolution of every single protein, this was an unfeasible task for more than a thousand mammalian proteins for four time points. It is also more informative to examine the regulation of whole modules of proteins rather than changes in only one single protein. We used statistical and bioinformatic methods to structure the dataset and interpret the dynamics of whole groups of proteins. I characterized these protein modules using functional databases to obtain an overview of what is happening in the host cells.

A fuzzy clustering of the host proteins showed 6 groups of genes that showed different dynamics altogether. This clustering approach has the advantage that one can filter the data *posterior* to the clustering based on the similarity it has to other data points (Section 2.2.1). This is an important feature for proteome data, because *a priori* filtering by e.g. minimum fold change would discard a substantial amount of data and other filtering methods from micro array systems cannot be used in this case. Due to the similar dynamics of the proteins in a cluster it is likely that these proteins are somehow functionally related to each other. To find out whether there are any functions that can be related to the dynamics, we performed different enrichment analyses of the proteins in each cluster (Section 2.2.2). We present the most significant terms combined with the dynamics of the cluster cores in Figure 5.2.3.

First of all we could not detect the overall breakdown of cellular protein production that we expected to happen, but a rather differentiated modulation of protein expression. The

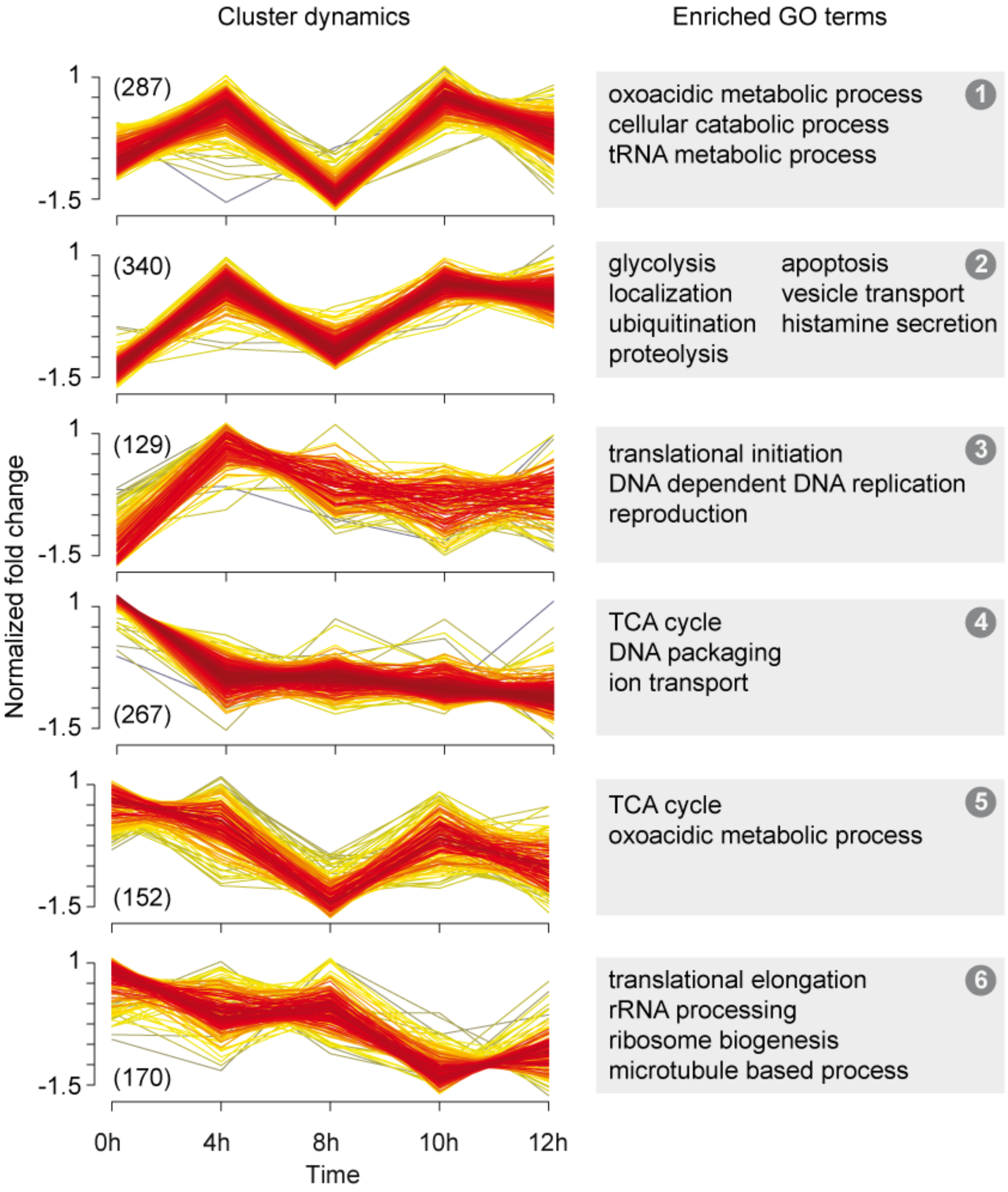


Figure 5.2.3: Proteins are clustered by their changes over time. All normalized temporal profiles were clustered by a fuzzy clustering algorithm to find modules of coregulated proteins. We performed enrichment tests for GO terms on each cluster for all proteins with a membership value >0.5 (n = number in brackets). The most significant terms are represented on the right panel.

10 hours post infection (p.i.) time point is of special interest, because infection is already established at this time, but the cells are still in a healthy condition. To get a more detailed picture of the changes at this time point, we performed an additional analysis called proteomic phenotyping (Section 2.2.3) that shows the differences between the uninfected cell and the 10 hour state in greater detail (Figure 5.2.1 and 5.2.5). The analysis confirmed the results of the dynamic clustering, and showed further differences.

Metabolism in the Early Influenza Infection

Clusters with a general downward trend in concentration like cluster 5 (Figure 5.2.3) show enriched GO terms like "TCA cycle", "cellular respiration", "respiratory electron transport chain", or "hypotonic response", all related to a downregulation of respiratory processes. On the other hand the clusters containing proteins that are produced in higher amounts (e.g. cluster 2) are connected to terms like "glycolysis" or "glucose metabolic process". On top of that, the analysis of the 10 hour time point showed an upregulation of pathways like "synthesis and degradation of ketone bodies" and "pentose phosphate way".

These results suggested to take a closer look at the metabolic pathways. The functional annotations of the cellular proteins show an increase in glycolytic and other metabolic pathway enzymes (Figure 5.2.4). As suggested by the analysis, also all measured concentrations of enzymes in the pentose phosphate pathway and the nucleotide synthesis increase over time. Strikingly, only the core glycolytic enzymes increase in abundance, but nearly all subunits of the pyruvate dehydrogenase complex (PDC) are less expressed over the course of infection, slowing the conversion of pyruvate to acetyl coenzyme A down. This would lead to an accumulation of intermediate metabolites of glycolysis and a lower flux through the TCA cycle as reported by Ritter et al. (2010). Downstream enzymes in the TCA cycle are nearly constant, while even further downstream enzymes in oxidative phosphorylation all decrease in concentration (Figure C.0.1). Changing the balance from oxidative phosphorylation to glycolysis energy production leads to a faster, but much more inefficient production of ATP. This redistribution of resources to different ATP producing pathways is a remarkable process in the infected cell. It is difficult to untangle what is the cause of this and what is the effect.

Ribosomal Proteins Become Less Abundant in the Course of Infection

The ribosome plays a special role in viral replication, because the virus needs it to produce the large amount of proteins to be able to generate new virions. This is achieved by different mechanisms like preventing cellular mRNA from leaving the nucleus (details in Section 5.1.1). This process needs to be regulated tightly to not disrupt the cells' protein production completely and keep it alive to produce viral offspring. In our data we see a general decrease in ribosomal proteins. We were able to quantify changes in 64 ribosomal proteins (of 79 known mammalian ribosomal proteins). Only 13 of these proteins increased in con-

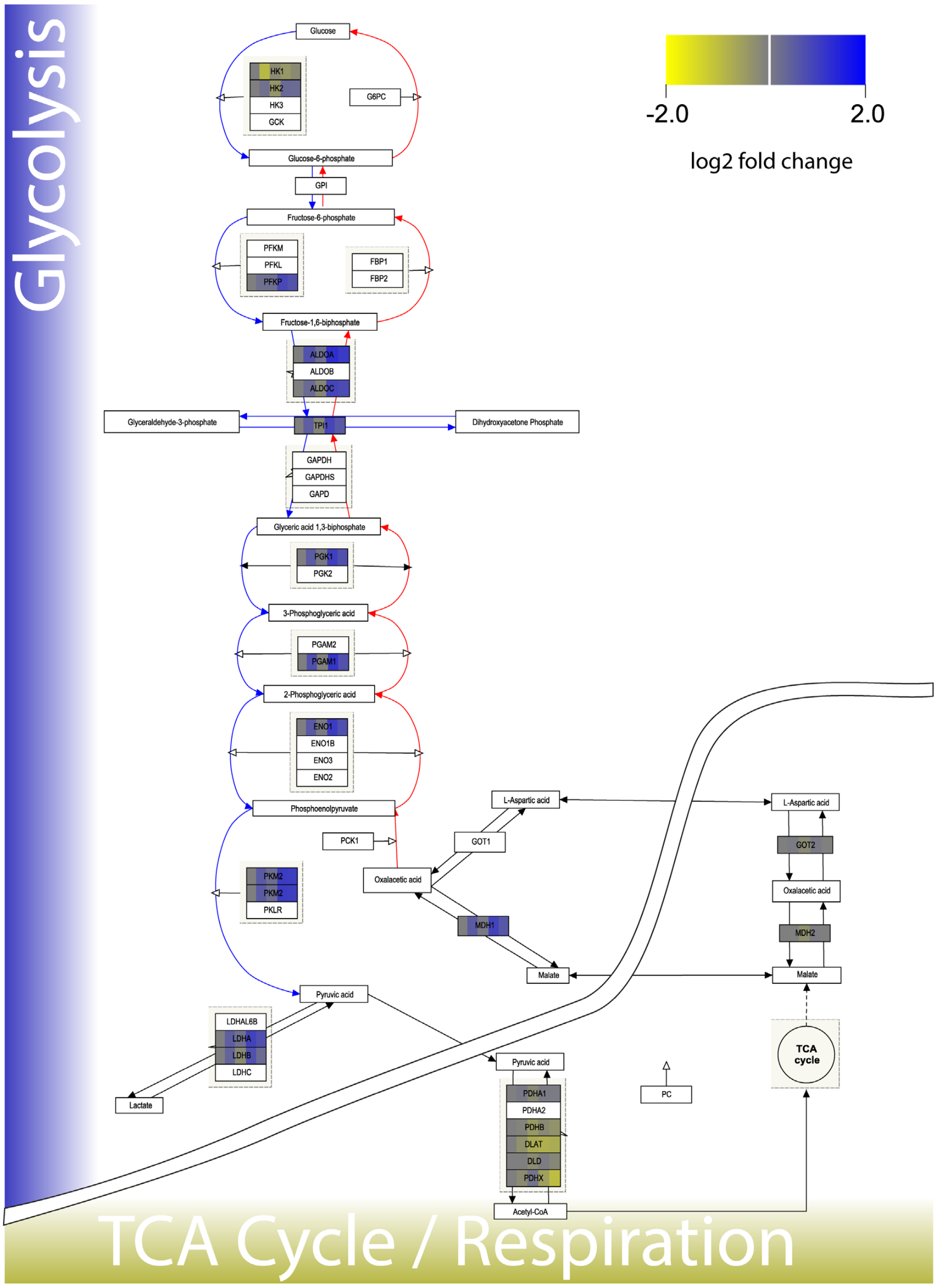


Figure 5.2.4: The quantified glycolysis proteins follow a similar pattern and most seem to be upregulated over the course of infection.

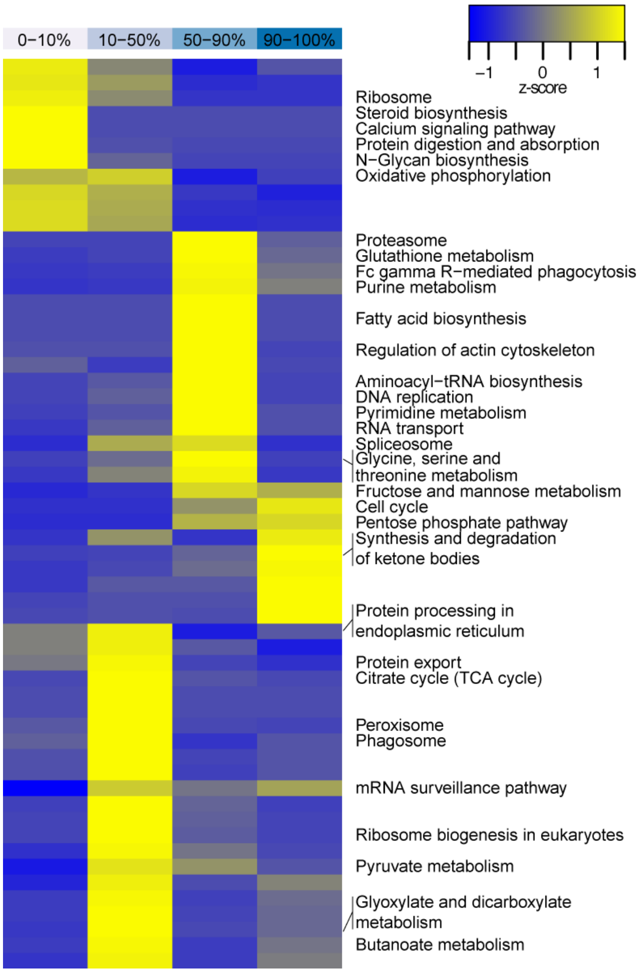


Figure 5.2.5: Proteomic phenotyping of the influenza A/PR/8 infected MDCK cell proteome at 10 hrs. p.i. using KEGG annotations. Quantiles of the quantification histogram are indicated at the top of the heatmap. Each quantile was analyzed separately for KEGG pathways and clustered by the z-transformed p-values. The most prominent representatives of all enriched biological processes of each quantile were selected and annotated.

centration over the course of the measurements, while the rest decreased uniformly. Those ribosomal proteins that increased over time stand out by their associated extra-ribosomal functions. Seven of the 13 proteins have been reported to possess such functions, while only 5 of the 51 remaining proteins do. Additionally, 3 of the proteins (RPS7, RPL23, and RPL5) are directly connected to p53 activation (Warner and McIntosh, 2009).

5.2.3 A Literature Derived Network Helps Interpreting Proteome Data

The large collection of data we gathered using the SILAC approach proves hard to interpret on its own. The picture changes when making use of already existing studies. There have been high-throughput studies before, generating a lot of data on virus host interaction using multiple techniques. These siRNA knockdown screens looked for genes that strongly affect infection and virus replication (e.g. Karlas et al., 2010; König et al., 2009). These valueable datasets were combined and used by Watanabe, Watanabe, and Kawaoka

(2010) to generate interaction networks of host and virus proteins. The authors created different networks showing the interactions inbetween host genes connected to the virus, between the vRNP and host genes, and between the virus proteins and the host. These networks contain only proteins that the authors found in at least two of the included studies and thereby have a higher accuracy than the original networks. We merged all these networks and used database identifiers to map our genes on the network. The outcome was a network of 65 proteins (Figure C.0.3). Topological analysis shows that most proteins interact with the vRNP and connections to other proteins are comparatively rare. This is due to many known interaction points in replication, translation, and transport of the vRNP. The interactions to the single proteins with cellular proteins are most prominent for the NS1 and the NP protein, which both also interact with the splicing and replication machinery. This network shows the connections between the dynamic changes in viral and host proteins that are evidently important for virus replication and enables us to show a global temporal picture of host proteins connected to their functions in viral replication (Figure 5.2.6).

5.2.4 Virus Host Interactions

As mentioned in the introduction of this chapter (Section 5.1.1), virus replication is a complex process that involves a large number of cellular proteins. Thus, we will describe it again step by step and relate it to the changes we found in both measured proteomes. In the following description we will refer to steps one to six marked in Figure 5.2.6. The virus enters the cell via endocytosis (1). For this step we detected an upregulation of proton pumps (ATP6V1) in the endosomal membrane, which leads to acidification and virus escape from the endosome. (2) The escaped vRNP enter the nucleus via the nuclear pore complex. We saw an early increase in the concentration of karyopherin β 1 (importin β) as well as two subunits of karyopherin α (importin α), which are both essential for this process (Mosammaparast and Pemberton, 2004). They all increased in concentration with roughly the same dynamics (Figure 5.2.6 KPNB1, KPNA3, KPNA6). (3) When the vRNP enters the nucleus, it starts to transcribe new vRNA and mRNA. We quantified 14 proteins that are connected to this process in one way or another. These proteins are mostly related to RNA processing. For example SRNP, PRPF4, and RBMX (among others), are components of the spliceosome. The dynamics of these proteins differ strongly, although belonging to the same functional group. (4) Nuclear export of vRNPs is mediated by NEP in conjunction with cellular proteins like nucleoporins. We were able to quantify 5 proteins related to this process and viral replication, which increased in concentration without exception, but showed a rather late activation. (5) Translation of proteins depends on the ribosome and on translation factors. Our data showed that the translation initiation factor complex EIF3 is strongly activated (six subunits with similar dynamics). This complex is responsible for the correct assembly of ribosomes and the recognition of AUG codons on the mRNA (Dong and Zhang, 2006). As mentioned earlier, most ribosomal proteins on the other hand decrease as well as EIF2 over the course of infection. (6) After translation, the virus proteins have to be transported to the cell membrane via the actin fibers, for which we also found an increased translation.

5.3 Discussion

The experiments and the extensive functional analysis presented in this chapter are the first proteome wide study of dynamics of influenza A infection covering both, host and virus proteins. The wide array of changes we detected in the proteome — albeit on a rather low level of change in most single proteins — tells us a lot about the interactions that are going on between virus and host. For this knowledge to be accessible, theoretical analysis is a prerequisite. Only in combination with existing data from functional databases and screening experiments could our data live up to its full potential.

Viral proteins mostly behave as expected and increased in abundance after infection quite rapidly. The two different dynamics of viral proteins were more surprising. Timing of viral protein expression that we observed can be explained when looking at the functions the proteins have in the course of the infection cycle. The first group with fast and early translation, mainly consists of structural proteins, or those that need to be present in high concentrations at an early phase of the infection cycle. HA, M1, and NP are the main structural proteins of the viral particle, and are therefore present in relatively high concentration directly after virus entry and need to be translated later to form new viral particles. NS1 on the other hand is purely non structural, but needs to be present in the early phase to counteract the hosts interferon response. It influences a number of important host processes like mRNA splicing (Marión et al., 1997), nuclear export (Qiu and Krug, 1994), mRNA polyadenylation (Nemeroff et al., 1998), and translation (Luna et al., 1995). One hypothesis could be that due to its wide array of binding partners NS1 needs to be present in high concentrations.

The second group consists of proteins that mainly serve enzymatic or catalytic functions. NS2/NEP is mainly responsible for exporting vRNPs from the nucleus, while the polymerase subunits PA, PB1, and PB2 as well as NA are enzymes. These proteins are not needed in high copy numbers in early replication, as they perform their functions multiple times. NA even serves no known functions inside the host cell and is only needed after virus budding. When virus budding starts, the requirements change and the enzymes switch their function to structural proteins, as the virus needs to include them into the budding particles. This explains the steep increase in copy numbers of the proteins in this group once the first particles start to bud after 10 hours p.i..

It appears that the interactions between host and viral systems are very complex. Activation of the majority of proteins interacting with the viral proteins and temporal inhibition of immune response genes could suggest that the virus is able to manipulate protein concentrations in the host for its own profit. Another interaction process seems to be the

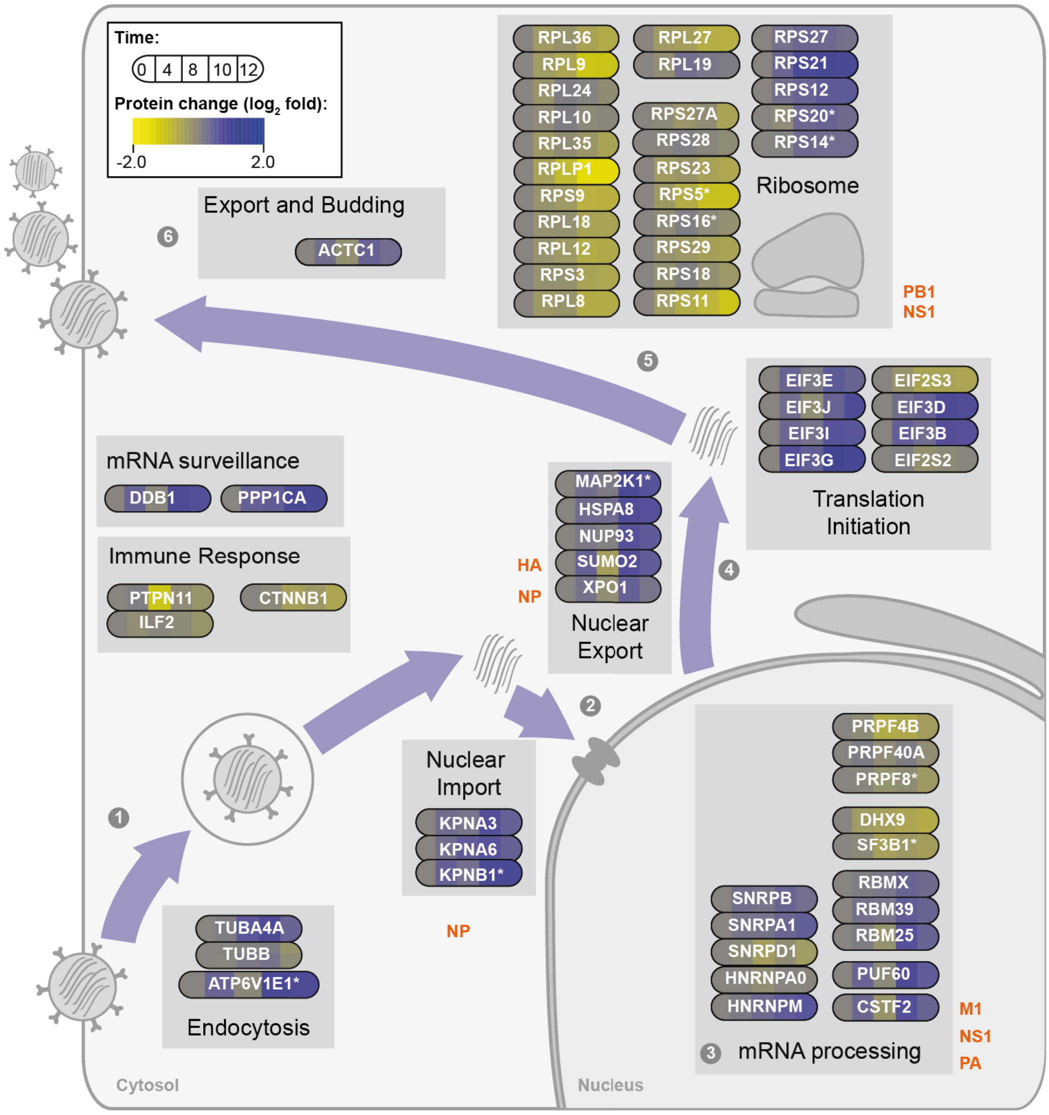


Figure 5.2.6: Schematic depictions of the viral life cycle and the connected host proteins. Protein symbols are divided into equal parts, each representing one time point and colored according to its relative concentration changes. Presented proteins were the overlap between proteins included in our dataset and genes interacting with viral proteins by Watanabe, Watanabe, and Kawaoka, 2010. These proteins were grouped by their function and put into context by the schematic graphics of the infection cycle.

shift from respiration to glycolysis that is also known from proliferating cells or tumor cells (e.g. Alberghina et al., 2012; Salminen and Kaarniranta, 2010). This could be a result of immune system induced apoptosis and mitochondrial degradation, but whether this dynamic change is only a cellular stress response or induced by the virus to gain a faster ATP production in a short time, remains to be elucidated.

One explanation could also be the competition of virus and host genes for other resources necessary for transcription and translation, e.g. tRNAs or amino acids. One could speculate that it is beneficial for the virus to keep the cell from producing the large machinery needed for cellular respiration, in order to save these resources for the viral proteins and mRNAs. It has to be mentioned that in a simple model of viral replication Sidorenko and Reichl (2004) suggested that these resources are no limiting factor for viral reproduction.

Another important resource for protein production are ribosomes. We see a correllated loss in concentration of ribosomal proteins that would lead to an even greater scarcity in ribosomes. The changes in the ribosomal proteins are even more surprising, as their mRNA as well as the proteins themselves are reported to generally have a very long halflife $\geq$100h (Schwanhäusser et al., 2011). This would suggest an even stronger breakdown in the production (or maybe active degradation) of these proteins than it seems on first sight. This lack of ribosomes could in the further course of infection lead to a bottleneck in viral production. Nucleolar changes and interactions to viral proteins, which could possibly lead to inhibition of ribosome production, have been shown before (e.g. Emmott et al., 2010; Murayama et al., 2007) in influenza infected MDCK cells (the same system we used) and should be investigated with a longer timecourse in the future. In this context, the differential regulation of ribosomal proteins is also an interesting result, as the expression of the group of ribosomal proteins with known extraribosomal functions (Warner and McIntosh, 2009) seems to be rather activated than inhibited.

The outcome of this study is only a first glance into the complex dynamic interactions and gives a birds eye view of the processes involved in infection and their timing. The amount of measured proteins is only a subset of the expressed proteome, which is biased towards abundant proteins due to experimental limitations. The combined measurement of viral and host proteins could also lead to a misidentification of proteins due to peptide similarities. A repeated study with longer mass spectrometry measurements and more time points would be helpful to follow the hints we got from this study more closely. Especially for the metabolic proteins (glycolysis, ATPase) data it would be interesting to extend the timecourse to 24h p.i. to be able to compare all time points to the metabolic data from Ritter et al. (2010). Another interesting follow up experiment would be a combination of this study with a RNAseq or microarray study to distinguish whether regulation happens on a transcriptional or translational level. From a bioinformatics perspective follow up experiments would be more informative if they are done in human cell lines, because the annotation of the human genome is much better than that of *Canis familiaris*. This would result in a more precise functional annotation and also in a better identification of proteins.

More detailed data would also allow modeling and parameterizations of existing models (Heldt, Frensing, and Reichl, 2012; Sidorenko and Reichl, 2004), which would be still very vague with the presented dataset. A dynamic model of the replication process highlighting the bottlenecks of infection based on real data would be a big step towards new anti-viral therapies, as it could identify new targets for anti-viral drugs using Systems Biology tools (Schulz, Bakker, and Klipp, 2009). The observed changes in ribosomal protein abundance and metabolism should be a center of attention for the model and data from other resources should be included in the model construction and parameterization.

Science is facts; just as houses are made of stones, so is science made of facts; but a pile of stones is not a house and a collection of facts is not necessarily science.

Henri Poincare

6

Discussion and Outlook

The goal of this thesis was to apply a Systems Biology approach to different eukaryotic systems and identify new important aspects in their regulation by data analysis and dynamic modeling. From a methodological perspective we also tried to find bottlenecks in the modeling process and widen these by the development of specialized software. The work was split up into three major parts. We investigated (1) the yeast signaling response to different stimuli via the MAP kinase system, (2) the response of human fibroblast cells to a viral transduction system carrying Yamanaka factors for reprogramming, and (3) the interaction between the influenza A virus and its host.

Major properties of all these seemingly different systems could be successfully investigated using similar theoretical methods that generated new insights about their regulation. For (1) we built two working models of different scope using ODE and Boolean frameworks and in parallel developed and verified the functionality of two software applications. In (2) we analyzed dynamic transcription profiles gathered during early reprogramming, and were able to identify roadblocks in the process of iPS generation. We used these findings to formulate a probabilistic Boolean model describing the process of reprogramming and the associated epigenetic regulations. For (3) we performed a functional analysis of the proteomes of virus and host in parallel over the course of the infection cycle. This analysis lead to new insights about the virus-host-interactions and the dynamics of infection.

6.1 From Data to Models

Mathematical models can advance our understanding of biological systems and clarify verbal hypothesis about these. Nevertheless, models can only be as good as the data they are based upon, and there is also no machine or software that could directly transform the available data into a useful model. One might say that data driven modeling is supposed to be propelled by data, but has to be steered by the modelers themselves.

To be able to build a model based on data, this data needs to be analyzed and put into context. Especially when using high-throughput data as presented in Chapters 4 and 5, this analysis is a crucial point for scientific progress. Functional analysis is needed to structure data and make it usable for hypothesis generation and construction of topologies and models (Chapter 5.2.2). Functional annotations of datasets were a large part of this Thesis and proved very useful for hypotheses generation, which is one step of the typical workflow of Systems Biology (Figure 1.3.1). A good example for this process can be found in chapter 4: Functional annotation of the dataset (and literature data) in combination with structural analysis lead to the realization that epigenetic marks and immune response block the early phases of reprogramming, which in turn lead to development of the modular model of epigenetic dynamics in reprogramming.

Although we were able to find new modeling approaches by analyzing the data extensively, we were not able to make further use of the different high-throughput datasets in terms of parameterization or quantification. This was mostly due to the low temporal resolution and the low precision of these approaches. This kind of data is mostly qualitative and can only be a first step for dynamic modeling to be able to generate abstract birds-eye-models. These models should establish a general understanding and be able to identify the most important parts of the system, which can then be investigated experimentally by more specific methods. The results could then be used to improve the level of detail of models in these parts and so on.

6.2 Modeling Software Catalyzes Progress

In the course of this work, models helped immensely to gain an initial understanding of the different systems we worked with already during initial literature studies. They proved to be an efficient way of scientific communication, which is still improving by advancing standardization efforts (Hucka et al., 2003). Standardization is not limited to file formats, but also aims towards visual model representation (Le Novère et al., 2009), which is an important feature for communication. Facilitating communication between modelers and experimenters is one of the most important features of the software we present in Chapter 3. This is done in different ways with both developed tools. *ModelMaGe* simplifies the process of generation, documentation, and validation of different model alternatives. This makes it easier to find a useful model, and also to present falsified models that lead to the end product of a working model. This helps other researchers to verify the results and understand the reasoning behind different model versions. As I mentioned before, annotation proved indispensable when dealing with large amounts of data. We strongly believe that the same holds true for further use of mathematical models, which is why *ModelMaGe* also supports automatic (although still limited) annotation of the generated models.

The way Boolean *rxncon* helps to communicate modeling results is a completely different approach. The most important aspects in this tool are its easy accessibility (freely available, without installation) and its standard compliant, dynamic, and interactive visualization features powered by the open Biographer framework (Figure 3.3.1). Integrating visualization into the model generation process will largely improve the mutual understanding of modelers and experimenters, while also highlighting problems in early modeling steps that can be discussed before further steps are taken. Additionally accepted standards for Boolean models are still missing, which hinders portability. The *rxncon* framework (Section 3.3) could be a first step towards a better adaptation of existing, or the creation of a new standard.

The other important goal besides communication, that both tools have in common, is to speed up the cycle of Systems Biology research at the modeling part. Often this part is the most non-linear and includes many internal loops before it leads on to the next phase. In many cases, models need to be created, parameterized, tested, and analyzed many times before the work can be presented and passed on to the experimenters. Both tools proved to help shorten and accelerate this inner loop in the presented examples and *ModelMaGe* has already been actively used in other projects (e.g. Klotz et al., 2011). Further iterations of the tools are currently under development.

6.3 Biological Advances

6.3.1 Somatic Cell Reprogramming

Somatic cell reprogramming is a rapidly changing field with thousands of published articles each year. However, we were able to add some new details to the view of the early stages of the reprogramming process. The most notable part is the effect of the innate immune system that we saw in the microarray data by upregulation of immune related genes. The second novelty we found in the data was the early onset of changes in morphological pathways. The importance of MET has been reported before (Li et al., 2010; Samavarchi-Tehrani et al., 2010), but has not been shown for such early time points in a viral reprogramming system.

Analysis of the data also has revealed the clearly existing epigenetic blocks of Yamanaka factor target genes, which have been shown before (Lister et al., 2009), but not in the context of early reprogramming. So far, there has been no analysis of the strength of the epigenetic inhibition for each target, which we roughly estimated with the NCA analysis.

Based on these experimental findings and our stochastic Boolean model we were able to propose possible methods for improving somatic cell reprogramming using, immune suppression, histone modification enzymes or their inhibitors. Our model also showed that the interplay between different epigenetic marks might be important for the velocity of chromatin remodeling during reprogramming. In the model we were able to improve reprogramming by faster DNA methylation and downregulation of somatic genes. DNA demethylation of pluripotency genes also proved to be an important step in the model, which might explain the reported improvements in efficiency of reprogramming by P53 knockout and faster proliferation (Hong et al., 2009; Kawamura et al., 2009), because demethylation is tightly coupled to cell cycle and proliferation (Cedar and Bergman, 2009).

6.3.2 Influenza Infection

In the analysis of proteomic data of the influenza A infection and its host we could find interesting dynamic changes in various host pathways that have not been reported before. Additionally we could show different abundances and dynamics of viral proteins during the infection cycle. These differences in viral proteins resemble previously reported stoichiometries and are nicely aligned with the proteins' reported functions during infection (Bouvier and Palese, 2008).

An interesting new finding of our analysis are the changes in the host's metabolic pathways that were not known before. The increase in glycolytic enzymes and the slight decrease in respiratory proteins probably leads to changes in the flux through the pathways and causes the major changes in the ATP metabolism of the cell that have been reported before (Ritter et al., 2010).

The second major change in the proteome of the host is the decrease of the ribosomal proteins over the course of infection. This could be caused by a deregulation of protein production by a strong production of viral proteins and lack of capacity of ribosomes to produce enough of their own proteins to sustain a steady state. Influences of the viral infection on the nucleolus and possibly ribosomal subunit production have been reported before (Emmott et al., 2010).

By mapping the dynamic data onto established interaction networks of infection and relating it to metabolome data, we have also taken the first steps towards data driven, detailed modeling of the infection process.

6.3.3 Different Levels - Different Perspective

As described in the introduction, the selection of a modeling framework is a crucial step in model creation. We had to gather a lot of literature data and knowledge about the specific system before we could choose a way to model it to answer the most pressing questions. The question one poses as well as the type of data available defines which framework is appropriate. Let us take the two yeast models in chapter 3 as an example: On the one hand, it would have been unfeasible to build and simulate the whole MAP kinase network as a kinetic ODE model, but the Boolean approach proved to be helpful in advancing the network structure and identify missing parts in the large network. On the other hand, only the quantitative modeling in combination with the detailed data for the HOG model enabled us to discriminate between the different alternatives and find the integrating feedback that predicted the additional experiments.

The more coarse grained approach for the large scale model can now be used to iteratively add more detail to the whole model or parts of it. The integration of our software into *rxncon* makes the switching between formalisms a lot easier than it used to be, this means we could take out one of the pathways and model it quantitatively once data becomes available and reintegrate the additional knowledge into the whole system.

6.4 All Levels are Equal

In the previous chapters we have discussed different regulatory systems. Although all of them act in roughly the same space (the cell), they all work on different timescales, from short lived signaling events (HOG activation in seconds and adaptation in minutes), over protein concentrations changing over the course of hours, to epigenetic marks on the chromatin that can last over generations of cells and sometimes organisms.

As expected, we have seen that all of these levels are important for the whole system to survive and that they are interconnected to a large degree. We found examples for such interconnections in each of the systems we investigated. The HOG pathway and its perfect adaptation via glycerol accumulation is a beautiful, well studied and simple example for the interplay of different levels, where signaling induces glycerol production via protein expression which results in silencing of the signal. The ribosome downregulation we found during the virus infection might be the opposite example of such an interplay, in which mRNA abundances influence protein production in such a way that the system gets out of control and protein production finally collapses because of a lack of ribosomes. The interplay between protein concentrations and epigenetic modifications described by our Boolean model of reprogramming are another example of an interplay of regulatory levels that is substantial for the development of the organism and for the transfer of short lived signals into long time marking.

In summary, we see the importance of a combination of upward and downward causation as it was proposed by Noble (2008) in all the systems we examined. To find the exact mechanisms behind interconnections of regulatory layers and to quantify their importance should be one of the primary goals of Systems Biology in the coming years. There are numerous unresolved questions in the specific systems we examined and there are also basic mechanisms that are not understood in this context. To name only some questions the results of this work pose: To which extend does the increase in protein concentration during influenza infection affect the flux of metabolites through glycolysis? How is the lower expression of genes of focal adhesion pathway affecting signaling and morphology in the cells? What is the exact mechanism connecting the expression of Yamanaka factors and epigenetic reprogramming events? To us, one of the most challenging general question is, how the proteins regulating chromatin structure find their targets and how do so few general modifying enzymes control the structure of the whole genome?

6.5 Outlook and Further Work

Our models generated a number of testable hypothesis and the data analyses gave hints where to focus further research. In Chapter 4.2 we showed the negative influence of immune response on iPS generation, which should further be investigated to produce better reprogramming protocols. Screenings with libraries of immune suppressors for increasing efficiency would be one possible way to go forward. Our model of epigenetic changes in reprogramming would benefit from better data on the single modules of gene regulation. With such data we could use the model as a scaffold for more detailed models of transcription factor interactions inside the modules to be able to specify the update rule probabilities based on data. It should also be tested whether our model can reflect recent single cell data on reprogramming that profiled genetic variation between cells during reprogramming (Buganim et al., 2012). This data could be directly compared to the different probability distributions our model produces. Using similar single cell techniques, one could also test if the early repression of somatic genes is really a predestining step for efficient reprogramming as our model suggests.

The new aspects of Boolean modeling we propose in Section 3.3.2 and 4.2.2 deserve further development and theoretical investigation. The model in Section 4.2.2 could also be analyzed further by for example calculating influence factors for all the nodes, which gives the relative importance of one node in the network. It would also be interesting to analyze single Boolean networks from the PBN using methods for attractor landscape generation like the one presented by Choi et al. (2012).

The PBN approach we used could also be included into the Boolean *rxncon* software to make PBN generation more user friendly and faster. The presented software tools are under constant development and will be improved in usability and featureset. *ModelMaGe* is currently being closely integrated with *Copasi* to make the parameter fitting even simpler and the whole tool more reliable. We hope that the open source licensing (LGPL) of both tools will help to attract further users and developers.

In the influenza A study we managed to extract interesting features from the large body of data. To test which potential impact these features might have on infection, we could add them to existing models of influenza A infection (Heldt, Frensing, and Reichl, 2012; Sidorenko and Reichl, 2004). On the contrary, to be able to improve these models by parameter fitting with our data, we would certainly need to reduce them to a fraction of their current sizes. A combination of the new features of the host reaction and a simplified general model of infection would be one possible way to find out how these changes in the host might impact infection. Nevertheless, to parameterize such a model in a meaningful way we would need more data with a higher time resolution for the important proteins. The high-troughput study identified candidates of host proteins that showed strong fluctuations and are responsible for key events in the infection cycle (e.g. lysosomal proton pump ATP6V). These could potentially be measured with precise low-throughput techniques to verify our results. To verify the hypothesis about ribosme loss, another MS study would be needed that could be more targeted towards ribosomal proteins.

6.6 Concluding Remarks

The presented work once again highlights the importance of a Systems Biological approach for the understanding of complex biological systems as eukaryotic cells. None of the examined systems could be understood in isolation, and the effects we found in experiments would be hard to explain without a broader context and will require additional work on both experimental and theoretical sides.

Bibliography

Adewumi, O, B Aflatoonian, and L Ahrlund-Richter (July 2007). "Characterization of human embryonic stem cell lines by the International Stem Cell Initiative". In: *Nature* 25.7, pp. 803–816.

Akaike, H (1974). "A new look at the statistical model identification". In: *Automatic Control, IEEE Transactions on*.

Alberghina, Lilia et al. (Jan. 2012). "Cancer cell growth and survival as a system-level property sustained by enhanced glycolysis and mitochondrial metabolic remodeling." In: *Frontiers in physiology* 3, p. 362.

Albert, István et al. (Jan. 2008). "Boolean network simulations for life scientists." In: *Source code for biology and medicine* 3, p. 16.

Albert, Réka and Hans G Othmer (July 2003). "The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster". In: *Journal of Theoretical Biology* 223.1, pp. 1–18.

Ang, Yen-Sin et al. (July 2011). "Stem cells and reprogramming: breaking the epigenetic barrier?" In: *Trends in pharmacological sciences* 32.7, pp. 394–401.

Artyomov, M.N., Alexander Meissner, and A.K. Chakraborty (May 2010). "A model for genetic and epigenetic regulatory networks identifies rare pathways for transcription factor induced pluripotency". In: *PLoS computational biology* 6.5, e1000785.

Ashburner, M et al. (May 2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." In: *Nature genetics* 25.1, pp. 25–9.

Bao, MZ, MA Schwartz, and GT Cantin (2004). "Pheromone-dependent destruction of the Tec1 transcription factor is required for MAP kinase signaling specificity in yeast". In: *Cell*.

Bardwell, Lee (Feb. 2005). "A walk-through of the yeast mating pheromone response pathway". In: *Peptides* 26.2, pp. 339–350.

Baudin, F et al. (July 1994). "Structure of influenza virus RNP. I. Influenza virus nucleoprotein melts secondary structure in panhandle RNA and exposes the bases to the solvent." In: *The EMBO journal* 13.13, pp. 3158–65.

Bauer, Amy L et al. (June 2010). "Receptor cross-talk in angiogenesis: mapping environmental cues to cell phenotype using a stochastic, Boolean signaling network model." In: *Journal of theoretical biology* 264.3, pp. 838–46.

Berg, Debbie L.C. van den et al. (2010). "An Oct4-Centered Protein Interaction Network in Embryonic Stem Cells". In: *Cell Stem Cell* 6.4, pp. 369–381.

Bhutani, Nidhi, David M. Burns, and Helen M. Blau (Sept. 2011). "DNA demethylation dynamics." In: *Cell* 146.6, pp. 866–72.

Botstein, David and Gerald R Fink (Nov. 2011). "Yeast: an experimental organism for 21st Century biology." In: *Genetics* 189.3, pp. 695–704.

Bouvier, Nicole and Peter Palese (Sept. 2008). "The biology of influenza viruses". In: *Vaccine* 26.null, pp. D49–D53.

Box, GEP (1976). "Science and statistics". In: *Journal of the American Statistical Association*.

Boyer, Laurie A. et al. (Sept. 2005). "Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells". In: *Cell* 122.6, pp. 947–956.

Buganim, Yosef et al. (Sept. 2012). "Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase." In: *Cell* 150.6, pp. 1209–22.

Burnham, Kenneth P. and David R. Anderson (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach (Google eBook)*. Springer, p. 488.

Cedar, Howard and Yehudit Bergman (2009). "Linking DNA methylation and histone modification: patterns and paradigms." In: *Nature Reviews Genetics* 10.5, pp. 295–304.

Chang, Rui, Robert Shoemaker, and Wei Wang (Dec. 2011). "Systematic Search for Recipes to Generate Induced Pluripotent Stem Cells". In: *PLoS Computational Biology* 7.12. Ed. by Denis Thieffry, e1002300.

Chavez, Lukas et al. (2009a). "BMC Genomics". In: *BMC Genomics* 14, pp. 1–14.

Chavez, Lukas et al. (2009b). "In silico identification of a core regulatory network of OCT4 in human embryonic stem cells using an integrated approach". In: *BMC Genomics* 10.1, p. 314.

Cheung, Timothy K W and Leo L M Poon (Apr. 2007). "Biology of influenza a virus." In: *Annals of the New York Academy of Sciences* 1102, pp. 1–25.

Chickarmane, Vijay and Carsten Peterson (2008). "A Computational Model for Understanding Stem Cell, Trophectoderm and Endoderm Lineage Determination". In: *PLoS ONE* 3.10, e3478.

Chickarmane, Vijay, Carl Troein, and UA Nuber (2006). "Transcriptional dynamics of the embryonic stem cell switch". In: *PLoS computational* 2.9, e123.

Choi, M. et al. (Nov. 2012). "Attractor Landscape Analysis Reveals Feedback Loops in the p53 Network That Control the Cellular Response to DNA Damage". In: *Science Signaling* 5.251, ra83–ra83.

Cleveland, William S. (Dec. 1979). "Robust Locally Weighted Regression and Smoothing Scatterplots". In: *Journal of the American Statistical Association* 74.368, p. 829.

Cox, Jürgen and Matthias Mann (Dec. 2008). "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification." In: *Nature biotechnology* 26.12, pp. 1367–72.

Crick, F (Aug. 1970). "Central dogma of molecular biology." In: *Nature* 227.5258, pp. 561–3.

Cros, Jerome and Peter Palese (Sept. 2003). "Trafficking of viral genomic RNA into and out of the nucleus: influenza, Thogoto and Borna disease viruses". In: *Virus Research* 95.1-2, pp. 3–12.

Di Cara, Alessandro et al. (Jan. 2007). "Dynamic simulation of regulatory networks using SQUAD." In: *BMC bioinformatics* 8.1, p. 462.

Djuric, Ugljesa and James Ellis (Jan. 2010). "Epigenetics of induced pluripotency, the seven-headed dragon." In: *Stem cell research & therapy* 1.1, p. 3.

Dodd, Ian B. et al. (May 2007). "Theoretical Analysis of Epigenetic Cell Memory by Nucleosome Modification". In: *Cell* 129.4, pp. 813–822.

Dong, Zizheng and Jian-Ting Zhang (Sept. 2006). "Initiation factor eIF3 and regulation of mRNA translation, cell growth, and cancer." In: *Critical reviews in oncology/hematology* 59.3, pp. 169–80.

Duff, Campbell et al. (Feb. 2012). "Mathematical modelling of stem cell differentiation: the PU.1-GATA-1 interaction." In: *Journal of mathematical biology* 64.3, pp. 449–68.

Dunn, J.C. (1973). "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters". In: *Journal of Cybernetics*.

Emmott, Edward et al. (2010). "Quantitative Proteomics Using SILAC Coupled to LC-MS / MS Reveals Changes in the Nucleolar Proteome in Influenza A Virus-Infected Cells research articles". In: *Journal of Proteome Research*, pp. 5335–5345.

Epsztejn-Litman, Silvina et al. (Nov. 2008). "De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes." In: *Nature structural & molecular biology* 15.11, pp. 1176–83.

Evans, MJ and MH Kaufman (1981). "Establishment in culture of pluripotential cells from mouse embryos." In: *Nature* 292.(5819), pp. 154–6.

Falcon, S and R Gentleman (Jan. 2007). "Using GOstats to test gene lists for GO term association." In: *Bioinformatics (Oxford, England)* 23.2, pp. 257–8.

Flöttmann, Max (2008). "Autmatic Model Generation in Saccharomyces Cerevisiae with ModelMaGe". Master. FU Berlin.

Flöttmann, Max, Till Scharp, and Edda Klipp (Jan. 2012). "A stochastic model of epigenetic dynamics in somatic cell reprogramming." In: *Frontiers in physiology* 3.June, p. 216.

Flöttmann, Max et al. (2008). "ModelMage: A Tool for Automatic Model Generation, Selection and Management". In: *Genome Informatics*, pp. 52–63.

Flöttmann, Max et al. (2012). "Reaction-contingency based bipartite Boolean modelling". In: *under review*.

Fouchier, Ron A M et al. (Mar. 2005). "Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls." In: *Journal of virology* 79.5, pp. 2814–22.

Francastel, C et al. (Nov. 2000). “Nuclear compartmentalization and gene activity.” In: *Nature reviews. Molecular cell biology* 1.2, pp. 137–43.

Fujii, Yutaka et al. (Feb. 2003). “Selective incorporation of influenza virus RNA segments into virions.” In: *Proceedings of the National Academy of Sciences of the United States of America* 100.4, pp. 2002–7.

Fuks, F et al. (Jan. 2000). “DNA methyltransferase Dnmt1 associates with histone deacetylase activity.” In: *Nature genetics* 24.1, pp. 88–91.

Funahashi, A. et al. (Aug. 2008). “CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks”. In: *Proceedings of the IEEE* 96.8, pp. 1254–1265.

Futschik, Matthias E and Bronwyn Carlisle (Aug. 2005). “Noise-robust soft clustering of gene expression time-course data.” In: *Journal of bioinformatics and computational biology* 3.4, pp. 965–88.

García-Rodríguez, Luis J et al. (Nov. 2005). “Cell integrity signaling activation in response to hyperosmotic shock in yeast.” In: *FEBS letters* 579.27, pp. 6186–90.

Garg, Abhishek et al. (June 2009). “Modeling stochasticity and robustness in gene regulatory networks.” In: *Bioinformatics (Oxford, England)* 25.12, pp. i101–9.

Gentleman, Robert C et al. (Jan. 2004). “Bioconductor: open software development for computational biology and bioinformatics.” In: *Genome biology* 5.10, R80.

Halley, Julianne D, Frank R Burden, and David a Winkler (May 2009). “Stem cell decision making and critical-like exploratory networks.” In: *Stem cell research* 2.3, pp. 165–77.

Hanna, Jacob H, Krishanu Saha, and Rudolf Jaenisch (2010). “Pluripotency and Cellular Reprogramming : Facts , Hypotheses , Unresolved Issues”. In: *Cell* 143.4, pp. 508–525.

Hanna, Jacob et al. (Nov. 2009). “Direct cell reprogramming is a stochastic process amenable to acceleration”. In: *Nature* advance on.

Hao, Nan et al. (Apr. 2007). “A systems-biology analysis of feedback inhibition in the Sho1 osmotic-stress-response pathway.” In: *Current biology : CB* 17.8, pp. 659–67.

Hashimshony, Tamar et al. (June 2003). “The role of DNA methylation in setting up chromatin structure during development.” In: *Nature genetics* 34.2, pp. 187–92.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning*. Springer, pp. 139–189.

Hawkins, R. David et al. (May 2010). “Distinct Epigenomic Landscapes of Pluripotent and Lineage-Committed Human Cells”. In: *Cell Stem Cell* 6.5, pp. 479–491.

Heldt, Frank S, Timo Frensing, and Udo Reichl (Aug. 2012). “Modeling the intracellular dynamics of influenza virus replication to understand the control of viral RNA synthesis.” In: *Journal of virology* 86.15, pp. 7806–17.

Hochedlinger, Konrad and Rudolf Jaenisch (Feb. 2002). “Monoclonal mice generated by nuclear transfer from mature B and T donor cells.” In: *Nature* 415.6875, pp. 1035–8.

Hockemeyer, Dirk et al. (Sept. 2008). “A drug-inducible system for direct reprogramming of human somatic cells to pluripotency.” In: *Cell stem cell* 3.3, pp. 346–53.

Hohmann, Stefan (Dec. 2009). “Control of high osmolarity signalling in the yeast Saccharomyces cerevisiae.” In: *FEBS letters* 583.24, pp. 4025–9.

Hong, Hyenjong et al. (Aug. 2009). “Suppression of induced pluripotent stem cell generation by the p53-p21 pathway.” In: *Nature* 460.7259, pp. 1132–5.

Hoops, Stefan et al. (Dec. 2006). “COPASI–a COmplex PAthway SImulator.” In: *Bioinformatics (Oxford, England)* 22.24, pp. 3067–74.

Hotta, Akitsu and James Ellis (2008). “Retroviral vector silencing during iPS cell induction: An epigenetic beacon that signals distinct pluripotent states”. In: *Journal of Cellular Biochemistry* 105.4, pp. 940–948.

Hsu, M.-T. (Nov. 1987). “Genomic RNAs of Influenza Viruses are Held in a Circular Conformation in Virions and in Infected Cells by a Terminal Panhandle”. In: *Proceedings of the National Academy of Sciences* 84.22, pp. 8140–8144.

Huang, Sui et al. (May 2007). “Bifurcation dynamics in lineage-commitment in bipotent progenitor cells.” In: *Developmental biology* 305.2, pp. 695–713.

Huangfu, Danwei et al. (July 2008a). “Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds.” In: *Nature biotechnology* 26.7, pp. 795–7.

Huangfu, Danwei et al. (Nov. 2008b). “Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2.” In: *Nature biotechnology* 26.11, pp. 1269–75.

Hucka, M. et al. (Mar. 2003). “The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models”. In: *Bioinformatics* 19.4, pp. 524–531.

Ichida, JK et al. (Nov. 2009). “A small-molecule inhibitor of Tgf- signaling replaces Sox2 in reprogramming by inducing Nanog”. In: *Cell Stem Cell* 5.5, pp. 491–503.

Ivanova, Natalia et al. (2006). “Dissecting self-renewal in stem cells with RNA interference”. In: *Nature* 442.7102, pp. 533–538.

Jagger, B W et al. (July 2012). “An overlapping protein-coding region in influenza A virus segment 3 modulates the host response.” In: *Science (New York, N.Y.)* 337.6091, pp. 199–204.

Joshi-Tope, G and M Gillespie (2005). “Reactome: a knowledgebase of biological pathways”. In: *Nucleic acids ...*.

Kalmar, Tibor et al. (July 2009). “Regulated Fluctuations in Nanog Expression Mediate Cell Fate Decisions in Embryonic Stem Cells”. In: *PLoS Biol* 7.7, e1000149.

Kanehisa, M. (Jan. 2000). “KEGG: Kyoto Encyclopedia of Genes and Genomes”. In: *Nucleic Acids Research* 28.1, pp. 27–30.

Kao, K.C. et al. (Jan. 2004). “Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis”. In: *Proceedings of the National Academy of Sciences of the United States of America* 101.2, p. 641.

Karlas, Alexander et al. (Feb. 2010). “Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication.” In: *Nature* 463.7282, pp. 818–22.

Kauffman, Stuart and Carsten Peterson (Dec. 2003). “Random Boolean network models and the yeast transcriptional network”. In: *Proceedings of the* 100.25, pp. 14796–14799.

Kauffmann, Stuart (Oct. 1969). “Homeostasis and Differentiation in Random Genetic Control Networks”. In: *Nature* 224.5215, pp. 177–178.

Kawamura, Teruhisa et al. (Aug. 2009). “Linking the p53 tumour suppressor pathway to somatic cell reprogramming.” In: *Nature* 460.7259, pp. 1140–4.

Kell, D B (July 1979). “On the functional proton current pathway of electron transport phosphorylation. An electrodic view.” In: *Biochimica et biophysica acta* 549.1, pp. 55–99.

Kim, Jonghwan et al. (Mar. 2008). “An extended transcriptional network for pluripotency of embryonic stem cells”. In: *Cell* 132.6, pp. 1049–1061.

Kirkpatrick, S and MP Vecchi (1983). “Optimization by simmulated annealing”. In: *Science*.

Kitano, H (2002). “System Biology: A Brief Overview.” In: *Science* 295, pp. 1662–1664.

Klipp, E, B Nordlander, and R Krüger (2005). “Integrative model of the response of yeast to osmotic shock”. In: *Nature …*.

Klipp, Edda et al. (Apr. 2007). “Systems biology standards–the community speaks.” In: *Nature biotechnology* 25.4, pp. 390–1.

Klipp, Edda et al. (2009). *Systems Biology*. Wiley-Blackwell, p. 592.

Klotz, Christian et al. (Jan. 2011). “A helminth immunomodulator exploits host signaling events to regulate cytokine production in macrophages.” In: *PLoS pathogens* 7.1. Ed. by Thomas A. Wynn, e1001248.

König, Matthias, Andreas Dräger, and Hermann-Georg Holzhütter (Sept. 2012). “CySBML: a Cytoscape plugin for SBML.” In: *Bioinformatics (Oxford, England)* 28.18, pp. 2402–3.

König, Renate et al. (2009). “Human host factors required for influenza virus replication”. In: *Nature* 463.7282, pp. 813–817.

Krause, F et al. (2010). “Annotation and merging of SBML models with semanticSBML”. In: … .

Lamb, R A and R M Krug (1996). *Orthomyxoviridae: The viruses and their Replication*. 3rd ed., pp. 1353–1445.

Laurent, Louise et al. (Mar. 2010). “Dynamic changes in the human methylome during differentiation”. In: *Genome Research* 20.3, pp. 320–331.

Laver, W G et al. (Sept. 1984). “Influenza virus neuraminidase with hemagglutinin activity.” In: *Virology* 137.2, pp. 314–23.

Le Novère, Nicolas et al. (Aug. 2009). “The Systems Biology Graphical Notation.” In: *Nature biotechnology* 27.8, pp. 735–41.

Lehembre, Francois et al. (Oct. 2008). “NCAM-induced focal adhesion assembly: a functional switch upon loss of E-cadherin.” In: *The EMBO journal* 27.19, pp. 2603–15.

Levin, David E (Dec. 2011). "Regulation of cell wall biogenesis in Saccharomyces cerevisiae: the cell wall integrity signaling pathway." In: *Genetics* 189.4, pp. 1145–75.

Li, E, T H Bestor, and R Jaenisch (June 1992). "Targeted mutation of the DNA methyltransferase gene results in embryonic lethality." In: *Cell* 69.6, pp. 915–26.

Li, Ronghui et al. (June 2010). "A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts". In: *Cell Stem Cell*, pp. 51–63.

Liang, Jiancong et al. (June 2008). "Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells." In: *Nature cell biology* 10.6, pp. 731–9.

Liao, James C. et al. (Dec. 2003). "Network component analysis: Reconstruction of regulatory signals in biological systems". In: *Proceedings of the National Academy of Sciences of the United States of America* 100.26, pp. 15522–15527.

Lister, Ryan et al. (Nov. 2009). "Human DNA methylomes at base resolution show widespread epigenomic differences". In: *Nature* 462.7271, pp. 315–322.

Lister, Ryan et al. (Feb. 2011). "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells". In: *Nature*.

Loh, Yuin-Han et al. (Apr. 2006). "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells". In: *Nat Genet* 38.4, pp. 431–440.

Luna, S de la et al. (Apr. 1995). "Influenza virus NS1 protein enhances the rate of translation initiation of viral mRNAs." In: *Journal of virology* 69.4, pp. 2427–33.

MacArthur, Ben D. BD, Colin P. CP Please, and Richard O. C. Oreffo (2008). "Stochasticity and the molecular mechanisms of induced pluripotency". In: *PLoS One* 3.8, e3086.

MacQueen, J. B. (1967). "Some Methods for Classification and Analysis of MultiVariate Observations". In: *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by L. M. Le Cam and J. Neyman. Vol. 1. University of California Press, pp. 281–297.

Mah, Nancy et al. (Jan. 2011). "Molecular Insights into Reprogramming-Initiation Events Mediated by the OSKM Gene Regulatory Network." In: *PloS one* 6.8, e24351.

Maiwald, Thomas and Jens Timmer (2008). "Dynamical modeling and multi-experiment fitting with PottersWheel". In: *Bioinformatics* 24.18, pp. 2037–2043.

Manney, TR (1983). "Expression of the BAR1 gene in Saccharomyces cerevisiae: induction by the alpha mating pheromone of an activity associated with a secreted protein." In: *Journal of bacteriology*.

Marión, R M et al. (Oct. 1997). "Influenza virus NS1 protein interacts with viral transcription-replication complexes in vivo." In: *The Journal of general virology* 78 (Pt 10, pp. 2447–51.

Marión, Rosa M et al. (Aug. 2009). "A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity." In: *Nature* 460.7259, pp. 1149–53.

Maximov, Alexander (1909). “Der Lymphozyt als gemeinsame Stammzelle der verschiedenen Blutelemente in der embryonalen Entwicklung und im postfetalen Leben der Säugetiere”. In: *Folia Haematologica* 8.1909, pp. 125–134.

Meissner, Alexander (2010). “Epigenetic modifications in pluripotent and differentiated cells”. In: *Nature biotechnology* 28.10, pp. 1079–1088.

Meshorer, Eran et al. (Jan. 2006). “Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells.” In: *Developmental cell* 10.1, pp. 105–16.

Mettetal, JT and D Muzzey (2008). “The frequency dependence of osmo-adaptation in Saccharomyces cerevisiae”. In: *Science ...* .

Mikkelsen, Tarjei S. et al. (July 2008). “Dissecting direct reprogramming through integrative genomic analysis”. In: *Nature* 454.7200, pp. 49–55.

Monk, M, R L Adams, and A Rinaldi (May 1991). “Decrease in DNA methylase activity during preimplantation development in the mouse.” In: *Development (Cambridge, England)* 112.1, pp. 189–92.

Mosammaparast, Nima and Lucy F Pemberton (Oct. 2004). “Karyopherins: from nuclear-transport mediators to nuclear-function regulators.” In: *Trends in cell biology* 14.10, pp. 547–56.

Murayama, Rikinori et al. (Nov. 2007). “Influenza A virus non-structural protein 1 (NS1) interacts with cellular multifunctional protein nucleolin during infection.” In: *Biochemical and biophysical research communications* 362.4, pp. 880–5.

Müssel, Christoph, Martin Hopfensitz, and Hans A Kestler (May 2010). “BoolNet–an R package for generation, reconstruction and analysis of Boolean networks.” In: *Bioinformatics (Oxford, England)* 26.10, pp. 1378–80.

Muzzey, D and CA Gómez-Uribe (2009). “A systems-level analysis of perfect adaptation in yeast osmoregulation”. In: *Cell*.

Nemeroff, M E et al. (June 1998). “Influenza virus NS1 protein interacts with the cellular 30 kDa subunit of CPSF and inhibits 3’end formation of cellular pre-mRNAs.” In: *Molecular cell* 1.7, pp. 991–1000.

Nishimura, Shigeko et al. (2000). “A GATA Box in the GATA-1 Gene Hematopoietic Enhancer Is a Critical Element in the Network of GATA Factors and Sites That Regulate This Gene A GATA Box in the GATA-1 Gene Hematopoietic Enhancer Is a Critical Element in the Network of GATA Factors and Sites”. In: *Society*.

Nishino, Koichiro et al. (May 2011). “DNA Methylation Dynamics in Human Induced Pluripotent Stem Cells over Time.” In: *PLoS genetics* 7.5, e1002085.

Niwa, Hitoshi (Feb. 2007). “How is pluripotency determined and maintained?” In: *Development* 134.4, pp. 635–46.

Niwa, Hitoshi et al. (Dec. 2005). “Interaction between Oct3/4 and Cdx2 Determines Trophectoderm Differentiation”. In: *Cell* 123.5, pp. 917–929.

Noble, D (2008). “Claude Bernard, the first systems biologist, and the future of physiology”. In: *Experimental Physiology*.

Noble, Dennis (1960). "Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations." In: *Nature* 188, pp. 495–7.

Okita, Keisuke et al. (Nov. 2008). "Generation of mouse induced pluripotent stem cells without viral vectors." In: *Science (New York, N.Y.)* 322.5903, pp. 949–53.

Okuno, Yutaka et al. (2005). "Potential Autoregulation of Transcription Factor PU . 1 by an Upstream Regulatory Element Potential Autoregulation of Transcription Factor PU . 1 by an Upstream Regulatory Element". In: *Society*.

ONeill, R E, J Talon, and P Palese (Jan. 1998). "The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins." In: *The EMBO journal* 17.1, pp. 288–96.

Ong, S.-E. (May 2002). "Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics". In: *Molecular & Cellular Proteomics* 1.5, pp. 376–386.

Ong, Shao-En and Matthias Mann (Oct. 2005). "Mass spectrometry-based proteomics turns quantitative." In: *Nature chemical biology* 1.5, pp. 252–62.

Orlando, David A et al. (June 2008). "Global control of cell-cycle transcription by coupled CDK and network oscillators." In: *Nature* 453.7197, pp. 944–7.

Ou, Jing-Ni et al. (May 2007). "Histone deacetylase inhibitor Trichostatin A induces global and gene-specific DNA demethylation in human cancer cell lines." In: *Biochemical pharmacology* 73.9, pp. 1297–307.

Pan, Cuiping et al. (Mar. 2009). "Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions." In: *Molecular & cellular proteomics : MCP* 8.3, pp. 443–50.

Pannell, D et al. (Nov. 2000). "Retrovirus vector silencing is de novo methylase independent and marked by a repressive histone code." In: *The EMBO journal* 19.21, pp. 5884–94.

Papp, Bernadett and Kathrin Plath (Mar. 2011). "Reprogramming to pluripotency: stepwise resetting of the epigenetic landscape." In: *Cell research* 21.3, pp. 486–501.

Plotch, Stephen J. et al. (Mar. 1981). "A unique cap(m7GpppXm)-dependent influenza virion endonuclease cleaves capped RNAs to generate the primers that initiate viral RNA transcription". In: *Cell* 23.3, pp. 847–858.

Qiu, Y and R M Krug (Apr. 1994). "The influenza virus NS1 protein is a poly(A)-binding protein that inhibits nuclear export of mRNAs containing poly(A)." In: *J. Virol.* 68.4, pp. 2425–2432.

Qu, K and P Ortoleva (Feb. 2008). "Understanding stem cell differentiation through self-organization theory." In: *Journal of theoretical biology* 250.4, pp. 606–20.

Ralston, A and J Rossant (Aug. 2005). "Genetic regulation of stem cell origins in the mouse embryo." In: *Clinical genetics* 68.2, pp. 106–12.

Rekhtman, Natasha et al. (1999). “GATA-1 : functional antagonism in erythroid cells Direct interaction of hematopoietic transcription factors PU . 1 and GATA-1 : functional antagonism in erythroid cells”. In: *Genes & Development*, pp. 1398–1411.

Richardson, J. C. and R. K. Akkina (Mar. 1991). “NS2 protein of influenza virus is found in purified virus and phosphorylated in infected cells”. In: *Archives of Virology* 116.1-4, pp. 69–80.

Rideout, W M, K Eggan, and Rudolf Jaenisch (Aug. 2001). “Nuclear cloning and epigenetic reprogramming of the genome.” In: *Science (New York, N.Y.)* 293.5532, pp. 1093–8.

Ritter, Joachim B et al. (Jan. 2010). “Metabolic effects of influenza virus infection in cultured animal cells: Intra- and extracellular metabolite profiling.” In: *BMC systems biology* 4, p. 61.

Roeder, Ingo and Ingmar Glauche (Aug. 2006). “Towards an understanding of lineage specification in hematopoietic stem cells: a mathematical model for the interaction of transcription factors GATA-1 and PU.1.” In: *Journal of theoretical biology* 241.4, pp. 852–65.

Saez-Rodriguez, Julio et al. (Aug. 2007). “A logical model provides insights into T cell receptor signaling.” In: *PLoS computational biology* 3.8, e163.

Salminen, Antero and Kai Kaarniranta (July 2010). “Glycolysis links p53 function with NF-kappaB signaling: impact on cancer and aging process.” In: *Journal of cellular physiology* 224.1, pp. 1–6.

Samavarchi-Tehrani, Payman et al. (July 2010). “Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming”. In: *Cell Stem Cell* 7.1, pp. 64–77.

Schaber, Jörg et al. (Oct. 2010). “Biophysical properties of Saccharomyces cerevisiae and their relationship with HOG pathway activation.” In: *European biophysics journal : EBJ* 39.11, pp. 1547–56.

Schaber, Jörg et al. (Jan. 2011). “Automated ensemble modeling with modelMaGe: analyzing feedback mechanisms in the Sho1 branch of the HOG pathway.” In: *PloS one* 6.3. Ed. by Alan Ruttenberg, e14791.

Schaber, Jörg et al. (Nov. 2012). “Modelling reveals novel roles of two parallel signalling pathways and homeostatic feedbacks in yeast”. In: *Molecular Systems Biology* 8.

Schulz, Marvin, Barbara M Bakker, and Edda Klipp (Jan. 2009). “TIde: a software for the systematic scanning of drug targets in kinetic network models.” In: *BMC bioinformatics* 10.1, p. 344.

Schwanhäusser, Björn et al. (May 2011). “Global quantification of mammalian gene expression control.” In: *Nature* 473.7347, pp. 337–42.

Shapira, Sagi D et al. (Dec. 2009). “A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection.” In: *Cell* 139.7, pp. 1255–67.

Shapiro, G I and R M Krug (July 1988). “Influenza virus RNA replication in vitro: synthesis of viral template RNAs and virion RNAs in the absence of an added primer.” In: *J. Virol.* 62.7, pp. 2285–2290.

Shi, Y et al. (2008). “A Combined Chemical and Genetic Approach for the Generation of Induced Pluripotent Stem Cells”. In: *Cell Stem Cell* 2.6, pp. 525–528.

Shmulevich, I. (Feb. 2002). “Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks”. In: *Bioinformatics* 18.2, pp. 261–274.

Sidorenko, Y and U Reichl (Oct. 2004). “Structured model of influenza virus replication in MDCK cells.” In: *Biotechnology and bioengineering* 88.1, pp. 1–14.

Sturm, Marc et al. (Jan. 2008). “OpenMS - an open-source software framework for mass spectrometry.” In: *BMC bioinformatics* 9, p. 163.

Szallasi, Zoltan, Jörg Stelling, and Vipul Periwal (2010). *System Modeling in Cellular Biology: From Concepts to Nuts and Bolts (Bradford Books)*. The MIT Press, p. 464.

Takahashi, Kazutoshi and Shinya Yamanaka (Aug. 2006a). “Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.” In: *Cell* 126.4, pp. 663–76.

— (Aug. 2006b). “Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.” In: *Cell* 126.4, pp. 663–76.

Takahashi, Kazutoshi et al. (Nov. 2007). “Induction of pluripotent stem cells from adult human fibroblasts by defined factors”. In: *Cell* 131.5, pp. 861–872.

Tarca, Adi Laurentiu et al. (Jan. 2009). “A novel signaling pathway impact analysis.” In: *Bioinformatics (Oxford, England)* 25.1, pp. 75–82.

Thomson, J. A. (Nov. 1998). “Embryonic Stem Cell Lines Derived from Human Blastocysts”. In: *Science* 282.5391, pp. 1145–1147.

Thomson, John P et al. (Apr. 2010). “CpG islands influence chromatin structure via the CpG-binding protein Cfp1.” In: *Nature* 464.7291, pp. 1082–6.

Thomson, Ty M et al. (Dec. 2011). “Scaffold number in yeast signaling system sets tradeoff between system output and dynamic range.” In: *Proceedings of the National Academy of Sciences of the United States of America* 108.50, pp. 20265–70.

Tiger, Carl-Fredrik et al. (Apr. 2012). “A framework for mapping, visualisation and automatic model creation of signal-transduction networks”. In: *Molecular Systems Biology* 8.578, pp. 1–20.

Tsuji-Takayama, Kazue et al. (Oct. 2004). “Demethylating agent, 5-azacytidine, reverses differentiation of embryonic stem cells.” In: *Biochemical and biophysical research communications* 323.1, pp. 86–90.

Twardziok, S, H Siebert, and A Heyl (2010). “Stochasticity in reactions: a probabilistic Boolean modeling approach”. In: *... of the 8th International Conference on ...*.

Varga, Zsuzsanna T et al. (June 2011). “The influenza virus protein PB1-F2 inhibits the induction of type I interferon at the level of the MAVS adaptor protein.” In: *PLoS pathogens* 7.6, e1002067.

Venter, J C et al. (Feb. 2001). "The sequence of the human genome." In: *Science (New York, N.Y.)* 291.5507, pp. 1304–51.

Vierbuchen, Thomas et al. (Feb. 2010). "Direct conversion of fibroblasts to functional neurons by defined factors". In: *Nature* 463.7284, pp. 1035–1041.

Waddington, Conrad (1940). *Organisers & genes*. Cambridge: Cambridge University Press.

— (1957). *The strategy of the genes*. London: George Allen & Unwin.

Wang, Jin et al. (July 2010). "The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation." In: *Biophysical journal* 99.1, pp. 29–39.

Wang, Ying and James Adjaye (Dec. 2010). "A Cyclic AMP Analog, 8-Br-cAMP, Enhances the Induction of Pluripotency in Human Fibroblast Cells." In: *Stem cell reviews*.

Warner, Jonathan R and Kerri B McIntosh (Apr. 2009). "How common are extraribosomal functions of ribosomal proteins?" In: *Molecular cell* 34.1, pp. 3–11.

Watanabe, Tokiko, Shinji Watanabe, and Yoshihiro Kawaoka (June 2010). "Cellular networks involved in the influenza virus life cycle." In: *Cell host & microbe* 7.6, pp. 427–39.

Watson, J D and F H Crick (Apr. 1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." In: *Nature* 171.4356, pp. 737–8.

Wernig, Marius et al. (July 2007). "In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state." In: *Nature* 448.7151, pp. 318–24.

Wittmann, Dominik M et al. (Jan. 2009). "Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling." In: *BMC systems biology* 3.1, p. 98.

Wu, Z et al. (Feb. 1999). "Cross-regulation of C/EBP alpha and PPAR gamma controls the transcriptional pathway of adipogenesis and insulin sensitivity." In: *Molecular cell* 3.2, pp. 151–8.

Zhou, Hongyan et al. (May 2009). "Generation of induced pluripotent stem cells using recombinant proteins." In: *Cell stem cell* 4.5, pp. 381–4.

Zhou, Qing, Hiram Chipperfield, and DA Melton (Oct. 2007). "A gene regulatory network in mouse embryonic stem cells". In: *Proceedings of the National Academy of Sciences* 104.42, pp. 16438–16443.

A

Software

A.1 Implementation

All software written for this work was implemented in *Python* and *R*. Grphical user interfaces were implemented using *Javascript* with *jQuery* and visualizations with *Biographer* and *D3.js*. *ModelMaGe* is available at `http://www.modelmage.org` and Boolean *rxncon* as part of the *rxncon*framework on `http://www.rxncon.org`.

A.2 *ModelMaGe*

Ordinary differential equation system for the master model mentioned used in section 3.2:

$$
\begin{aligned}
\frac{\mathrm{d}Sho1}{\mathrm{d}t} &= -k1_{v1} \cdot Sho1 \cdot Signal \cdot OuterOsmo + \frac{V_{v2} \cdot Sho1a}{Km_{v2} + Sho1a} + \frac{V_{v4} \cdot Sho1i}{Km_{v4} + Sho1i} \\
\frac{\mathrm{d}Sho1a}{\mathrm{d}t} &= k1_{v1} \cdot Sho1 \cdot Signal \cdot OuterOsmo - \frac{V_{v2} \cdot Sho1a}{Km_{v2} + Sho1a} - \frac{vmax_{v3}P - Hog1 \cdot Sho1a}{Km_{v3} + Sho1a} \\
\frac{\mathrm{d}Sho1i}{\mathrm{d}t} &= \frac{vmax_{v3}P - Hog1 \cdot Sho1a}{Km_{v3} + Sho1a} - \frac{V_{v4} \cdot Sho1i}{Km_{v4} + Sho1i} \\
\frac{\mathrm{d}Ste11}{\mathrm{d}t} &= -\frac{vmax_{v5} \cdot Sho1a \cdot Ste11}{Km_{v5} + Ste11} + \frac{V_{v6} \cdot PSte11}{Km_{v6} + P - Ste11} \\
\frac{\mathrm{d}PSte11}{\mathrm{d}t} &= \frac{vmax_{v5} \cdot Sho1a \cdot Ste11}{Km_{v5} + Ste11} - \frac{V_{v6} \cdot PSte11}{Km_{v6} + PSte11} \\
\frac{\mathrm{d}Pbs2}{\mathrm{d}t} &= -\frac{vmax_{v7} \cdot PSte11 \cdot Pbs2}{Km_{v7} + Pbs2} + \frac{V_{v8} \cdot PPbs2}{Km_{v8} + PPbs2} \\
\frac{\mathrm{d}PPbs2}{\mathrm{d}t} &= \frac{vmax_{v7} \cdot PSte11 \cdot Pbs2}{Km_{v7} + Pbs2} - \frac{V_{v8} \cdot PPbs2}{Km_{v8} + PPbs2} \\
\frac{\mathrm{d}Hog1}{\mathrm{d}t} &= -k1_{v9} \cdot Hog1 \cdot PPbs2 \cdot Sho1a \cdot Signal + \frac{V_{v10} \cdot PHog1}{Km_{v10} + PHog1} \\
\frac{\mathrm{d}PHog1}{\mathrm{d}t} &= k1_{v9} \cdot Hog1 \cdot PPbs2 \cdot Sho1a \cdot Signal - \frac{V_{v10} \cdot PHog1}{Km_{v10} + PHog1} \\
\frac{\mathrm{d}signal}{\mathrm{d}t} &= \begin{cases} osmo_{out} - osmo_{in}, osmo_{out} > osmo_{in} \\ 0, else \end{cases} \\
\frac{\mathrm{d}osmo_{in}}{\mathrm{d}t} &= k_{v11} \cdot PHog1 + v_{v12} - \frac{k_{v13} \cdot osmo_{in}}{1 + \left(\frac{signal}{ki_{v13}}\right)^{h}}
\end{aligned}
$$

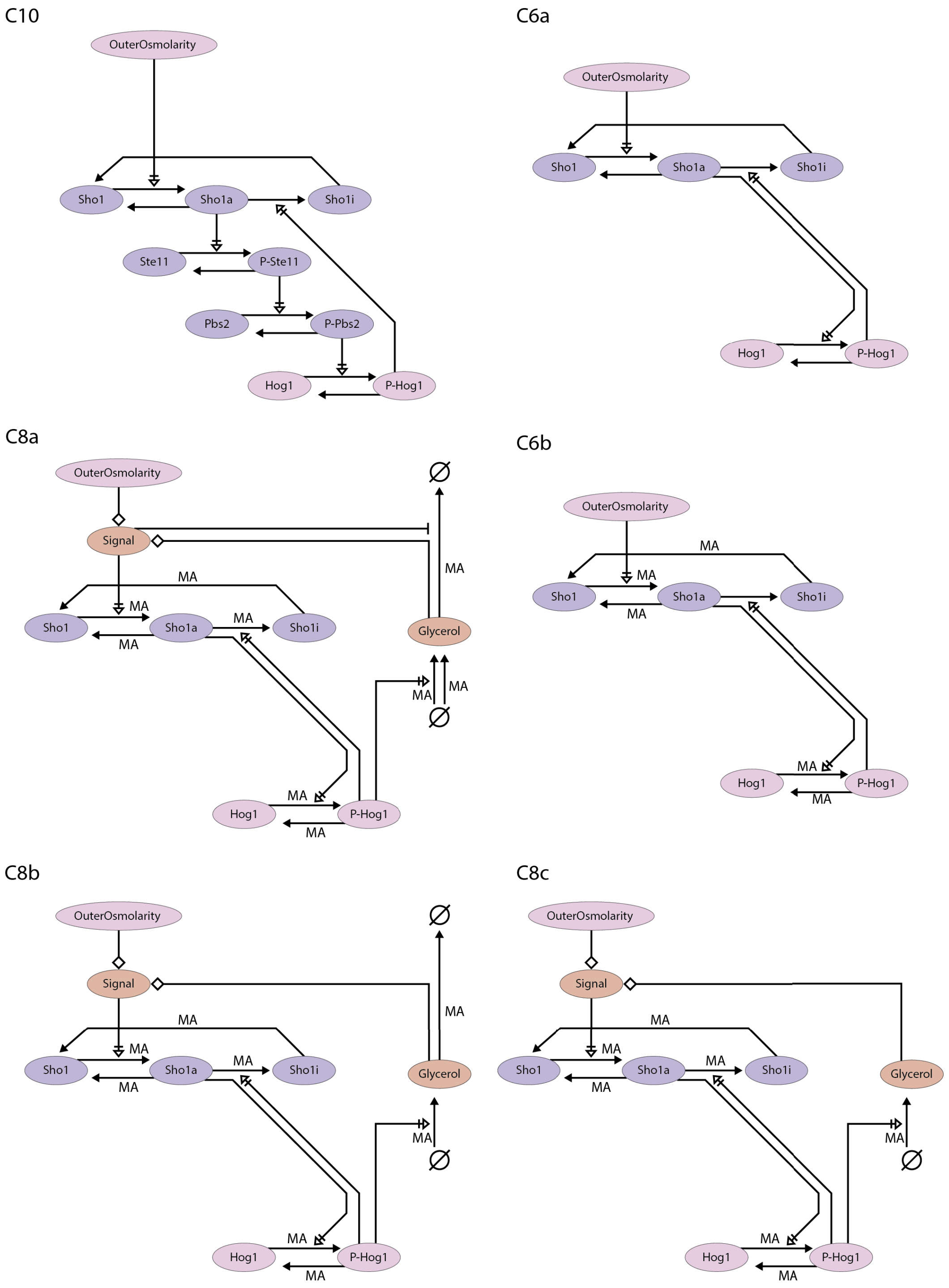


Figure A.2.1: Structure of candidate models C10, C6a, C6b, C8a, C8b, and C8c.

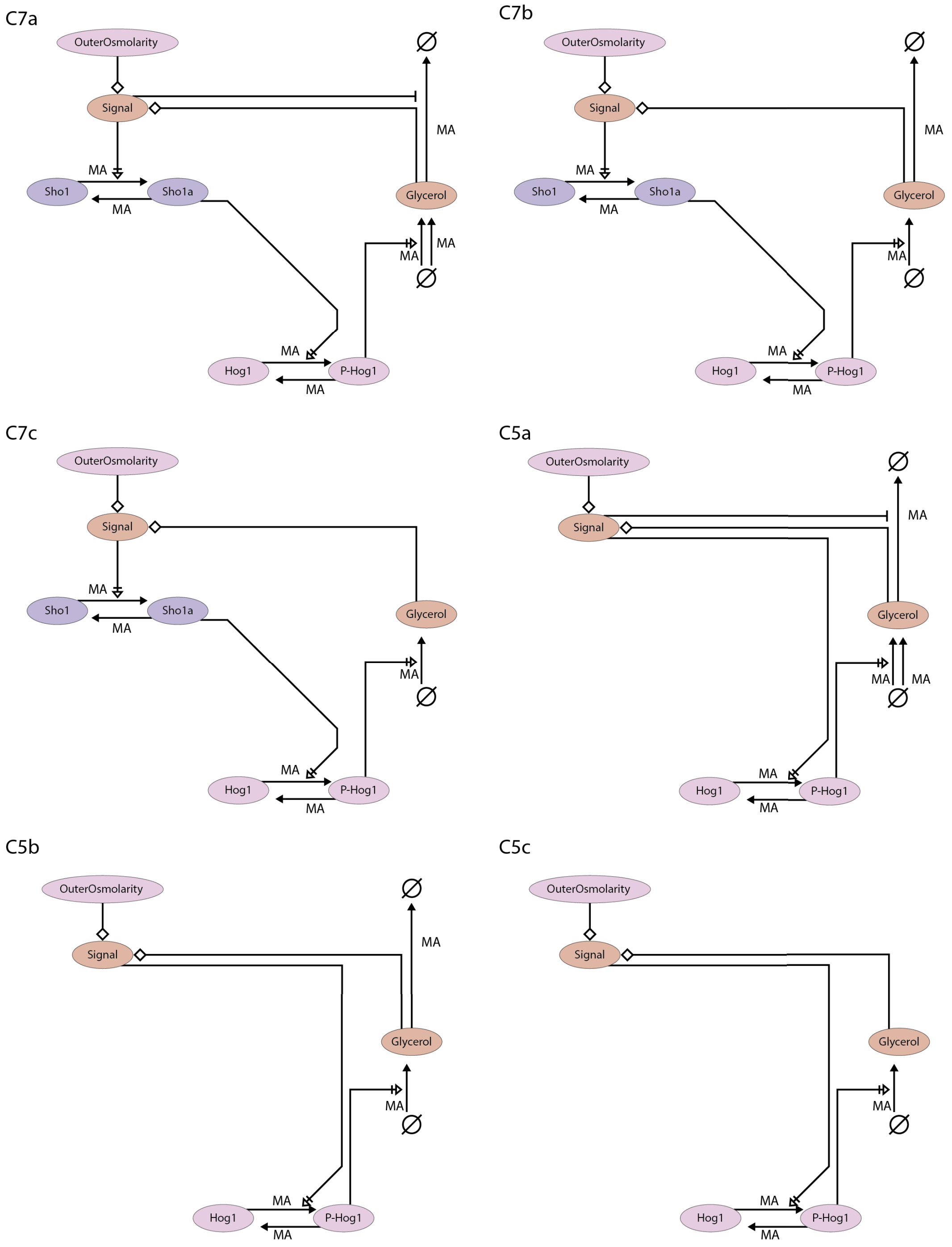
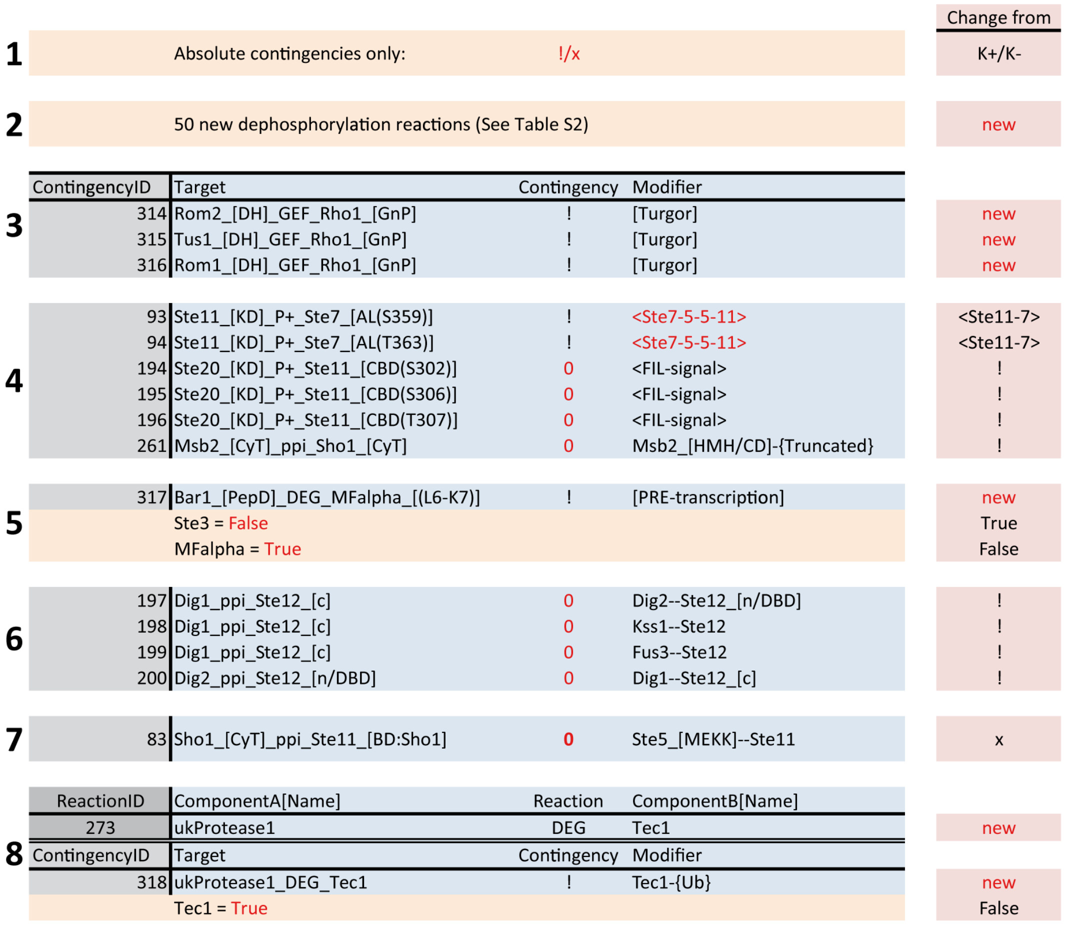
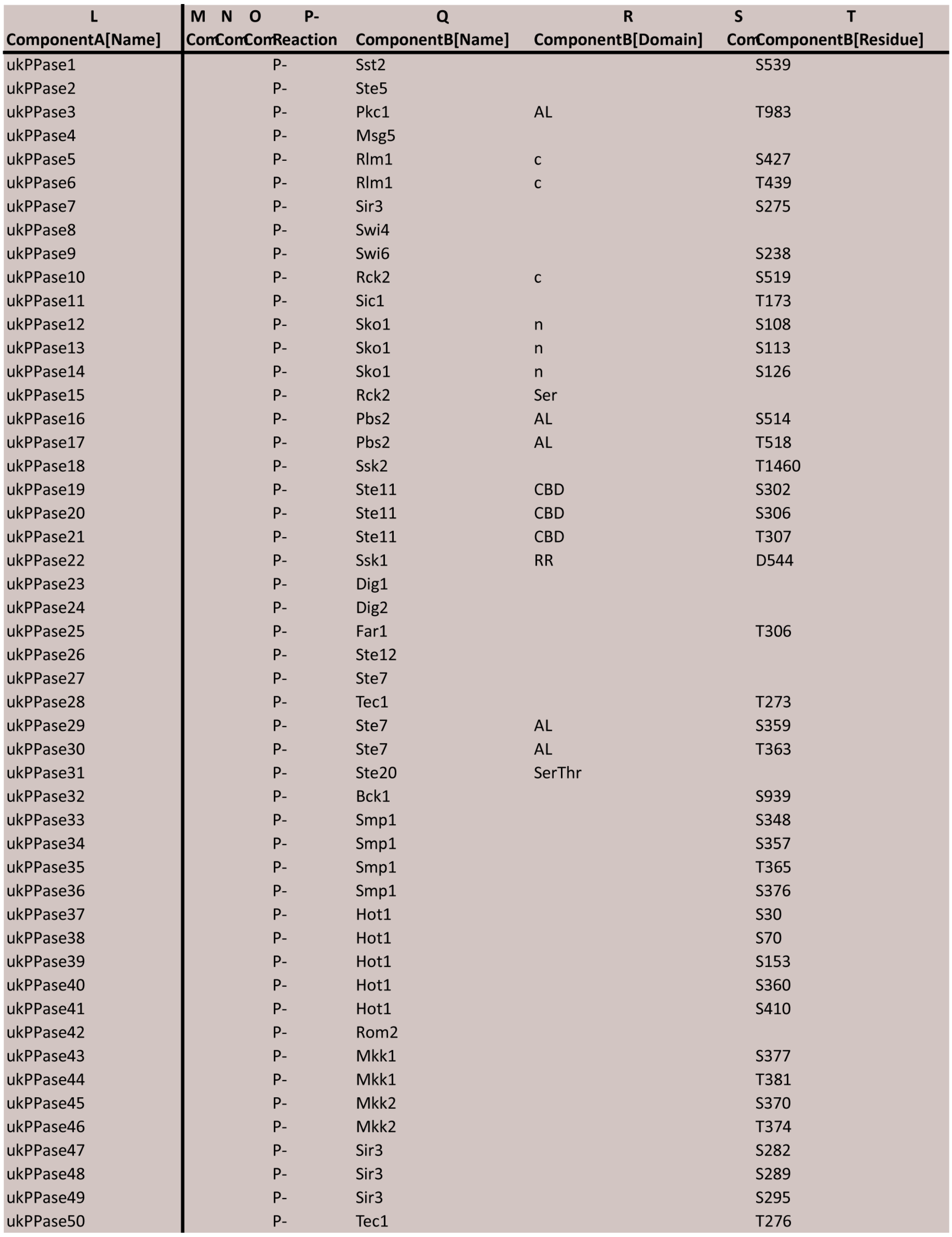


Figure A.2.2: Structure of candidate models C7a, C7b, C7c, C5a, C5b, and C5c.



					Change from
1		Absolute contingencies only:	!/x		K+/K-
2		50 new dephosphorylation reactions (See Table S2)			new
	ContingencyID	Target	Contingency	Modifier	
3	314	Rom2_[DH]_GEF_Rho1_[GnP]	!	[Turgor]	new
	315	Tus1_[DH]_GEF_Rho1_[GnP]	!	[Turgor]	new
	316	Rom1_[DH]_GEF_Rho1_[GnP]	!	[Turgor]	new
4	93	Ste11_[KD]_P+_Ste7_[AL(S359)]	!	<Ste7-5-5-11>	<Ste11-7>
	94	Ste11_[KD]_P+_Ste7_[AL(T363)]	!	<Ste7-5-5-11>	<Ste11-7>
	194	Ste20_[KD]_P+_Ste11_[CBD(S302)]	0	<FIL-signal>	!
	195	Ste20_[KD]_P+_Ste11_[CBD(S306)]	0	<FIL-signal>	!
	196	Ste20_[KD]_P+_Ste11_[CBD(T307)]	0	<FIL-signal>	!
	261	Msb2_[CyT]_ppi_Sho1_[CyT]	0	Msb2_[HMH/CD]-{Truncated}	!
5	317	Bar1_[PepD]_DEG_MFalpha_[(L6-K7)]	!	[PRE-transcription]	new
		Ste3 = False			True
		MFalpha = True			False
6	197	Dig1_ppi_Ste12_[c]	0	Dig2--Ste12_[n/DBD]	!
	198	Dig1_ppi_Ste12_[c]	0	Kss1--Ste12	!
	199	Dig1_ppi_Ste12_[c]	0	Fus3--Ste12	!
	200	Dig2_ppi_Ste12_[n/DBD]	0	Dig1--Ste12_[c]	!
7	83	Sho1_[CyT]_ppi_Ste11_[BD:Sho1]	0	Ste5_[MEKK]--Ste11	x
	ReactionID	ComponentA[Name]	Reaction	ComponentB[Name]	
8	273	ukProtease1	DEG	Tec1	new
	ContingencyID	Target	Contingency	Modifier	
	318	ukProtease1_DEG_Tec1	!	Tec1-{Ub}	new
		Tec1 = True			False

Table A.2.1: The necessary changes in the MAPK model for a signal transduction in all pathways. Changes are given in the *rxncon* format.



L	M	N	O	P-	Q	R	S	T
ComponentA[Name]	Com	Com	Com	Reaction	ComponentB[Name]	ComponentB[Domain]	Com	ComponentB[Residue]
ukPPase1				P-	Sst2			S539
ukPPase2				P-	Ste5			
ukPPase3				P-	Pkc1	AL		T983
ukPPase4				P-	Msg5			
ukPPase5				P-	Rlm1	c		S427
ukPPase6				P-	Rlm1	c		T439
ukPPase7				P-	Sir3			S275
ukPPase8				P-	Swi4			
ukPPase9				P-	Swi6			S238
ukPPase10				P-	Rck2	c		S519
ukPPase11				P-	Sic1			T173
ukPPase12				P-	Sko1	n		S108
ukPPase13				P-	Sko1	n		S113
ukPPase14				P-	Sko1	n		S126
ukPPase15				P-	Rck2	Ser		
ukPPase16				P-	Pbs2	AL		S514
ukPPase17				P-	Pbs2	AL		T518
ukPPase18				P-	Ssk2			T1460
ukPPase19				P-	Ste11	CBD		S302
ukPPase20				P-	Ste11	CBD		S306
ukPPase21				P-	Ste11	CBD		T307
ukPPase22				P-	Ssk1	RR		D544
ukPPase23				P-	Dig1			
ukPPase24				P-	Dig2			
ukPPase25				P-	Far1			T306
ukPPase26				P-	Ste12			
ukPPase27				P-	Ste7			
ukPPase28				P-	Tec1			T273
ukPPase29				P-	Ste7	AL		S359
ukPPase30				P-	Ste7	AL		T363
ukPPase31				P-	Ste20	SerThr		
ukPPase32				P-	Bck1			S939
ukPPase33				P-	Smp1			S348
ukPPase34				P-	Smp1			S357
ukPPase35				P-	Smp1			T365
ukPPase36				P-	Smp1			S376
ukPPase37				P-	Hot1			S30
ukPPase38				P-	Hot1			S70
ukPPase39				P-	Hot1			S153
ukPPase40				P-	Hot1			S360
ukPPase41				P-	Hot1			S410
ukPPase42				P-	Rom2			
ukPPase43				P-	Mkk1			S377
ukPPase44				P-	Mkk1			T381
ukPPase45				P-	Mkk2			S370
ukPPase46				P-	Mkk2			T374
ukPPase47				P-	Sir3			S282
ukPPase48				P-	Sir3			S289
ukPPase49				P-	Sir3			S295
ukPPase50				P-	Tec1			T276

Table A.2.2: List of reactions where a phosphatase needed to be added to enable the pathways to be responsive to signaling.

Table A.2.3: The small HOG model in BooxXklean terms as converted by my export tool.

$$\begin{aligned}
\text{Sln1-\{P\}} &= \text{Sln1\_AP\_Sln1} \lor \lnot\text{Sln1\_PT\_Ypd1} \land \text{Sln1-\{P\}} \\
\text{Ssk1-\{P\}} &= \text{Ypd1\_PT\_Ssk1} \lor \lnot\text{PPase\_P-\_Ssk1} \land \text{Ssk1-\{P\}} \\
\text{Hot1-\{P\}} &= \text{Hog1\_P+\_Hot1} \lor \lnot\text{PPase\_P-\_Hot1} \land \text{Hot1-\{P\}} \\
\text{Hog1-\{P\}} &= \text{Pbs2\_P+\_Hog1} \lor \lnot\text{PPase\_P-\_Hog1} \land \text{Hog1-\{P\}} \\
\text{Ypd1-\{P\}} &= \text{Sln1\_PT\_Ypd1} \lor \lnot\text{Ypd1\_PT\_Ssk1} \land \text{Ypd1-\{P\}} \\
\text{Pbs2-\{P\}} &= \text{Ssk2\_P+\_Pbs2} \lor \lnot\text{PPase\_P-\_Pbs2} \land \text{Pbs2-\{P\}} \\
\text{Ssk1--Ssk2} &= \text{Ssk1\_ppi\_Ssk2}
\end{aligned}$$

$$\begin{aligned}
\text{Hog1\_P+\_Hot1} &= \mathit{Hog1} \land \mathit{Hot1} \land \text{Hog1-\{P\}} \\
\text{Ssk1\_ppi\_Ssk2} &= \mathit{Ssk1} \land \mathit{Ssk2} \land \lnot\text{Ssk1-\{P\}} \\
\text{PPase\_P-\_Hot1} &= \mathit{PPase} \land \mathit{Hot1} \land \text{Hot1-\{P\}} \\
\text{PPase\_P-\_Pbs2} &= \mathit{PPase} \land \mathit{Pbs2} \land \text{Pbs2-\{P\}} \\
\text{PPase\_P-\_Ssk1} &= \mathit{PPase} \land \mathit{Ssk1} \land \text{Ssk1-\{P\}} \\
\text{PPase\_P-\_Hog1} &= \mathit{PPase} \land \mathit{Hog1} \land \text{Hog1-\{P\}} \\
\text{Ypd1\_PT\_Ssk1} &= \mathit{Ypd1} \land \mathit{Ssk1} \land \text{Ypd1-\{P\}} \\
\text{Pbs2\_P+\_Hog1} &= \mathit{Pbs2} \land \mathit{Hog1} \land \text{Pbs2-\{P\}} \\
\text{Sln1\_PT\_Ypd1} &= \mathit{Sln1} \land \mathit{Ypd1} \land \text{Sln1-\{P\}} \\
\text{Ssk2\_P+\_Pbs2} &= \mathit{Ssk2} \land \mathit{Pbs2} \land \text{Ssk1--Ssk2} \\
\text{Sln1\_AP\_Sln1} &= \mathit{Sln1} \land \mathit{Sln1} \land \text{Hot1-\{P\}}
\end{aligned}$$

```

Hot1 = True
Sln1 = True
Ssk2 = True
Ypd1 = True
PPase = True
Hog1 =True
Pbs2 = True
Ssk1 = True

Sln1-_P_ = False
Ssk1-_P_ = False
Hot1-_P_ = False
Hog1-_P_ = False
Ypd1-_P_ = False
Pbs2-_P_ = False
Ssk1--Ssk2 = False

Hog1_P+_Hot1 = False
Ssk1_ppi_Ssk2 = False
PPase_P-_Hot1 = False
PPase_P-_Pbs2 = False
PPase_P-_Ssk1 = False
PPase_P-_Hog1 = False
Ypd1_PT_Ssk1 = False
Pbs2_P+_Hog1 = False
Sln1_PT_Ypd1 = False
Ssk2_P+_Pbs2 = False
Sln1_AP_Sln1 = False

Sln1-_P_ *= (Sln1_AP_Sln1) or (not (Sln1_PT_Ypd1) and Sln1-_P_)
Ssk1-_P_ *= (Ypd1_PT_Ssk1) or (not (PPase_P-_Ssk1) and Ssk1-_P_)
Hot1-_P_ *= (Hog1_P+_Hot1) or (not (PPase_P-_Hot1) and Hot1-_P_)
Hog1-_P_ *= (Pbs2_P+_Hog1) or (not (PPase_P-_Hog1) and Hog1-_P_)
Ypd1-_P_ *= (Sln1_PT_Ypd1) or (not (Ypd1_PT_Ssk1) and Ypd1-_P_)
Pbs2-_P_ *= (Ssk2_P+_Pbs2) or (not (PPase_P-_Pbs2) and Pbs2-_P_)
Ssk1--Ssk2 *= (Ssk1_ppi_Ssk2)

Hog1_P+_Hot1 *= Hog1 and Hot1 and Hog1-_P_
Ssk1_ppi_Ssk2 *= Ssk1 and Ssk2 and not (Ssk1-_P_)
PPase_P-_Hot1 *= PPase and Hot1 and Hot1-_P_
PPase_P-_Pbs2 *= PPase and Pbs2 and Pbs2-_P_
PPase_P-_Ssk1 *= PPase and Ssk1 and Ssk1-_P_
PPase_P-_Hog1 *= PPase and Hog1 and Hog1-_P_
Ypd1_PT_Ssk1 *= Ypd1 and Ssk1 and Ypd1-_P_
Pbs2_P+_Hog1 *= Pbs2 and Hog1 and Pbs2-_P_
Sln1_PT_Ypd1 *= Sln1 and Ypd1 and Sln1-_P_
Ssk2_P+_Pbs2 *= Ssk2 and Pbs2 and Ssk1--Ssk2
Sln1_AP_Sln1 *= Sln1 and Sln1 and Hot1-_P_
```

Listing A.1: The small example HOG model shown in Figure 3.3.1 formulated in the *BooleanNet* format as exported by our software.

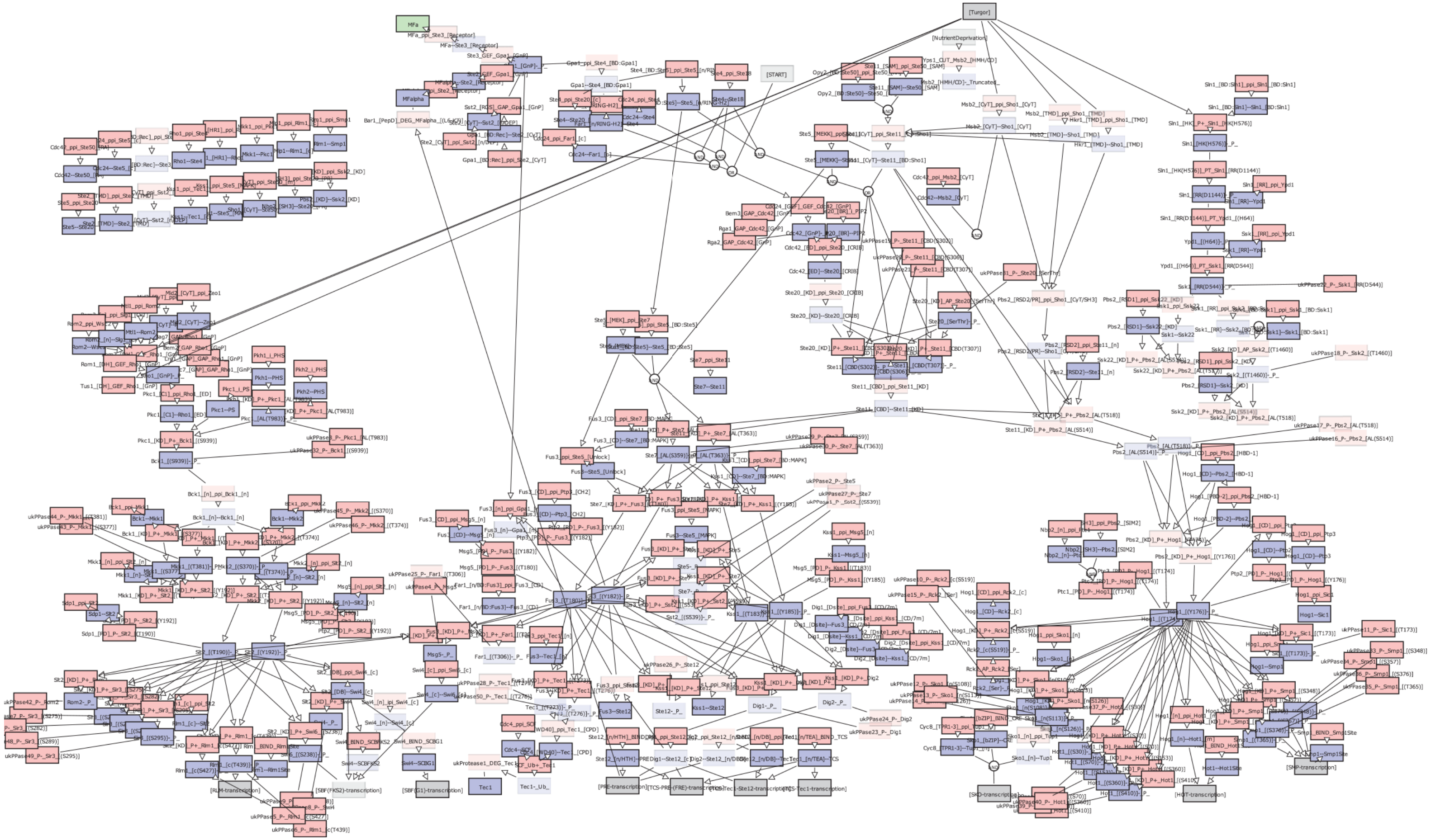


Figure A.2.3: The complete network structure of the MAPK model. This visualization was done by the *rxncon* extension I implemented.

B
Pluripotency

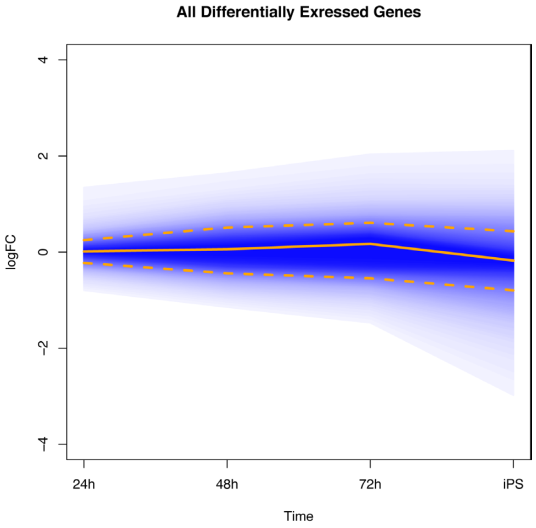


Figure B.0.1: General changes in the measured genes over time. The stronger changes in the iPS cells are clearly visible compared to the first time points.

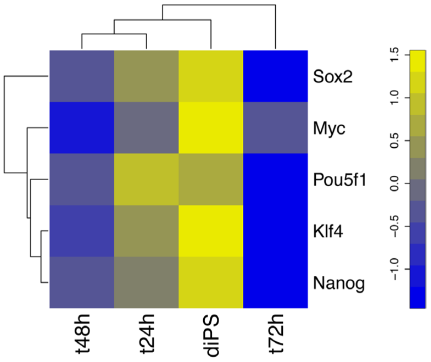


Figure B.0.2: Heatmap of the reprogramming factors TFA.

Table B.0.1: Spia 24h

Name	ID	pSize	NDE	pNDE	tA	pPERT	pG	pGFdr	pGFWER	Status
Herpes simplex infection	5168	120	16	8,13E+06	145.167.277.855.853	0,017	2,96E+06	3,05E+08	3,05E+08	Activated
Influenza A	5164	107	14	6,76E+07	135.155.582.823.667	0,008	1,08E+07	5,15E+08	1,12E+09	Activated
Measles	5162	79	12	8,47E+07	118.717.922.209.127	0,009	1,50E+07	5,15E+08	1,55E+09	Activated
Cytosolic DNA-sensing pathway	4623	32	7	1,47E+09	304.359.834.469.842	0,124	2,59E+09	0,000666879	0,002667516	Activated
Neuroactive ligand-receptor interaction	4080	84	9	0,000327395	0	1	0,002954523	0,046780151	0,304315849	Inhibited
RIG-I-like receptor signaling pathway	4622	36	5	0,002268388	342.736.012.907.208	0,146	0,002984912	0,046780151	0,307445949	Activated
African trypanosomiasis	5143	17	3	0,00911654	-153.274.351.429.464	0,039	0,003179234	0,046780151	0,327461054	Inhibited
Pertussis	5133	47	5	0,00731608	510.662.275.377.167	0,156	0,008874335	0,114257069	0,914056549	Activated
Hepatitis C	5160	89	8	0,00220803	0,98184935	0,712	0,011720658	0,134136425	1	Activated
Toll-like receptor signaling pathway	4620	54	5	0,013022155	604.870.281.171.912	0,204	0,018411642	0,189639917	1	Activated
Jak-STAT signaling pathway	4630	80	6	0,017961819	124.344.033.998.375	0,228	0,026610884	0,240395176	1	Activated
Intestinal immune network for IgA production	4672	19	3	0,012501974	0,676532749	0,348	0,028007205	0,240395176	1	Activated
Calcium signaling pathway	4020	96	7	0,012651252	245.300.903.042.485	0,432	0,033936099	0,252466499	1	Activated
Cytokine-cytokine receptor interaction	4060	108	8	0,007211367	0,726520418	0,768	0,034315835	0,252466499	1	Activated
Legionellosis	5134	37	4	0,015241732	174.548.878.211.752	0,466	0,042241487	0,290058209	1	Activated
Pathways in cancer	5200	236	13	0,008985098	0,929140242	0,905	0,047260423	0,304238976	1	Activated
Leishmaniasis	5140	41	3	0,090822764	387.215.086.521.367	0,111	0,056425896	0,327888398	1	Activated
Complement and coagulation cascades	4610	26	2	0,146743616	-75.124.575.702.227	0,07	0,057300885	0,327888398	1	Inhibited
Bladder cancer	5219	35	4	0,012582336	-0,498420338	0,918	0,063077963	0,34194896	1	Inhibited
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	5412	45	2	0,329069757	113.460.265.733.333	0,039	0,068733289	0,353976441	1	Activated

Table B.0.2: Spia 48h

Name	ID	pSize	NDE	pNDE	tA	pPERT	pG	pGFdr	pGFWER	Status
Measles	5162	79	21	0,000179205	113.591.673.438.573	0,009	2,31E+09	0,002587393	0,002959901	
Influenza A	5164	107	27	6,13E+09	907.834.083.439.528	0,048	4,04E+09	0,002587393	0,005174786	
Focal adhesion	4510	149	31	0,000777785	-170.100.748.977.737	0,046	0,00040208	0,017155431	0,051466293	
Pertussis	5133	47	15	0,000168344	275.214.352.262.328	0,513	0,000894435	0,028621923	0,114487693	
Legionellosis	5134	37	10	0,0076864	705.153.355.813.764	0,055	0,003706996	0,079446506	0,474495471	
Colorectal cancer	5210	46	10	0,034644496	-524.547.152.598.461	0,013	0,003920735	0,079446506	0,501854027	
Herpes simplex infection	5168	120	21	0,034019539	14.208.172.734.239	0,015	0,004378583	0,079446506	0,560458592	
Cytosolic DNA-sensing pathway	4623	32	9	0,008528367	267.212.320.759.415	0,069	0,004965407	0,079446506	0,635572051	
Jak-STAT signaling pathway	4630	80	18	0,003913694	221.546.907.094.042	0,201	0,006409427	0,091156297	0,820406673	
Malaria	5144	28	8	0,011671612	0	NA	0,011671612	0,149108152	1	
African trypanosomiasis	5143	17	5	0,038730495	-218.829.816.224.671	0,045	0,012813982	0,149108152	1	
Calcium signaling pathway	4020	96	19	0,012916803	508.531.379.245.982	0,171	0,015716125	0,167638662	1	
Toll-like receptor signaling pathway	4620	54	13	0,007358513	45.373.304.046.586	0,429	0,021334287	0,200125087	1	
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	5412	45	10	0,030123473	0,980504018	0,108	0,021888681	0,200125087	1	
Neuroactive ligand-receptor interaction	4080	84	17	0,014612806	0,694304329	0,262	0,025135489	0,207087721	1	
Regulation of autophagy	4140	14	1	0,821332327	-410.069.506.910.884	0,005	0,026673362	0,207087721	1	
Pathways in cancer	5200	236	31	0,250487878	-204.021.777.480.654	0,017	0,027503838	0,207087721	1	
Epithelial cell signaling in Helicobacter pylori infection	5120	53	11	0,037534467	-325.319.763.448.358	0,174	0,039389722	0,280104688	1	
Cytokine-cytokine receptor interaction	4060	108	21	0,011323635	-0,809966561	0,772	0,050174997	0,331329755	1	
Leishmaniasis	5140	41	9	0,041460226	305.797.833.369.821	0,219	0,051770274	0,331329755	1	

Table B.0.3: Spia 72h

Name	ID	pSize	NDE	pNDE	tA	pPERT	pG	pGFdr	pGFWER	Status
Focal adhesion	4510	149	46	2,19E+08	-16.477.384.029.386	0,142	4,26E+09	0,005451967	0,005451967	Inhibited
Measles	5162	79	22	0,011102071	646.720.531.446.174	0,095	0,008284118	0,387644997	1	Activated
Pathways in cancer	5200	236	48	0,104701932	-294.801.158.777.347	0,019	0,014362919	0,387644997	1	Inhibited
Complement and coagulation cascades	4610	26	11	0,002184383	-0,066519579	0,995	0,015499893	0,387644997	1	Inhibited
Pathogenic Escherichia coli infection	5130	43	15	0,003646495	-261.847.168.375.479	0,662	0,016961776	0,387644997	1	Inhibited
Cytosolic DNA-sensing pathway	4623	32	12	0,004613709	0,645253988	0,567	0,018170859	0,387644997	1	Activated
Calcium signaling pathway	4020	96	27	0,004555743	0,909690557	0,84	0,025125857	0,414713276	1	Activated
Regulation of actin cytoskeleton	4810	147	31	0,117568337	-16.768.134.267.283	0,034	0,026071064	0,414713276	1	Inhibited
GnRH signaling pathway	4912	63	15	0,106907599	-193.543.195.665.804	0,045	0,030485743	0,414713276	1	Inhibited
Colorectal cancer	5210	46	12	0,080811023	-431.225.601.576.103	0,064	0,032399475	0,414713276	1	Inhibited
Gap junction	4540	58	17	0,014369968	-521.573.258.567.982	0,428	0,037463277	0,435936315	1	Inhibited
Jak-STAT signaling pathway	4630	80	21	0,024892828	205.252.927.482.209	0,306	0,044768949	0,460007984	1	Activated
Viral myocarditis	5416	41	8	0,400247359	633.512.484.974.459	0,024	0,054229105	0,460007984	1	Activated
Prostate cancer	5215	60	16	0,040552094	-854.950.523.907.358	0,243	0,05537901	0,460007984	1	Inhibited
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	5412	45	11	0,13272529	128.873.636.303.763	0,075	0,055841587	0,460007984	1	Activated
Legionellosis	5134	37	7	0,448512265	737.047.156.704.618	0,023	0,057500998	0,460007984	1	Activated
Vibrio cholerae infection	5110	44	11	0,117429843	-164.680.205.198.217	0,099	0,063412164	0,477456292	1	Inhibited
African trypanosomiasis	5143	17	3	0,574040432	-226.239.462.610.882	0,025	0,075255766	0,512103333	1	Inhibited
Melanoma	5218	46	10	0,250521582	-146.463.999.172.906	0,058	0,076015339	0,512103333	1	Inhibited
Epithelial cell signaling in Helicobacter pylori infection	5120	53	14	0,057060579	-340.023.615.473.008	0,289	0,084183551	0,536040459	1	Inhibited

Table B.0.4: Enriched GO terms cluster 1

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0051496	0.000	69.384	0	3	11	positive regulation of stress fiber assembly
GO:0071356	0.000	Inf	0	2	2	cellular response to tumor necrosis factor
GO:0048646	0.000	6.133	2	9	312	anatomical structure formation involved in morphogenesis
GO:0032231	0.000	42.671	0	3	16	regulation of actin filament bundle assembly
GO:0023046	0.000	3.169	9	20	1564	signaling process
GO:0051241	0.000	10.426	1	5	97	negative regulation of multicellular organismal process
GO:0022610	0.000	4.396	3	10	482	biological adhesion
GO:0031589	0.000	9.120	1	5	110	cell-substrate adhesion
GO:0032970	0.000	13.206	0	4	61	regulation of actin filament-based process
GO:0033144	0.000	90.500	0	2	6	negative regulation of steroid hormone receptor signaling pathway
GO:0051130	0.001	7.936	1	5	131	positive regulation of cellular component organization
GO:0051495	0.001	18.270	0	3	34	positive regulation of cytoskeleton organization
GO:0007165	0.001	2.849	8	17	1369	signal transduction
GO:0014020	0.001	17.295	0	3	35	primary neural tube formation
GO:0007015	0.002	8.721	1	4	90	actin filament organization
GO:0072175	0.002	13.822	0	3	43	epithelial tube formation
GO:0030336	0.002	12.559	0	3	47	negative regulation of cell migration
GO:0006936	0.003	7.487	1	4	104	muscle contraction
GO:0048731	0.003	2.487	9	17	1529	system development
GO:0040007	0.004	3.947	2	7	353	growth
GO:0065008	0.004	2.675	6	13	1032	regulation of biological quality
GO:0032501	0.004	2.401	12	20	2233	multicellular organismal process
GO:0021915	0.004	10.034	0	3	58	neural tube development
GO:0042246	0.005	22.592	0	2	18	tissue regeneration
GO:0016331	0.005	9.679	0	3	60	morphogenesis of embryonic epithelium
GO:0048514	0.005	5.014	1	5	194	blood vessel morphogenesis
GO:0006979	0.005	6.277	1	4	123	response to oxidative stress
GO:0048729	0.005	4.907	1	5	198	tissue morphogenesis
GO:0045765	0.005	9.349	0	3	62	regulation of angiogenesis
GO:0032319	0.006	20.077	0	2	20	regulation of Rho GTPase activity
GO:0002396	0.006	Inf	0	1	1	MHC protein complex assembly
GO:0002502	0.006	Inf	0	1	1	peptide antigen assembly with MHC class I protein complex
GO:0003065	0.006	Inf	0	1	1	positive regulation of heart rate by epinephrine
GO:0006701	0.006	Inf	0	1	1	progesterone biosynthetic process
GO:0018125	0.006	Inf	0	1	1	peptidyl-cysteine methylation
GO:0030836	0.006	Inf	0	1	1	positive regulation of actin filament depolymerization
GO:0035491	0.006	Inf	0	1	1	positive regulation of leukotriene production involved in inflammatory response
GO:0045578	0.006	Inf	0	1	1	negative regulation of B cell differentiation
GO:0060316	0.006	Inf	0	1	1	positive regulation of ryanodine-sensitive calcium-release channel activity
GO:0061044	0.006	Inf	0	1	1	negative regulation of vascular wound healing
GO:0007264	0.006	4.009	2	6	293	small GTPase mediated signal transduction
GO:0016192	0.007	3.179	3	8	502	vesicle-mediated transport
GO:0043086	0.009	4.331	1	5	223	negative regulation of catalytic activity
GO:0048870	0.009	3.652	2	6	320	cell motility
GO:0034614	0.010	15.046	0	2	26	cellular response to reactive oxygen species
GO:0045807	0.010	15.046	0	2	26	positive regulation of endocytosis
GO:0034097	0.010	7.340	0	3	78	response to cytokine stimulus

Table B.0.5: Enriched GO terms cluster 2

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0051301	0.000	8.119	1	9	274	cell division
GO:0007010	0.000	7.095	2	10	354	cytoskeleton organization
GO:0060236	0.000	135.850	0	2	5	regulation of mitotic spindle organization
GO:0022402	0.001	4.399	2	9	485	cell cycle process
GO:0046579	0.001	58.193	0	2	9	positive regulation of Ras protein signal transduction
GO:0007015	0.001	10.263	0	4	89	actin filament organization
GO:0007051	0.002	14.857	0	3	45	spindle organization
GO:0030029	0.002	6.615	1	5	177	actin filament-based process
GO:0007017	0.003	5.853	1	5	189	microtubule-based process
GO:0030097	0.003	5.853	1	5	189	hemopoiesis
GO:0042060	0.003	7.358	1	4	119	wound healing
GO:0032570	0.003	27.130	0	2	17	response to progesterone stimulus
GO:0046777	0.003	11.124	0	3	59	protein amino acid autophosphorylation
GO:0007126	0.004	10.554	0	3	62	meiosis
GO:0051321	0.004	10.554	0	3	62	meiotic cell cycle
GO:0030048	0.004	23.932	0	2	19	actin filament-based movement
GO:0007067	0.005	5.087	1	5	216	mitosis
GO:0031247	0.005	Inf	0	1	1	actin rod assembly
GO:0051545	0.005	Inf	0	1	1	negative regulation of elastin biosynthetic process
GO:0060209	0.005	Inf	0	1	1	estrus
GO:0070926	0.005	Inf	0	1	1	regulation of ATP:ADP antiporter activity
GO:0002520	0.005	4.966	1	5	221	immune system development
GO:0048285	0.006	4.873	1	5	225	organelle fission
GO:0051146	0.006	8.884	0	3	73	striated muscle cell differentiation
GO:0051726	0.007	3.914	2	6	339	regulation of cell cycle

Table B.0.6: Enriched GO terms cluster 3

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0019725	0.000	4.544	2	9	273	cellular homeostasis
GO:0048016	0.000	123.121	0	2	4	inositol phosphate-mediated signaling
GO:0070989	0.000	123.121	0	2	4	oxidative demethylation
GO:0006869	0.001	7.169	1	5	94	lipid transport
GO:0030001	0.001	4.254	2	8	255	metal ion transport
GO:0006873	0.002	4.503	2	7	209	cellular ion homeostasis
GO:0034383	0.002	41.020	0	2	8	low-density lipoprotein particle clearance
GO:0048878	0.003	3.578	2	8	300	chemical homeostasis
GO:0015674	0.004	5.123	1	5	129	di-, tri-valent inorganic cation transport
GO:0016044	0.005	3.318	3	8	322	cellular membrane organization
GO:0055066	0.005	4.845	1	5	136	di-, tri-valent inorganic cation homeostasis
GO:0046942	0.007	5.777	1	4	91	carboxylic acid transport
GO:0006309	0.008	17.563	0	2	16	DNA fragmentation involved in apoptotic nuclear change
GO:0001920	0.008	Inf	0	1	1	negative regulation of receptor recycling
GO:0002023	0.008	Inf	0	1	1	reduction of food intake in response to dietary excess
GO:0007518	0.008	Inf	0	1	1	myoblast cell fate determination
GO:0007527	0.008	Inf	0	1	1	adult somatic muscle development
GO:0009822	0.008	Inf	0	1	1	alkaloid catabolic process
GO:0019556	0.008	Inf	0	1	1	histidine catabolic process to glutamate and formamide
GO:0032493	0.008	Inf	0	1	1	response to bacterial lipoprotein
GO:0042245	0.008	Inf	0	1	1	RNA repair
GO:0042737	0.008	Inf	0	1	1	drug catabolic process
GO:0045795	0.008	Inf	0	1	1	positive regulation of cell volume
GO:0060100	0.008	Inf	0	1	1	positive regulation of phagocytosis, engulfment
GO:0070634	0.008	Inf	0	1	1	transepithelial ammonium transport
GO:0070684	0.008	Inf	0	1	1	seminal clot liquefaction
GO:0071221	0.008	Inf	0	1	1	cellular response to bacterial lipopeptide
GO:0015909	0.009	16.390	0	2	17	long-chain fatty acid transport
GO:0006879	0.010	15.364	0	2	18	cellular iron ion homeostasis
GO:0006940	0.010	15.364	0	2	18	regulation of smooth muscle contraction
GO:0007200	0.010	15.364	0	2	18	activation of phospholipase C activity by G-protein coupled receptor protein signaling pathway coupled to IP3 second messenger
GO:0006875	0.010	5.175	1	4	101	cellular metal ion homeostasis

Table B.0.7: Enriched GO terms cluster 4

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0019432	0.003	30.492	0	2	9	triglyceride biosynthetic process
GO:0006956	0.005	23.711	0	2	11	complement activation
GO:0046460	0.005	23.711	0	2	11	neutral lipid biosynthetic process
GO:0006364	0.006	6.126	1	4	75	rRNA processing
GO:0007188	0.006	8.980	0	3	39	G-protein signaling, coupled to cAMP nucleotide second messenger
GO:0051605	0.006	8.980	0	3	39	protein maturation by peptide bond cleavage
GO:0045104	0.007	19.395	0	2	13	intermediate filament cytoskeleton organization
GO:0046504	0.007	19.395	0	2	13	glycerol ether biosynthetic process
GO:0051346	0.008	8.078	0	3	43	negative regulation of hydrolase activity
GO:0019932	0.008	4.380	1	5	130	second-messenger-mediated signaling
GO:0034660	0.008	3.675	2	6	186	ncRNA metabolic process
GO:0043154	0.009	16.407	0	2	15	negative regulation of caspase activity
GO:0006657	0.010	Inf	0	1	1	CDP-choline pathway

Table B.0.8: Enriched GO terms cluster 5

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006641	0.006	8.912	0	3	31	triglyceride metabolic process
GO:0010888	0.006	20.633	0	2	10	negative regulation of lipid storage
GO:0009225	0.008	18.338	0	2	11	nucleotide-sugar metabolic process
GO:0009266	0.008	5.488	1	4	65	response to temperature stimulus
GO:0001958	0.009	16.502	0	2	12	endochondral ossification
GO:0030279	0.009	16.502	0	2	12	negative regulation of ossification
GO:0030835	0.009	16.502	0	2	12	negative regulation of actin filament depolymerization
GO:0006638	0.009	7.557	0	3	36	neutral lipid metabolic process
GO:0007044	0.009	7.557	0	3	36	cell-substrate junction assembly

Table B.0.9: Enriched GO terms cluster 6

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006414	0.001	8.509	1	5	88	translational elongation
GO:0042273	0.002	39.356	0	2	9	ribosomal large subunit biogenesis
GO:0006446	0.002	13.291	0	3	34	regulation of translational initiation
GO:0008152	0.007	2.167	39	49	5214	metabolic process
GO:0045446	0.007	18.042	0	2	17	endothelial cell differentiation
GO:0000027	0.008	Inf	0	1	1	ribosomal large subunit assembly
GO:0006043	0.008	Inf	0	1	1	glucosamine catabolic process
GO:0006747	0.008	Inf	0	1	1	FAD biosynthetic process
GO:0009078	0.008	Inf	0	1	1	pyruvate family amino acid metabolic process
GO:0009107	0.008	Inf	0	1	1	lipoate biosynthetic process
GO:0015889	0.008	Inf	0	1	1	cobalamin transport
GO:0034141	0.008	Inf	0	1	1	positive regulation of toll-like receptor 3 signaling pathway
GO:0034145	0.008	Inf	0	1	1	positive regulation of toll-like receptor 4 signaling pathway
GO:0042726	0.008	Inf	0	1	1	riboflavin and derivative metabolic process
GO:0042851	0.008	Inf	0	1	1	L-alanine metabolic process

Table B.0.10: Enriched GO terms cluster 7

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0009615	0.000	6.580	2	11	102	response to virus
GO:0051704	0.000	2.980	9	24	483	multi-organism process
GO:0009607	0.000	4.167	4	14	204	response to biotic stimulus
GO:0003417	0.000	156.584	0	3	4	growth plate cartilage development
GO:0001937	0.000	26.245	0	4	12	negative regulation of endothelial cell proliferation
GO:0006952	0.000	3.292	6	17	303	defense response
GO:0002544	0.000	31.301	0	3	8	chronic inflammatory response
GO:0008218	0.000	Inf	0	2	2	bioluminescence
GO:0018401	0.000	Inf	0	2	2	peptidyl-proline hydroxylation to 4-hydroxy-L-proline
GO:0048870	0.000	2.887	6	16	320	cell motility
GO:0032496	0.001	5.553	1	7	74	response to lipopolysaccharide
GO:0030198	0.001	5.739	1	6	62	extracellular matrix organization
GO:0018208	0.001	103.716	0	2	3	peptidyl-proline modification
GO:0007160	0.001	4.702	2	7	86	cell-matrix adhesion
GO:0060021	0.001	9.527	0	4	26	palate development
GO:0042060	0.002	3.835	2	8	119	wound healing
GO:0001867	0.002	51.852	0	2	4	complement activation, lectin pathway
GO:0022617	0.002	51.852	0	2	4	extracellular matrix disassembly
GO:0002673	0.002	14.217	0	3	14	regulation of acute inflammatory response
GO:0006986	0.002	5.977	1	5	49	response to unfolded protein
GO:0043542	0.003	5.844	1	5	50	endothelial cell migration
GO:0051271	0.003	5.716	1	5	51	negative regulation of cellular component movement
GO:0009617	0.003	3.944	2	7	101	response to bacterium
GO:0060348	0.003	7.221	1	4	33	bone development
GO:0006621	0.004	34.563	0	2	5	protein retention in ER lumen
GO:0060560	0.004	11.167	0	3	17	developmental growth involved in morphogenesis
GO:0006955	0.005	2.377	6	14	332	immune response
GO:0042493	0.005	3.053	3	9	166	response to drug
GO:0007155	0.005	2.119	9	18	481	cell adhesion
GO:0010955	0.005	25.919	0	2	6	negative regulation of protein maturation by peptide bond cleavage
GO:0018298	0.005	25.919	0	2	6	protein-chromophore linkage
GO:0031622	0.005	25.919	0	2	6	positive regulation of fever
GO:0033687	0.005	25.919	0	2	6	osteoblast proliferation
GO:0048512	0.005	25.919	0	2	6	circadian behavior
GO:0050820	0.005	25.919	0	2	6	positive regulation of coagulation
GO:0010033	0.006	1.984	11	20	571	response to organic substance
GO:0006509	0.006	9.193	0	3	20	membrane protein ectodomain proteolysis
GO:0046824	0.006	9.193	0	3	20	positive regulation of nucleocytoplasmic transport
GO:0060350	0.006	9.193	0	3	20	endochondral bone morphogenesis
GO:0006878	0.007	20.733	0	2	7	cellular copper ion homeostasis
GO:0031650	0.007	20.733	0	2	7	regulation of heat generation
GO:0070613	0.007	20.733	0	2	7	regulation of protein processing
GO:0060326	0.008	5.654	1	4	41	cell chemotaxis
GO:0009266	0.008	4.374	1	5	65	response to temperature stimulus
GO:0019221	0.008	4.374	1	5	65	cytokine-mediated signaling pathway
GO:0023038	0.008	4.374	1	5	65	signal initiation by diffusible mediator
GO:0042476	0.008	5.505	1	4	42	odontogenesis
GO:0042330	0.008	3.675	2	6	92	taxis
GO:0051604	0.009	4.232	1	5	67	protein maturation
GO:0031638	0.009	17.275	0	2	8	zymogen activation
GO:0046782	0.009	17.275	0	2	8	regulation of viral transcription

Table B.0.11: Enriched GO terms cluster 8

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0033554	0.000	3.678	5	16	497	cellular response to stress
GO:0006281	0.000	5.220	2	8	174	DNA repair
GO:0051301	0.001	3.963	3	10	274	cell division
GO:0000012	0.001	97.699	0	2	4	single strand break repair
GO:0090304	0.001	2.066	24	38	2319	nucleic acid metabolic process
GO:0046006	0.002	48.837	0	2	6	regulation of activated T cell proliferation
GO:0006266	0.002	39.065	0	2	7	DNA ligation
GO:0033151	0.002	39.065	0	2	7	V(D)J recombination
GO:0060429	0.002	3.709	2	8	229	epithelium development
GO:0006265	0.003	32.550	0	2	8	DNA topological change
GO:0044260	0.003	1.876	38	51	3655	cellular macromolecule metabolic process
GO:0034641	0.004	1.846	30	42	2849	cellular nitrogen compound metabolic process
GO:0050789	0.004	1.848	41	54	3990	regulation of biological process
GO:0007059	0.004	6.567	1	4	66	chromosome segregation
GO:0042176	0.005	6.517	1	4	65	regulation of protein catabolic process
GO:0048524	0.005	24.407	0	2	10	positive regulation of viral reproduction
GO:0051983	0.005	24.407	0	2	10	regulation of chromosome segregation
GO:0007283	0.005	4.173	2	6	151	spermatogenesis
GO:0051054	0.005	9.536	0	3	34	positive regulation of DNA metabolic process
GO:0042130	0.006	21.692	0	2	11	negative regulation of T cell proliferation
GO:0043043	0.006	21.692	0	2	11	peptide biosynthetic process
GO:0010564	0.006	4.765	1	5	110	regulation of cell cycle process
GO:0031324	0.006	2.444	6	13	571	negative regulation of cellular metabolic process
GO:0010605	0.006	2.416	6	13	577	negative regulation of macromolecule metabolic process
GO:0051726	0.007	3.362	2	7	229	regulation of cell cycle
GO:0032945	0.008	17.744	0	2	13	negative regulation of mononuclear cell proliferation
GO:0070534	0.008	17.744	0	2	13	protein K63-linked ubiquitination
GO:0006310	0.008	5.514	1	4	76	DNA recombination
GO:0006302	0.008	7.984	0	3	40	double-strand break repair
GO:0022411	0.009	5.363	1	4	78	cellular component disassembly
GO:0008284	0.009	2.960	3	8	283	positive regulation of cell proliferation
GO:0042787	0.009	16.263	0	2	14	protein ubiquitination involved in ubiquitin-dependent protein catabolic process
GO:0060603	0.009	16.263	0	2	14	mammary gland duct morphogenesis
GO:0002237	0.009	5.291	1	4	79	response to molecule of bacterial origin
GO:0006260	0.009	3.569	2	6	175	DNA replication
GO:0009893	0.010	2.208	7	14	679	positive regulation of metabolic process

Table B.0.12: Enriched GO terms cluster 9

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0044271	0.000	7.458	2	10	295	cellular nitrogen compound biosynthetic process
GO:0034404	0.000	7.661	1	7	189	nucleobase, nucleoside and nucleotide biosynthetic process
GO:0009126	0.000	42.671	0	3	16	purine nucleoside monophosphate metabolic process
GO:0006164	0.000	7.966	1	6	153	purine nucleotide biosynthetic process
GO:0009161	0.000	29.175	0	3	22	ribonucleoside monophosphate metabolic process
GO:0009113	0.001	72.391	0	2	7	purine base biosynthetic process
GO:0009124	0.001	9.266	0	4	85	nucleoside monophosphate biosynthetic process
GO:0018130	0.002	14.553	0	3	41	heterocycle biosynthetic process
GO:0009260	0.002	8.620	1	4	91	ribonucleotide biosynthetic process
GO:0009168	0.002	32.881	0	2	13	purine ribonucleoside monophosphate biosynthetic process
GO:0009117	0.003	4.007	2	7	348	nucleotide metabolic process
GO:0009112	0.004	25.825	0	2	16	nucleobase metabolic process
GO:0002882	0.006	Inf	0	1	1	positive regulation of chronic inflammatory response to non-antigenic stimulus
GO:0006168	0.006	Inf	0	1	1	adenine salvage
GO:0006756	0.006	Inf	0	1	1	AMP phosphorylation
GO:0009088	0.006	Inf	0	1	1	threonine biosynthetic process
GO:0009159	0.006	Inf	0	1	1	deoxyribonucleoside monophosphate catabolic process
GO:0010701	0.006	Inf	0	1	1	positive regulation of norepinephrine secretion
GO:0032713	0.006	Inf	0	1	1	negative regulation of interleukin-4 production
GO:0042795	0.006	Inf	0	1	1	snRNA transcription from RNA polymerase II promoter
GO:0042796	0.006	Inf	0	1	1	snRNA transcription from RNA polymerase III promoter
GO:0045650	0.006	Inf	0	1	1	negative regulation of macrophage differentiation
GO:0046083	0.006	Inf	0	1	1	adenine metabolic process

Table B.0.13: General model structure. In bold, the part of the variable's update rule that reflects the modeled property referenced in the column Explanation. Column P contains the probabilities of the update rule

Represented property	Update rule	Probability	Explanation
Auto activation of gene modules	$m_e^A(t+1) = \mathbf{m_e^A(t)} \wedge \neg(m_e^B(t) \vee m_e^P(t)) \wedge \neg m_{m/hc}^A(t)$	0.5/0.5	Regulatory proteins are closely coregulated and are often connected by positive feedback loops. (Boyer et al., 2005; Chickarmane and Peterson, 2008; MacArthur, Please, and Oreffo, 2008)
Pluripotency module activating DNA methylation through variable *DNMT* expression	$dnmt(t+1) = \mathbf{m_e^P(t)} \vee m_e^E(t) \vee dnmt(t)$	0.99	*DNMT3* coregulated with Pluripotency genes. DNMT3 methylates unspecifically (Adewumi, Aflatoonian, and Ahrlund-Richter, 2007; Mah et al., 2011)
Mutual inhibition of gene modules	$m_e^A(t+1) = m_e^A(t) \wedge \boldsymbol{\neg(}m_e^B\boldsymbol{(}t\boldsymbol{)} \boldsymbol{\vee} m_e^P\boldsymbol{(}t\boldsymbol{))} \wedge \neg m_{m/hc}^A(t)$	0.5/0.5	Master Regulators inhibit other master regulators, competing lineages repress each other (MacArthur, Please, and Oreffo, 2008; Niwa et al., 2005; Ralston and Rossant, 2005)
Heterochromatin increases probability for DNA methylation	$m_m^A(t+1) = m_m^A(t) \vee dnmt(t) \wedge \mathbf{m_{hc}^A(t)}$	0.05	Interaction via G9a complex: DNMT3A/B bind to nucleosomes with methylated histones such as H3K9me and methylates DNA (Cedar and Bergman, 2009)
Heterochromatin formation is inhibited by appropriate gene module	$m_{hc}^A(t+1) = m_{hc}^A(t) \vee m_m^A(t) \wedge \boldsymbol{\neg} m_e^A\boldsymbol{(}t\boldsymbol{)}$	0.11	G9a binds specific sequences (Epsztejn-Litman et al., 2008)
DNA methylation increases probability for heterochromatin formation	$m_{hc}^A(t+1) = m_{hc}^A(t) \vee \mathbf{m_m^A(t)} \wedge \neg m_e^A(t)$	0.17	Promotes chromatin inheritance after mitosis (Thomson et al., 2010)
DNA demethylation slower than other factors	$m_m^A(t+1) = m_m^A(t) \wedge \mathbf{demeth(t)}$	0.02	Passive cell cycle dependent demethylation through variable DNMT1 activity after mitosis (Li, Bestor, and Jaenisch, 1992)
DNA demethylation is faster in euchromatin	$m_m^A(t+1) = m_m^A(t) \wedge (\mathbf{demeth(t)} \vee \mathbf{m_{hc}^A})$	0.03	Histone deacetylase (HDAC) inhibitor TSA induces global and specific DNA demethylation (Ou et al., 2007)
Methylation not necessary to downregulate retroviral gene expression	$m_e^E(t+1) = \neg m_{hc}^E(t) \boldsymbol{\neg \vee} m_m^E\boldsymbol{(}t\boldsymbol{)}$	0.5	Retroviral silencing is DNMT3A/B independent in the first 10 days of reprogramming (Pannell et al., 2000)
Retroviral gene demethylation is very slow in absence of DNMT3A/B or DNMT1	$m_m^E(t+1) = m_m^E(t) \wedge \boldsymbol{(}\neg demeth\boldsymbol{(}t\boldsymbol{)} \vee dnmt\boldsymbol{(}t\boldsymbol{))}$	0.001	
Retroviral gene heterochromatin dynamics	$m_{hc}^E(t+1) = m_{hc}^E(t) \vee m_e^P\boldsymbol{(}t\boldsymbol{)}$	0.1	A complex between HDAC and NANOG (NODE complex responsible for the silencing of developmental genes) could account for retroviral silencing (Hotta and Ellis, 2008; Liang et al., 2008)

Table B.0.14: Experimental findings from literature compared to simulation results.

Experimental Finding	Theoretical validation by our model
Somatic cells can be reprogrammed to iPSCs upon viral delivery of pluripotency factors with a very low efficiency (Takahashi and Yamanaka, 2006b)	Reprogramming experiment of our main model (Figure 4.2.12)
iPSCs can be re-differentiated into various kinds of tissues (all three germ layers) (Takahashi and Yamanaka, 2006b)	Differentiation experiment of our main model (Figure 4.2.11)
ESCs have more euchromatin and accumulate high condensed heterochromatin as differentiation progresses (Francastel et al., 2000)	In the differentiation of the pluripotent state, which still consists of a distribution across several different chromatin and methylation configurations, we can observe a transition to more sharply defined states, which mostly include heterochromatin and methylation compositions (Figure 4.2.11 A)
DNA methylation is essential for chromatin structure during development (Hashimshony et al., 2003)	In models lacking DNA methylation, differentiation as well as reprogramming are abolished and cells will not be able to pass to other states in the state space
Treatment of partially differentiated ES cells with the DNA demethylating agent 5-azacytidine (5-AzaC) induces de-differentiation (Tsuji-Takayama et al., 2004)	When starting from partly differentiated states in models with spontaneous demethylation mimicking 5-AzaC treatment, we observe de-differentiation and even efficient reprogramming
Knockdown of DnmtI reactivates retroviral genes (Wernig et al., 2007)	In models mimicking DnmtI knockdown (e.g. spontaneous demethylation in or no methylation in simulation from the iPS state leads to partial reactivation of retroviral genes
Dnmt3a and Dnmt3b are not required for retroviral silencing in the first 10 days of reprogramming (Hotta and Ellis, 2008; Pannell et al., 2000)	In models without dnmt activity we can still observe silencing of retroviral genes (results not explicitly shown)

Continued on next page

Table B.0.14: Experimental findings from literature compared to simulation results.

Experimental Finding	Theoretical validation by our model
The histone deacetylase (HDAC) inhibitor valproic acid is capable of enhancing reprogramming efficiency (Huangfu et al., 2008b)	In models where the probability for heterochromatin formation is down-regulated (mimicking inhibition of HDAC) we observe a slight increase in the reprogramming efficiency (Figure 4.2.14).

C
Influenza A

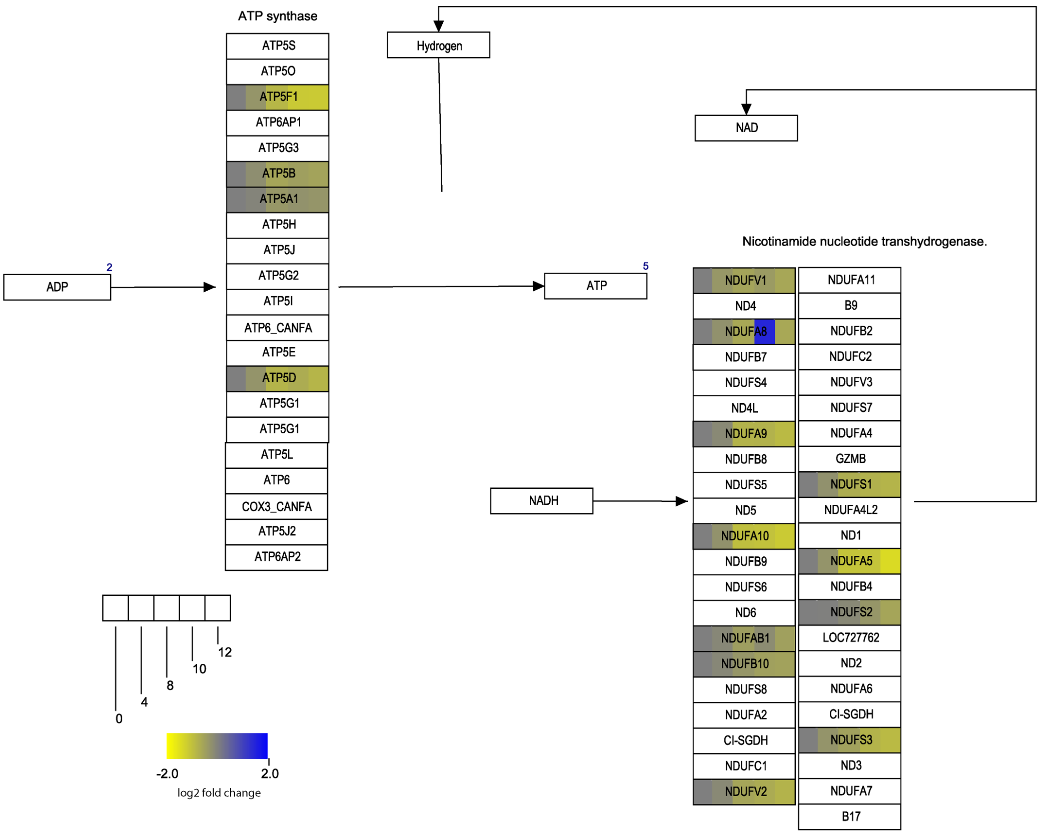


Figure C.0.1: Subunits of the ATP synthetase are downregulated in our experiments.

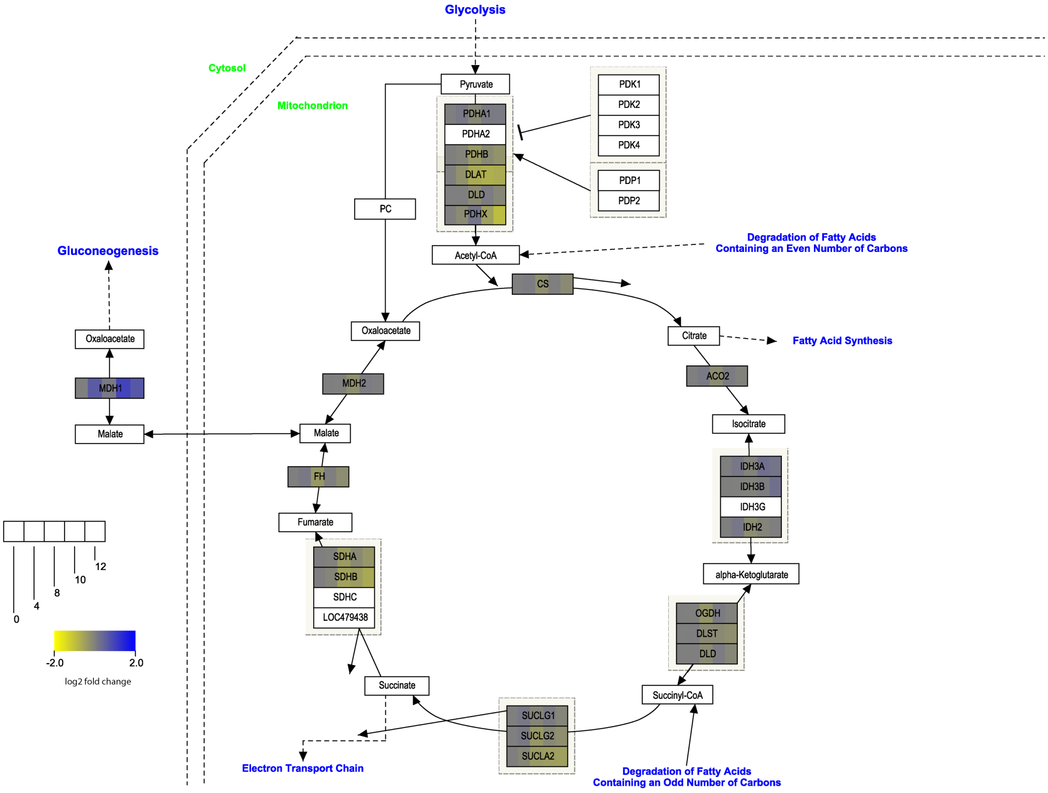


Figure C.0.2: TCA cycle and its regulation in the early phase of infection

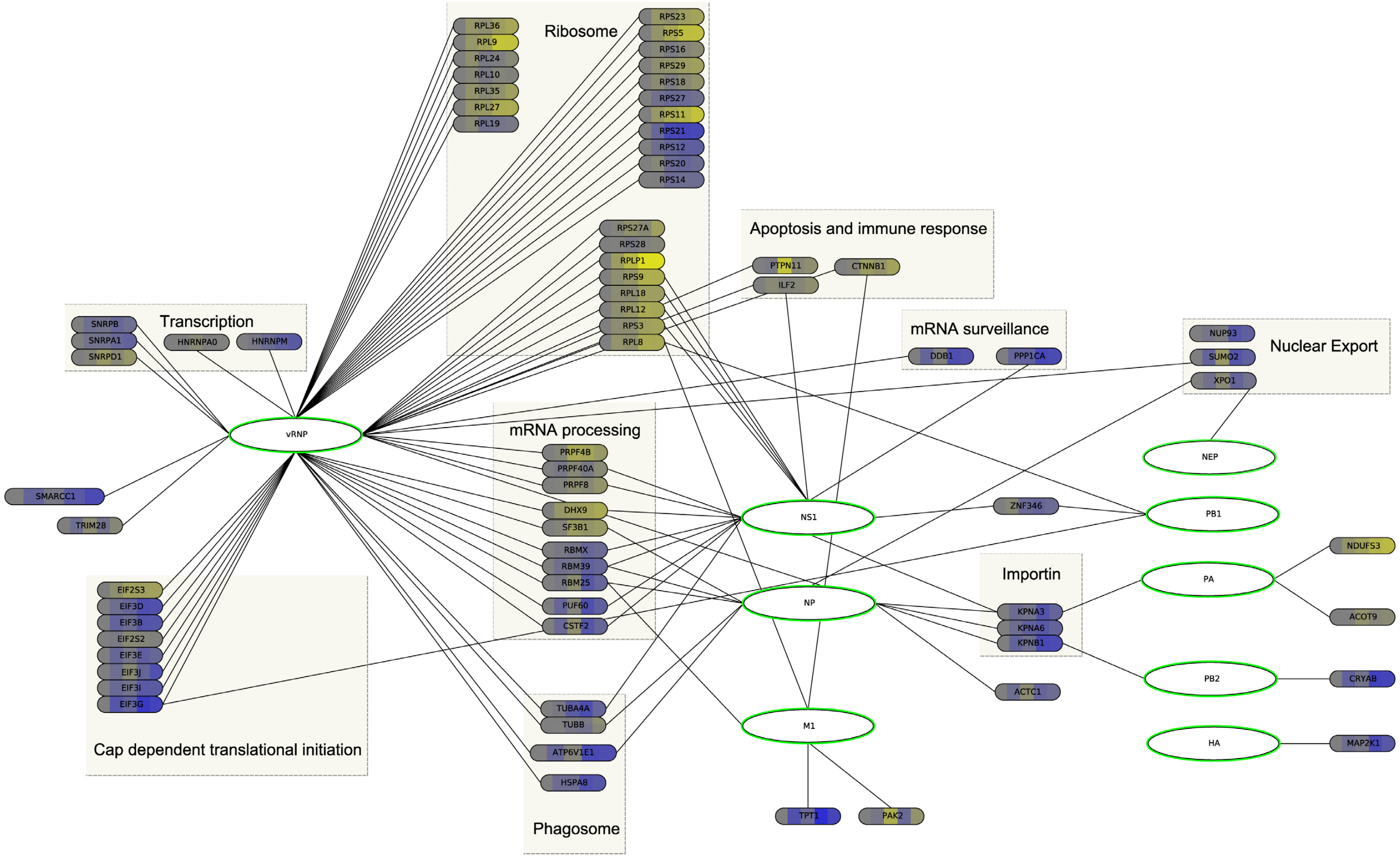


Figure C.0.3: The network of host and virus proteins was derived from a number of high-throughput studies by Watanabe, Watanabe, and Kawaoka (2010).

Table C.0.1: Enriched GO terms cluster 1

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006414	0	4.073	5	16	79	translational elongation
GO:0006270	0	29.847	0	4	6	DNA-dependent DNA replication initiation
GO:0000003	0.001	1.955	28	43	427	reproduction
GO:0042303	0.001	4.306	3	9	42	molting cycle
GO:0018996	0.001	4.175	3	9	42	molting cycle, collagen and cuticulin-based cuticle
GO:0010467	0.001	1.814	36	51	551	gene expression
GO:0006892	0.002	5.648	1	6	22	post-Golgi vesicle-mediated transport
GO:0042274	0.002	5.648	1	6	22	ribosomal small subunit biogenesis
GO:0009059	0.003	1.813	25	38	390	macromolecule biosynthetic process
GO:0033572	0.004	Inf	0	2	2	transferrin transport
GO:0051304	0.004	Inf	0	2	2	chromosome separation
GO:0007017	0.006	2.156	10	18	149	microtubule-based process
GO:0030855	0.008	4.093	2	6	28	epithelial cell differentiation
GO:0022404	0.008	6.606	1	4	13	molting cycle process
GO:0033205	0.008	11.086	0	3	7	cell cycle cytokinesis
GO:0035017	0.008	11.086	0	3	7	cuticle pattern formation

Table C.0.2: Enriched GO terms cluster 2

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0031397	0	10.836	6	25	37	negative regulation of protein ubiquitination
GO:0051352	0	11.318	6	24	35	negative regulation of ligase activity
GO:0051436	0	11.318	6	24	35	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
GO:0051437	0	9.996	7	25	38	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
GO:0051351	0	9.029	7	26	41	positive regulation of ligase activity
GO:0051438	0	8.459	7	26	42	regulation of ubiquitin-protein ligase activity
GO:0032446	0	5.462	12	35	69	protein modification by small protein conjugation
GO:0031145	0	9.564	6	24	37	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
GO:0031398	0	7.956	7	26	43	positive regulation of protein ubiquitination
GO:0006511	0	4.923	13	37	77	ubiquitin-dependent protein catabolic process
GO:0044257	0	4.6	14	38	82	cellular protein catabolic process
GO:0043632	0	4.682	14	37	79	modification-dependent macromolecule catabolic process
GO:0048523	0	2.389	56	95	327	negative regulation of cellular process
GO:0010498	0	5.404	10	29	57	proteasomal protein catabolic process
GO:0032270	0	4.221	13	33	74	positive regulation of cellular protein metabolic process
GO:0044092	0	4.184	12	32	72	negative regulation of molecular function
GO:0048522	0	2.185	58	94	341	positive regulation of cellular process
GO:0031399	0	3.363	18	40	103	regulation of protein modification process
GO:0032269	0	4.146	12	31	70	negative regulation of cellular protein metabolic process
GO:0009057	0	3.08	21	45	123	macromolecule catabolic process
GO:0044093	0	3	21	44	122	positive regulation of molecular function
GO:0050790	0	2.636	25	48	145	regulation of catalytic activity
GO:0051246	0	2.484	28	52	164	regulation of protein metabolic process
GO:0006508	0	2.525	26	48	149	proteolysis
GO:0022402	0	2.585	23	44	138	cell cycle process
GO:0031323	0	1.883	65	95	378	regulation of cellular metabolic process
GO:0010604	0	2.482	23	43	134	positive regulation of macromolecule metabolic process
GO:0043412	0	2	47	73	272	macromolecule modification
GO:0009056	0	5.86	4	14	27	catabolic process
GO:0000278	0	3.223	11	25	68	mitotic cell cycle
GO:0010605	0	2.158	26	45	154	negative regulation of macromolecule metabolic process
GO:0050789	0	1.666	126	155	789	regulation of biological process
GO:0012501	0	1.945	36	57	212	programmed cell death
GO:0019320	0	4.378	5	14	30	hexose catabolic process
GO:0032271	0	4.378	5	14	30	regulation of protein polymerization
GO:0006606	0	3.648	7	16	38	protein import into nucleus
GO:0044087	0	3.058	9	20	53	regulation of cellular component biogenesis
GO:0046164	0	4.118	5	14	31	alcohol catabolic process
GO:0007015	0	2.948	10	21	57	actin filament organization
GO:0033365	0	2.59	13	25	74	protein localization to organelle
GO:0007346	0	3.886	5	14	32	regulation of mitotic cell cycle
GO:0006096	0	4.971	4	11	22	glycolysis
GO:0016265	0	1.814	38	57	222	death
GO:0071822	0.001	1.86	32	49	186	protein complex subunit organization
GO:0044275	0.001	3.409	6	15	37	cellular carbohydrate catabolic process
GO:0030216	0.001	8.628	2	7	11	keratinocyte differentiation
GO:0006458	0.001	6.586	2	8	14	'de novo' protein folding
GO:0051086	0.001	Inf	1	4	4	chaperone mediated protein folding independent of cofactor
GO:0070271	0.001	1.813	30	45	173	protein complex biogenesis
GO:0072521	0.001	2.17	15	27	90	purine-containing compound metabolic process
GO:0006413	0.002	3.233	6	13	33	translational initiation

Continued on next page

Table C.0.2: Enriched GO terms cluster 2

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0044248	0.002	1.649	41	58	241	cellular catabolic process
GO:0009124	0.002	3.636	4	11	26	nucleoside monophosphate biosynthetic process
GO:0009117	0.002	1.96	20	32	115	nucleotide metabolic process
GO:0051533	0.002	12.264	1	5	7	positive regulation of NFAT protein import into nucleus
GO:0048610	0.002	2.054	17	28	97	cellular process involved in reproduction
GO:0016044	0.002	1.666	38	54	222	cellular membrane organization
GO:0008064	0.002	3.31	5	12	30	regulation of actin polymerization or depolymerization
GO:0000279	0.002	1.813	26	39	149	M phase
GO:0006974	0.003	2.1	15	25	85	response to DNA damage stimulus
GO:0044267	0.003	1.529	61	79	385	cellular protein metabolic process
GO:0016192	0.003	1.604	43	59	250	vesicle-mediated transport
GO:0042981	0.003	1.723	30	44	175	regulation of apoptosis
GO:0001821	0.004	19.574	1	4	5	histamine secretion
GO:0002349	0.004	19.574	1	4	5	histamine production involved in inflammatory response
GO:0002553	0.004	19.574	1	4	5	histamine secretion by mast cell
GO:0008633	0.004	19.574	1	4	5	activation of pro-apoptotic gene products
GO:0035094	0.004	19.574	1	4	5	response to nicotine
GO:0035308	0.004	19.574	1	4	5	negative regulation of protein dephosphorylation
GO:0046822	0.004	3.533	4	10	24	regulation of nucleocytoplasmic transport
GO:0006164	0.004	2.367	10	18	56	purine nucleotide biosynthetic process
GO:0010941	0.004	1.684	31	45	182	regulation of cell death
GO:0034621	0.004	1.684	31	45	182	cellular macromolecular complex subunit organization
GO:0007010	0.004	1.663	33	47	197	cytoskeleton organization
GO:0033043	0.004	2.101	13	23	78	regulation of organelle organization
GO:0034654	0.004	2.122	13	22	74	nucleobase-containing biosynthetic process
GO:0046686	0.004	2.975	5	12	32	response to cadmium ion
GO:0009161	0.005	4.921	2	7	14	ribonucleoside monophosphate metabolic process
GO:0031109	0.005	4.921	2	7	14	microtubule polymerization or depolymerization
GO:0051494	0.005	5.985	2	6	11	negative regulation of cytoskeleton organization
GO:0006183	0.005	Inf	1	3	3	GTP biosynthetic process
GO:0009220	0.005	Inf	1	3	3	pyrimidine ribonucleotide biosynthetic process
GO:0045138	0.005	Inf	1	3	3	tail tip morphogenesis
GO:0048566	0.005	Inf	1	3	3	embryonic digestive tract development
GO:0021761	0.005	8.171	1	5	8	limbic system development
GO:0007052	0.005	2.009	14	24	84	mitotic spindle organization
GO:0043624	0.005	3.295	4	10	25	cellular protein complex disassembly
GO:0030041	0.006	4.003	3	8	18	actin filament polymerization
GO:0010035	0.006	1.799	21	32	122	response to inorganic substance
GO:0051234	0.006	1.395	100	120	585	establishment of localization
GO:0009119	0.006	3.942	3	8	18	ribonucleoside metabolic process
GO:0031333	0.006	3.942	3	8	18	negative regulation of protein complex assembly
GO:0018130	0.006	1.975	15	24	85	heterocycle biosynthetic process
GO:0009150	0.007	2.156	11	19	63	purine ribonucleotide metabolic process
GO:0006006	0.007	2.342	9	16	50	glucose metabolic process
GO:0008360	0.007	3.415	4	9	22	regulation of cell shape
GO:0009314	0.007	2.405	8	15	46	response to radiation
GO:0007030	0.007	4.303	3	7	15	Golgi organization
GO:0042990	0.007	4.303	3	7	15	regulation of transcription factor import into nucleus
GO:0051650	0.007	4.303	3	7	15	establishment of vesicle localization
GO:0032984	0.008	3.087	4	10	26	macromolecular complex disassembly
GO:0043627	0.008	2.863	5	11	30	response to estrogen stimulus
GO:0043623	0.008	1.989	13	22	78	cellular protein complex assembly
GO:0007017	0.008	1.675	26	37	149	microtubule-based process
GO:0009260	0.009	2.419	7	14	43	ribonucleotide biosynthetic process
GO:0051707	0.009	1.966	13	22	78	response to other organism
GO:0009168	0.009	4.909	2	6	12	purine ribonucleoside monophosphate biosynthetic process
GO:0030837	0.009	4.909	2	6	12	negative regulation of actin filament polymerization
GO:0080090	0.009	1.596	32	44	201	regulation of primary metabolic process
GO:0048513	0.009	1.393	81	98	470	organ development
GO:0050793	0.009	1.672	25	36	145	regulation of developmental process
GO:0009408	0.009	2.395	7	14	43	response to heat
GO:0006000	0.01	9.781	1	4	6	fructose metabolic process
GO:0007021	0.01	9.781	1	4	6	tubulin complex assembly
GO:0030435	0.01	9.781	1	4	6	sporulation resulting in formation of a cellular spore
GO:0051220	0.01	9.781	1	4	6	cytoplasmic sequestering of protein
GO:0032535	0.01	2.477	7	13	39	regulation of cellular component size
GO:0006733	0.01	3.169	4	9	23	oxidoreduction coenzyme metabolic process
GO:0046496	0.01	3.169	4	9	23	nicotinamide nucleotide metabolic process
GO:0072524	0.01	3.169	4	9	23	pyridine-containing compound metabolic process
GO:0002275	0.01	6.124	2	5	9	myeloid cell activation involved in immune response
GO:0002444	0.01	6.124	2	5	9	myeloid leukocyte mediated immunity
GO:0043299	0.01	6.124	2	5	9	leukocyte degranulation
GO:0048205	0.01	6.124	2	5	9	COPI coating of Golgi vesicle

Table C.0.3: Enriched GO terms cluster 3

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006414	0	4.073	5	16	79	translational elongation
GO:0006270	0	29.847	0	4	6	DNA-dependent DNA replication initiation
GO:0000003	0.001	1.955	28	43	427	reproduction
GO:0042303	0.001	4.306	3	9	42	molting cycle
GO:0018996	0.001	4.175	3	9	42	molting cycle, collagen and cuticulin-based cuticle
GO:0010467	0.001	1.814	36	51	551	gene expression
GO:0006892	0.002	5.648	1	6	22	post-Golgi vesicle-mediated transport
GO:0042274	0.002	5.648	1	6	22	ribosomal small subunit biogenesis
GO:0009059	0.003	1.813	25	38	390	macromolecule biosynthetic process
GO:0033572	0.004	Inf	0	2	2	transferrin transport
GO:0051304	0.004	Inf	0	2	2	chromosome separation
GO:0007017	0.006	2.156	10	18	149	microtubule-based process
GO:0030855	0.008	4.093	2	6	28	epithelial cell differentiation
GO:0022404	0.008	6.606	1	4	13	molting cycle process
GO:0033205	0.008	11.086	0	3	7	cell cycle cytokinesis
GO:0035017	0.008	11.086	0	3	7	cuticle pattern formation

Table C.0.4: Enriched GO terms cluster 4

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0071824	0	5.528	4	15	35	protein-DNA complex subunit organization
GO:0006334	0	6.35	4	13	28	nucleosome assembly
GO:0046942	0	5.836	3	12	27	carboxylic acid transport
GO:0006333	0	4.409	5	15	40	chromatin assembly or disassembly
GO:0055085	0	3.545	7	18	57	transmembrane transport
GO:0006323	0	3.663	6	15	45	DNA packaging
GO:0006200	0	10.051	2	7	12	ATP catabolic process
GO:0006172	0	Inf	1	4	4	ADP biosynthetic process
GO:0009136	0	Inf	1	4	4	purine nucleoside diphosphate biosynthetic process
GO:0009188	0	Inf	1	4	4	ribonucleoside diphosphate biosynthetic process
GO:0006820	0	7.374	2	8	16	anion transport
GO:0022904	0	5.168	3	10	24	respiratory electron transport chain
GO:0009206	0	4.153	4	12	33	purine ribonucleoside triphosphate biosynthetic process
GO:0015992	0	5.409	3	9	21	proton transport
GO:0015711	0	10.728	1	6	10	organic anion transport
GO:0051234	0.001	1.627	73	96	585	establishment of localization
GO:0006839	0.001	3.787	4	12	35	mitochondrial transport
GO:0009142	0.001	3.787	4	12	35	nucleoside triphosphate biosynthetic process
GO:0015980	0.001	3.002	7	16	55	energy derivation by oxidation of organic compounds
GO:0008203	0.001	5.751	2	8	18	cholesterol metabolic process
GO:0034220	0.001	3.982	4	11	31	ion transmembrane transport
GO:0009179	0.001	28.414	1	4	5	purine ribonucleoside diphosphate metabolic process
GO:0019233	0.001	28.414	1	4	5	sensory perception of pain
GO:0033555	0.001	11.876	1	5	8	multicellular organismal response to stress
GO:0040008	0.001	1.861	25	40	203	regulation of growth
GO:0040010	0.002	1.98	19	31	148	positive regulation of growth rate
GO:0009853	0.002	Inf	0	3	3	photorespiration
GO:0015800	0.002	8.902	1	5	9	acidic amino acid transport
GO:0006119	0.003	4.416	3	8	21	oxidative phosphorylation
GO:0051240	0.003	2.748	6	14	51	positive regulation of multicellular organismal process
GO:0015986	0.003	5.01	2	7	17	ATP synthesis coupled proton transport
GO:0014911	0.003	14.198	1	4	6	positive regulation of smooth muscle cell migration
GO:0015837	0.003	3.806	3	9	26	amine transport
GO:0030001	0.004	3.101	5	11	37	metal ion transport
GO:0030155	0.004	4.552	2	7	18	regulation of cell adhesion
GO:0006816	0.005	3.822	3	8	23	calcium ion transport
GO:0007568	0.006	1.961	14	24	114	aging
GO:0019915	0.006	4.17	2	7	19	lipid storage
GO:0006814	0.006	9.46	1	4	7	sodium ion transport
GO:0019218	0.006	9.46	1	4	7	regulation of steroid metabolic process
GO:0051276	0.007	1.99	13	22	103	chromosome organization
GO:0009199	0.007	2.54	6	13	50	ribonucleoside triphosphate metabolic process
GO:0016126	0.007	4.754	2	6	15	sterol biosynthetic process
GO:0001662	0.007	21.219	1	3	4	behavioral fear response
GO:0006971	0.007	21.219	1	3	4	hypotonic response
GO:0046513	0.007	21.219	1	3	4	ceramide biosynthetic process
GO:0001817	0.007	5.927	1	5	11	regulation of cytokine production
GO:0008610	0.008	2.234	9	16	68	lipid biosynthetic process
GO:0022610	0.008	2.234	9	16	68	biological adhesion
GO:0009144	0.008	2.471	6	13	51	purine nucleoside triphosphate metabolic process
GO:0016049	0.008	2.874	4	10	35	cell growth
GO:0006812	0.009	3.752	3	7	21	cation transport

Table C.0.5: Enriched GO terms cluster 5

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0043436	0	2.601	13	28	181	oxoacid metabolic process
GO:0055114	0	2.555	13	26	169	oxidation-reduction process
GO:0006105	0	25.632	0	4	6	succinate metabolic process
GO:0002520	0.001	3.102	6	15	80	immune system development
GO:0048878	0.001	3.005	6	15	82	chemical homeostasis
GO:0040027	0.001	17.078	1	4	7	negative regulation of vulval development
GO:0061062	0.001	17.078	1	4	7	regulation of nematode larval development
GO:0043648	0.001	6.658	1	6	18	dicarboxylic acid metabolic process
GO:0065008	0.001	1.917	24	38	322	regulation of biological quality
GO:0032496	0.002	5.05	2	7	25	response to lipopolysaccharide
GO:0032787	0.002	2.783	6	14	81	monocarboxylic acid metabolic process
GO:0006084	0.002	4.781	2	7	26	acetyl-CoA metabolic process
GO:0051186	0.005	2.679	5	12	71	cofactor metabolic process
GO:0045454	0.005	4.551	2	6	23	cell redox homeostasis
GO:0050804	0.005	4.551	2	6	23	regulation of synaptic transmission
GO:0001837	0.005	Inf	0	2	2	epithelial to mesenchymal transition
GO:0048147	0.005	Inf	0	2	2	negative regulation of fibroblast proliferation
GO:0051588	0.005	Inf	0	2	2	regulation of neurotransmitter transport
GO:2000300	0.005	Inf	0	2	2	regulation of synaptic vesicle exocytosis
GO:0007399	0.006	1.867	17	27	225	nervous system development
GO:0006091	0.006	2.486	6	13	82	generation of precursor metabolites and energy
GO:0006457	0.006	2.722	5	11	64	protein folding
GO:0030097	0.006	2.588	5	12	73	hemopoiesis
GO:0006099	0.006	5.349	1	5	17	tricarboxylic acid cycle
GO:0030866	0.006	7.303	1	4	11	cortical actin cytoskeleton organization
GO:0048581	0.006	7.303	1	4	11	negative regulation of post-embryonic development
GO:0022904	0.007	4.296	2	6	24	respiratory electron transport chain
GO:0002562	0.007	12.715	0	3	6	somatic diversification of immune receptors via germline recombination within a single locus
GO:0006664	0.007	12.715	0	3	6	glycolipid metabolic process
GO:0055072	0.007	12.715	0	3	6	iron ion homeostasis
GO:0008406	0.008	4.068	2	6	25	gonad development
GO:0045333	0.009	4.743	1	5	19	cellular respiration
GO:0006873	0.01	2.525	5	11	68	cellular ion homeostasis

Table C.0.6: Enriched GO terms cluster 6

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0006414	0	23.644	7	48	79	translational elongation
GO:0009059	0	7.008	16	57	211	macromolecule biosynthetic process
GO:0000022	0	14.465	3	20	37	mitotic spindle elongation
GO:0044260	0	3.342	71	112	857	cellular macromolecule metabolic process
GO:0044249	0	3.437	28	59	361	cellular biosynthetic process
GO:0019538	0	2.864	50	86	589	protein metabolic process
GO:0071843	0	5.065	8	27	94	cellular component biogenesis at cellular level
GO:0006364	0	7.713	3	15	38	rRNA processing
GO:0007051	0	4.144	8	24	95	spindle organization
GO:0006396	0	2.627	18	36	209	RNA processing
GO:0007017	0	2.835	13	28	149	microtubule-based process
GO:0043487	0	13.202	1	7	13	regulation of RNA stability
GO:0042254	0	8.336	2	8	20	ribosome biogenesis
GO:0008152	0	2.178	112	132	1326	metabolic process
GO:0048255	0	13.498	1	6	11	mRNA stabilization
GO:0070925	0	7.567	2	8	20	organelle assembly
GO:0042273	0	11.242	1	6	12	ribosomal large subunit biogenesis
GO:0034660	0	3.161	6	16	75	ncRNA metabolic process
GO:0000028	0.001	22.245	1	4	6	ribosomal small subunit assembly
GO:0006417	0.001	4.07	3	10	38	regulation of translation
GO:0090304	0.001	1.751	39	56	465	nucleic acid metabolic process
GO:0010556	0.001	2.103	15	27	180	regulation of macromolecule biosynthetic process
GO:0022403	0.002	2.13	14	25	164	cell cycle phase
GO:0031326	0.002	1.965	17	29	205	regulation of cellular biosynthetic process
GO:0007010	0.004	1.88	19	30	220	cytoskeleton organization
GO:0010468	0.004	1.821	21	33	250	regulation of gene expression
GO:0071826	0.004	3.392	3	9	39	ribonucleoprotein complex subunit organization
GO:0000291	0.007	Inf	0	2	2	nuclear-transcribed mRNA catabolic process. exonucleolytic
GO:0006412	0.007	2.63	5	11	83	translation
GO:0006402	0.008	5.062	1	5	16	mRNA catabolic process
GO:0042157	0.01	11.045	1	3	6	lipoprotein metabolic process

Selbstständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Max Flöttmann

Berlin, den June 20, 2013

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Arno Pro, designed by Robert Slimbach in the style of book types from the Aldine Press in Venice, and issued by Adobe in 2007. A template, which can be used to format a PhD thesis with this look and feel, has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/ or from the author at suchow@post.harvard.edu.